

# PROGNOSTIC

---

## Point pROcesses: learninG, NOnparametric STatistics and appliCations

Processus ponctuels: apprentissage, statistiques non-paramétriques et applications

CSD Principale : 5

Aide totale demandée : 142 268 euros

Durée : 36 mois

## Contents

<b>1</b>	<b>Context and positioning of the proposal</b>	<b>3</b>
1.1	Birth of the project . . . . .	3
1.2	Constitution of the team . . . . .	3
<b>2</b>	<b>Scientific and technical description</b>	<b>4</b>
2.1	Background, State of Art . . . . .	4
2.1.1	State of the Art on counting processes . . . . .	5
2.1.2	Background, Mathematical methods involved . . . . .	6
2.1.3	Concentration inequalities . . . . .	6
2.1.4	Stochastic calculus . . . . .	7
2.1.5	Statistical learning theory . . . . .	7
2.1.6	High dimension, detection of non-zero components . . . . .	7
	LASSO . . . . .	7
	Multiple tests approach . . . . .	7
	Estimation approach . . . . .	8
2.2	Originality and novelty of the proposal . . . . .	8
<b>3</b>	<b>Scientific and technical programme, project management</b>	<b>9</b>
3.1	Detailed description of the work . . . . .	9
3.1.1	Task 1: Model selection for counting processes . . . . .	9
	Subtask 1: Semiparametric models . . . . .	10
3.1.2	Task 2: Learning for counting processes . . . . .	10
	Current work (almost completed) . . . . .	10
	Subtask 2: Implementation . . . . .	11
	Subtask 3: Go beyond the Cox model, learning . . . . .	12
	Subtask 4: Sparse aggregation with exponential weights . . . . .	12
3.1.3	Task 3: Dimension reduction, the single-index/multi-index approach . . . . .	13
	Single-index for censored data (Done work). . . . .	13
	Subtask 5: Change-point single-index model . . . . .	14
	Subtask 6: Single-index and aggregation . . . . .	14
3.1.4	Task 4: High dimension . . . . .	15
	Subtask 7: Multiple testing approach . . . . .	15

	Subtask 8: Estimation approach . . . . .	15
	Subtask 9: Detection of non-zero coefficients in the Cox model . . . . .	16
	Subtask 10: High dimensionality with correlations of covariates: the elastic-net approach . . . . .	16
3.1.5	Task 5: Other estimation settings . . . . .	18
	Subtask 11: Recurrent events . . . . .	18
	Subtask 12: SIM for recurrent events . . . . .	18
	Subtask 13: Time-dependent covariates . . . . .	19
	Subtask 14: Semi-Markov process . . . . .	20
3.1.6	Task 6: Some parallel developments . . . . .	21
	Subtask 15: SVM and sparsity . . . . .	21
	Subtask 16: Form of coordinate projections . . . . .	22
	Subtask 17: $\ell_1$ -penalization and Convex Asymptotic Geometry . . . . .	23
	Subtask 18: General study of the ERM procedure in the agnostic setup . . . . .	23
3.1.7	Task 7: Applications . . . . .	25
	Subtask 19: Applications in Biology . . . . .	25
	Subtask 20: Applications in Cancerology . . . . .	25
3.2	Planning of tasks . . . . .	26
<b>4</b>	<b>Consortium organisation and description</b>	<b>27</b>
4.1	Qualification of the principal investigator . . . . .	27
4.2	Contribution and qualification of each project participant . . . . .	27
<b>5</b>	<b>Scientific justification of requested budget</b>	<b>28</b>
<b>6</b>	<b>Annexes</b>	<b>28</b>
6.1	Bibliography . . . . .	28
6.2	CV of the members . . . . .	33
6.3	Involvement of project participants to other grants . . . . .	45

# 1 Context and positioning of the proposal

## 1.1 Birth of the project

The idea that gave birth to this project comes, somehow, from an applied statistical problem in medicine. In the late 2007, A. GUILLOUX (principal investigator, see Table 1) helped to analyze a dataset collected by CRSA E13 (INSERM, UPMC) (Director: A. DUVAL, see Table 1) on 192 patients with stage III colon cancer treated by surgery and chemotherapy. Among these patients, 123 were treated by fluorouracil and leucovorin alone (FL), and 69 by FL with oxaliplatin (FOLFOX). The duration between surgery and disease recurrence (or the last hospital contact, in case of no recurrence), the status at the end of the study (sick or not), the mutational status of a particular gene (p53), and some other covariates were recorded for each patient. One of the aim of this study was to statistically determine whether, or not, the mutational status of p53 had an influence on the response to chemotherapy. A Cox proportional hazards model (see Cox (1972a)) was fitted to assess the simultaneous effect of the covariates. On one hand, patients with wild type p53 gene do not benefit from the addition of oxaliplatin to FL, *ceteris paribus*. Patients with mutated p53 gene, on the other hand, had a statistically significant lower risk to experience a recurrence of the disease, if they have received the adjuvant FOLFOX. Those medical research results are given in Zaanani et al. (2008).

This might seem as “yet another Cox model for survival data”, but two details, which, in a certain way, gave birth to this project, have to be highlighted. First of all, in this study, the  $p$ -value for the Cox proportional hazards null hypothesis was 0.13. The Cox model has been accepted, but without a great trust in it. In front of such a  $p$ -value, a natural behavior for a statistician is to look beyond the Cox model, despite the nice properties and interpretability that the Cox model has. This is the first main problematic of this project, which is, in summary, **to investigate beyond the Cox model**. The second particularity of this statistical study is that the mutational status of a particular gene has an influence on the response to chemotherapy. The natural question is: what if the p53 gene were not suspected, in advance, to have a role in this particular process? Not less than 2000 genes might be deregulated in colon cancer tumoral cells. If Alex Duval and his team have asked which gene (or combination) is a **prognostic** factor for the disease-free time, i.e. responsible for a lower or greater risk of a disease recurrence, there would have been very few statistical answers, apart from a Cox model, see Tibshirani (1997). Statistical theory for **high-dimensional covariates in survival analysis** has indeed been only very partly investigated. This is be our second main problematic.

## 1.2 Constitution of the team

In this paragraph, we describe how the team has grown around these two main problematics. All the people named in this section are member of the team, see Table 1 from Section 4.2 below. In Figure 1 below we show the scientific collaborations already existing in the team, together with the corresponding papers.

The Cox model is a semiparametric modelization of the intensity of a counting process in presence of covariates. Since its introduction by Cox (1972a), it became very popular in applications, because of the nice interpretability of the hazard ratios and the good asymptotic behavior of the estimator in this model. However, this model, as it is semiparametric, may not fit well some dataset (see the example above). This led us to consider, as a first step, a purely nonparametric modelization of the intensity. We construct in Comte et al. (2008) and Gaïffas and Guilloux (2008) new estimation techniques based on recent nonparametric statistics methods, such as penalized model selection and aggregation. This is a joint work of F. COMTE, S. GAÏFFAS and A. GUILLOUX (principal investigator) for the first paper, and S. GAÏFFAS and A. GUILLOUX for the second. In Gaïffas and Guilloux (2008), we adapt tools of learning theory to inference for counting processes. These tools are studied by S. GAÏFFAS and G. LECUÉ in the regression model, see Gaïffas and Lecué (2008). G. BIAU is an expert of learning theory, and accepted to share some of his expertise with the team.

The fully nonparametric modelization has a major drawback: the quality of estimation deteriorates with the dimension of the covariates. This is the so-called “curse of dimensionality”. A solution to reduce the dimension of the covariates is to consider a semiparametric model, such as the single index model (SIM),

which happens to be a generalization of the Cox model. Adaptive estimation in this model is considered in [Gaïffas and Guilloux \(2008\)](#), again with tools coming from learning theory, whereas O. BOUAZIZ, O. LOPEZ study M-estimation techniques for this model in survival analysis, see [Lopez \(2008\)](#) and [Bouaziz and Lopez \(2008\)](#). For more common models (such as regression), semiparametric techniques are known to provide good results when the dimension of the covariates is of a “reasonable” size, i.e. much smaller than the sample size.

In medical applications (see the example above), the dimension of the covariates may be very high. The statistical inference in high dimension involves different tools, such as multi-tests,  $\ell_1$  penalization, etc. These methods are barely only studied in the linear regression model, and in particular not in the counting process setting, excepted for [Tibshirani \(1997\)](#). This is the reason why we want to develop such techniques for these models with the help of E. ROQUAIN and F. VILLERS, see for instance the Ph.D by [Roquain \(2007\)](#) and [Villers \(2007\)](#).

In some medical applications, the modelization and/or the object to be estimated may be more complicated than previously described. The most common examples are multi-states medical history (see for instance the Ph.D of P. SAINT-PIERRE) or recurrent events (see [Geffray and Guilloux \(2008a\)](#)). A. DUVAL (director of CRSA E13 (INSERM, UPMC), see Section 1.1), is also a member of the project.

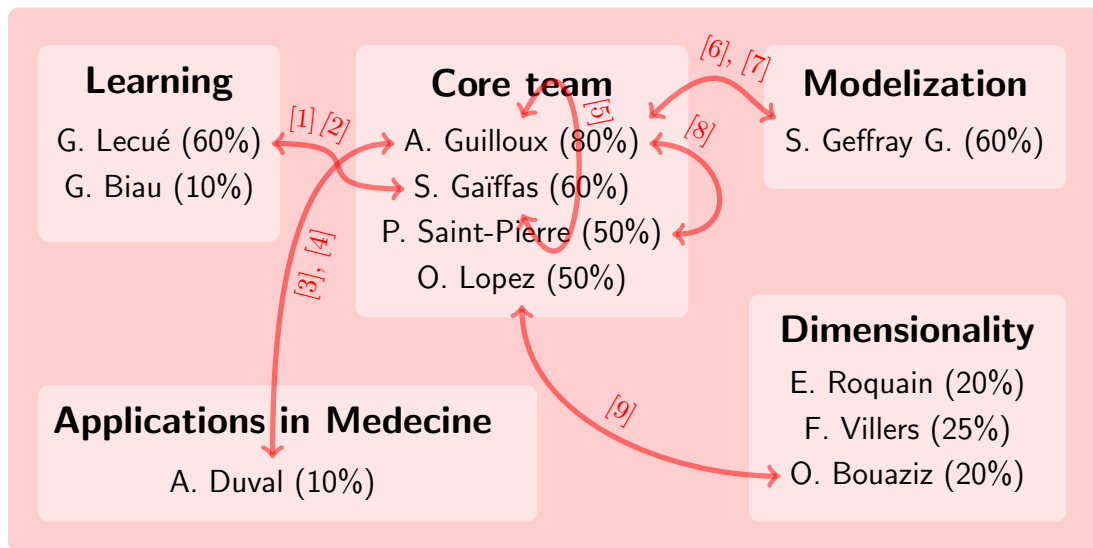


Figure 1: Already existing collaborations with the corresponding bibliography: [1]=[Gaïffas and Lecué \(2007\)](#), [2]=[Gaïffas and Lecué \(2008\)](#), [3]=[Zaanan et al. \(2008\)](#), [4]=[El-Bchiri et al. \(2008\)](#), [5]=[Gaïffas and Guilloux \(2008\)](#), [6]=[Geffray and Guilloux \(2005\)](#), [7]=[Geffray and Guilloux \(2008b\)](#), [8]=[Guilloux and Saint-Pierre \(2008\)](#), [9]=[Bouaziz and Lopez \(2008\)](#).

## 2 Scientific and technical description

### 2.1 Background, State of Art

The example described in Section 1.1 is an example of regression for survival data. Since the late 70’s, survival data are often seen as a particular case of counting processes, see in particular [Aalen \(1978\)](#). Hence, the mathematical background of the project is related to counting processes. Let us describe it here. Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and  $(\mathcal{F}_t)_{t \geq 0}$  a filtration satisfying the “usual conditions” (see [Jacod and Shiryaev \(1987\)](#)). Let  $N$  be a marker-dependent counting process, or equivalently a marker-dependent point

process on  $\mathbb{R}_+$ , with compensator  $\Lambda$  with respect to  $(\mathcal{F}_t)_{t \geq 0}$ , such that

$$N - \Lambda = M, \tag{1}$$

where  $M$  is a  $(\mathcal{F}_t)_{t \geq 0}$ -local martingale. The most popular model for inference for counting processes is to assume that  $N$  satisfies the Aalen multiplicative intensity model in the sense that:

$$\Lambda(X, t) = \int_0^t \alpha(X, z) Y(z) dz, \text{ for all } t \geq 0 \tag{2}$$

where  $X$  is a vector of covariates in  $\mathbb{R}^d$  which is  $\mathcal{F}_0$ -measurable, the process  $Y$  is nonnegative and predictable and  $\alpha$  is an unknown deterministic function called intensity. The estimation is carried on the basis of the observation of a  $n$ -sample  $(X_i, N^i(z), Y^i(z), z \leq \tau)$  for  $i = 1, \dots, n$ . In the paragraph below, we will present the state of the art when the model satisfies Equation 2, but also in other situations.

We would like to emphasize the parallel and differences with the standard regression model. The latter can be described as  $Y - f(X) = \varepsilon$ , where, in general,  $Y$  is in  $\mathbb{R}$ ,  $X$  is a vector of covariates in  $\mathbb{R}^d$ ,  $\varepsilon$  is such that  $\mathbb{E}(\varepsilon|X) = 0$  and  $f$  is an unknown regression function. The parallel with Equation 1 is obvious, the processes  $N$ ,  $\Lambda$ , and  $M$  playing resp. the roles of  $Y$ ,  $f(X)$ , and the error  $\varepsilon$ . The first difference is structural: model (1) has a “temporal direction”. The second difference is that the regression model has been the focus of a lot of researchers in statistics, until now, while model (1) did not benefit, apart from some rare exceptions, from the recent advances in statistical theory, e.g. learning theory or statistics in high dimension.

### 2.1.1 State of the Art on counting processes

There are many examples, crucial in practice, which fulfill the model model described in Equations 1 and 2. For the seek of conciseness, we restrict our presentation to the three following ones:

- **Censored data:** Let  $T$  be a nonnegative random variable (r.v.) and  $X$  a vector of covariates in  $\mathbb{R}^d$ , with respective cumulative distribution functions (c.d.f.)  $F_T$  and  $F_X$ . We consider in addition that the r.v.  $Y$  can be censored. We introduce the nonnegative r.v.  $C$ , with c.d.f.  $G$ , such that the observable r.v. are  $Z = T \wedge C$ ,  $\delta = I(T \leq C)$  and  $X$ . In this case, the processes to consider (see e.g. Andersen et al. (1993)) are given, for  $i = 1, \dots, n$  and  $z \geq 0$ , by:  $N^i(z) = I(Z_i \leq z, \delta_i = 1)$  and  $Y^i : Y^i(z) = I(Z_i \geq z)$ . The unknown intensity function  $\alpha$  to be estimated is, in this setting, the conditional hazard rate of the r.v.  $T$  given  $X = x$  defined, for all  $z > 0$  by:  $\alpha(x, z) = f_{T|X}(x, z)/(1 - F_{T|X}(x, z))$ , where  $f_{T|X}$  and  $F_{T|X}$  are respectively the conditional probability density function (p.d.f.) and the conditional c.d.f. of  $Y$  given  $X$ .
- **Nonhomogeneous Poisson process:** Let  $\eta^i$ , for  $i = 1, \dots, n$ , be a Cox process on  $\mathbb{R}_+$  with random mean-measure  $\Lambda^i$  given by  $\Lambda^i(t) = \int_0^t \alpha(X_i, z) dz$ . In this context the predictable process  $Y$  constantly equals 1.
- **Markov process:** Consider a  $n$ -sample of nonhomogeneous Markov processes  $P^1, \dots, P^n$  with finite state space  $\{1, \dots, k\}$  and denote by  $\alpha_{jl}$  the transition intensity from state  $j$  to state  $l$ . For individual  $i$  with covariate  $X_i$ , the r.v.  $N_{jl}^i(t)$  counts the number of observed direct transitions from  $j$  to  $l$  before time  $t$  (we allow the possibility of right-censoring for example). Conditionally on the initial state, the counting process  $N_{jl}^i$  verifies the Aalen multiplicative intensity model:  $N_{jl}^i(t) = \int_0^t \alpha_{jl}(X_i, z) Y_j^i(z) dz + M^i(t)$  for all  $t \geq 0$ , where  $Y_j^i(t) = I\{P^i(t-) = j\}$  for all  $t \geq 0$ , see Andersen et al. (1993) or Jacobsen (1982).

The model described in Equations 1 and 2 **when no covariates are present** has received a lot of attention, see Andersen et al. (1993) for a review on the estimation of  $\int \alpha$  by “empirical estimation” or  $\alpha$  by kernel methods. Since then, Stute (1994) and Csörgő (1996) can be seen has the most achieved developments on the estimation of  $\int \alpha$  for censored data. Whereas Antoniadis et al. (1999), Brunel and Comte (2005) introduced new estimation techniques, as wavelets and model selection, again for censored data. The nonparametric

estimation of the intensity of Poisson processes without covariates has been considered in several papers. We refer to [Reynaud-Bouret \(2003\)](#) and [Baraud and Birge \(2008\)](#) for the adaptive estimation of the intensity of nonhomogeneous Poisson processes in general spaces. For the general model described in Equations 1 and 2, [Ramlau-Hansen \(1983\)](#) proposed a kernel-type estimator, [Grégoire \(1993\)](#) studied least squares cross-validation. [Reynaud-Bouret \(2006\)](#) studied model selection.

**When covariates are present**, regression for censored data has received a lot of attention. Nonparametric estimation of the hazard rate in presence of covariates was initiated by [Beran \(1981\)](#). [Stute \(1996\)](#), [Dabrowska \(1987\)](#), [McKeague and Utikal \(1990\)](#) and [Li and Doss \(1995\)](#) extended his results. Many authors have considered semiparametric estimation of the hazard rate, beginning with [Cox \(1972b\)](#), see [Andersen et al. \(1993\)](#) for a review of the enormous literature on semiparametric models. We refer to [Huang \(1999\)](#) and [Linton et al. \(2003\)](#) for some recent developments. As far as we know, adaptive nonparametric estimation for censored data in presence of covariates has only been considered in [Brunel et al. \(2008\)](#), who constructed an optimal adaptive estimator of the conditional density. In this model, one can consider the single-index semiparametric model given by:

$$\alpha_0(t, x) = \beta_0(t, v_0^\top x). \quad (3)$$

In such a model,  $\beta_0$  is called *link* function and  $v_0$  is called *index*. On single-index models (mainly in regression) and the corresponding estimation problems (estimation of the link function, estimation of the index), see [Hristache et al. \(2001\)](#), [Delecroix et al. \(2003\)](#), [Xia and Härdle \(2006\)](#), [Delecroix et al. \(2006\)](#), [Geenens and Delecroix \(2005\)](#), [Gaïffas and Lecué \(2007\)](#) among many others. In the problem of estimating the intensity of a counting process in presence of covariates, two of the most famous and applied in practice models are special cases of the single-index model, as described in Equation (3):

- the Cox model (see [Cox \(1972b\)](#)), where there exists an unknown function  $\beta_0$  such that:

$$\alpha_0(t, x) = \beta_0(t) \exp(v_0^\top x), \quad (4)$$

- the Aalen model (see [Aalen \(1980\)](#)), which can be written as:

$$\alpha_0(t, x) = \beta_0(t) + v_0^\top x. \quad (5)$$

### 2.1.2 Background, Mathematical methods involved

In this section, we give short descriptions of the mathematical fields involved in this project. We shall provide a small number of references, as the goal is just to give the context. More details can be found in Section 3.1 below. The mathematics involved here are, of course, the ones usually involved in mathematical statistics: probability theory of course, more specifically stochastic calculus and concentration inequalities, but also approximation theory, which is naturally useful in nonparametric statistics, and learning theory, which provides a framework for the theoretical properties of algorithms. Of course, those fields are far from having disjoint intersections. The benefit of using them at the same time is fruitful, an example being for instance the construction of adaptive estimators, that share good properties from the theoretical and applied point of view. Note that we already obtained results involving this kind of mathematics, see [Comte et al. \(2008\)](#) and [Gaïffas and Guilloux \(2008\)](#).

### 2.1.3 Concentration inequalities

Concentration inequalities, namely a control on the probability of the deviation of a random variable from its expectation, is at the core of modern non-asymptotic statistics, see for instance [Massart \(2007\)](#). In learning theory, such inequalities are at the heart of various problems, and allow to derive new efficient algorithms, see, among others, [Cucker and Smale \(2002\)](#). For model selection problems, an important impetus (see [Massart \(2007\)](#)) was provided by the works of Michel Talagrand, see [Talagrand \(1996\)](#) and Michel Ledoux, see [Ledoux and Talagrand \(1991\)](#); [Ledoux \(2001\)](#). Roughly speaking, such inequalities are useful to give a non-asymptotic (fixed sample size) theoretical assessment of an estimation procedure.

#### 2.1.4 Stochastic calculus

The nature of the statistical models considered in this project involves stochastic calculus, since these models are continuously time-dependent. The corresponding mathematical objects are stochastic process, more precisely for our concerns point processes. For instance, the proof of concentration inequalities related to such objects involves tools from martingale theory, see for instance [Liptser and Shiriyayev \(1989\)](#), and examples of such results for statistics can be found in [Reynaud-Bouret \(2003, 2006\)](#) and [van de Geer \(1995\)](#), see also [Comte et al. \(2008\)](#).

#### 2.1.5 Statistical learning theory

At the beginning (e.g. [Vapnik \(1998, 2000\)](#)), learning theory was the problem of predicting a discrete output  $Y \in \{0, 1\}$  associated to some covariates  $X \in \mathcal{X}$  (where  $\mathcal{X}$  can be a very general space) from observations. Support vector machine (see [Steinwart and Scovel \(2007\)](#)) has been widely studied within the last decade. Nevertheless, we believe that a right calibration of the penalty (see [S. Mendelson \(2008\)](#)) combined with some  $\ell_1$ -based selection procedure could lead to improved classification procedures. These problems seem to be very hard to handle from both algorithmic and theoretical point of view.

On the other hand, several problems in statistics and machine learning can be resumed to the problem of construction of a procedure which is nearly as good as the best among the ones from a fixed dictionary. This is often called the ability to “mimic the oracle”, and this property is, from the theoretical point of view, proved by so-called oracle inequalities. Procedures such as empirical risk minimization, penalized model selection and aggregation are at the core of the project. As mentioned above, we already obtained results for such procedures in [Comte et al. \(2008\)](#) and [Gaïffas and Guilloux \(2008\)](#).

#### 2.1.6 High dimension, detection of non-zero components

**LASSO** In many statistical applications (the prediction of some genetic illness is an example), the number  $d$  of covariates  $X = (X_1, \dots, X_d)$  (for instance genes) can be much larger than the number of observations  $n$ . Fortunately, we can hope that most of these variables have no impact on the output, say, a variable  $Y \in \mathbb{R}$ , which is the variable that we are interested to predict. Most papers consider the case where  $Y$  is a linear function of  $X \in \mathbb{R}^d$  up to a noise  $\varepsilon$ :  $Y = \beta^\top X + \varepsilon$  (linear regression model). Saying that most of the covariates of  $X$  have no impact on the output means that most of the coordinates of  $\beta$  are equal to zero. This is the classical sparsity assumption that is at the core of many recent research papers (cf, for instance, [Candes and Tao \(2007\)](#), [Meinshausen and Bühlmann \(2006\)](#), [van de Geer \(2008\)](#), [Bickel et al. \(2008\)](#)). Classical Ordinary Least Squares estimation fails in the setup where  $d$  much larger than  $n$ . Nevertheless, several procedures have been introduced to tackle this high dimensional problem of estimation/selection: LASSO, grouped LASSO, elastic-net, adaptive LASSO, Dantzig selector, two step (forward and backward) procedures, etc. Many of them have been studied theoretically and empirically, but mainly in the linear regression model. Some exceptions can be found in [Tibshirani \(1997\)](#) where the Cox model is considered or [Bickel et al. \(2008\)](#) and [Meier et al. \(2008\)](#) where general dictionary are considered. However, even in the linear regression model, it remains a lot of work to do. In particular, the colinearity problem of the covariates induced by the high-dimensionality.

**Multiple tests approach** Current statistical problems, like the detection of differentially expressed genes in microarray experiments, or more generally the detection of significant variables among a large number of variables, involve the simultaneous tests of a huge number of null hypotheses. Testing each single hypothesis at some level  $\alpha$  produces the overall control to be multiplied by the number of hypotheses. Therefore, a multiple testing correction is needed.

The probably most used multiple testing correction is the Bonferroni correction, which take for each single test the level  $\alpha/m$  where  $\alpha$  is the wanted overall control and  $m$  the number of hypotheses to test. This procedure controls the family-wise error rate (FWER), that is the probability to have at least one wrongly rejected hypothesis. This thresholding is quite conservative, because it does not make a lot of rejection. A

more permissive error rate have been proposed by [Benjamini and Hochberg \(1995\)](#), called the false discovery rate (FDR) and defined as the average proportion of wrongly rejected hypothesis among all the rejected hypothesis. The FDR is very useful in practice, because it is adaptive to the number of rejections: for a given number of rejections, it tolerates a specific proportion of mistakes (on average). This generally results in more detected variables.

**Estimation approach** The detection of non-zero components of a high-dimensional vector can be tackle following two approaches: multiple testing or estimation method based on a model choice procedure. In the estimation approach, the aim is to estimate the number of zero components as well as their positions. More generally, model selection procedures are designed to estimate the mean of a vector. Multiple testing aims at controlling the False Discovery Rate while in model selection, a penalty function is choose to minimize some risk of the selected estimator.

The first heuristics in the domain are due to [Mallows \(1973\)](#) for the estimation of the mean in homoscedastic Gaussian regression with known variance. In Gaussian regression framework with common known variance [Barron et al. \(1999\)](#), and [Birgé and Massart \(2001\)](#) designed a model selection procedure to estimate the mean of the vector, and provided non-asymptotic upper-bound for the quadratic risk of the selected estimator.

## 2.2 Originality and novelty of the proposal

The whys and wherefores of this project can be resumed by the two following points.

- Our belief is that inference for counting processes has not **yet fully benefited from the recent advances in statistical theory**, such as the ones obtained in statistical learning theory or for statistics in high dimension, for instance. So, our aim is to obtain similar results for counting processes in order to approach (or even reach) the state of the art in regression for example. Indeed, a striking fact is that model (1) is of **crucial importance in many fields of application**, especially in medicine (more on that below), economics or actuarial sciences, and that, in these fields, there are available datasets that cannot be statistically studied because of the lack of methods adapted to model (1). High dimensional data is the first example that comes into mind. This is why we believe that the development of “state of the art” methods for this model is critical, and this is what this project is about. A lot of statistical mathematics is involved (see Section 3.1). This will lead to a lot of publications in statistics (in Section 3.1 below, each subtask should lead to at least one publication).
- Let us recall that the idea that gave birth to this project comes from collaborations between A. GUILLOUX (principal investigator) and CRSA E13 (INSERM, UPMC) research team, directed by A. DUVAL. This research team studies genomic and transcriptomic profiling of colorectal cancers. Colorectal cancers consist in a category of tumours occurring frequently in human. They constitute the **second cause of cancer mortality in industrialized countries**. The knowledge about the oncogenic processes and their consequences on gene expression remains to be clarified. The CRSA E13 research team constituted datasets, that consist of a large panel of tumor samples, for which they conduct an analysis of genomic (see Section 1.1) and expression alterations. This research team is supported by the Tumor ID-card program (CIT) from LNCC (**Ligue Nationale contre le Cancer**) and INCa (**Institut National du Cancer**). An example of dataset is given in Figure 2 below, which shows the transcriptomic profile (expression alterations) for 311 genes of colorectal tumoral cells for 120 patients, for which disease-free survival times (and status) and other clinical or biological variables are observed. We believe that theoretical and empirical developments from the first point will provide **methods and computer tools** to investigate the influence of the expression alteration of each of these 311 genes on the risk of a disease recurrence. This question is of great importance in cancerology, see in particular [Del Rio et al. \(2007\)](#).

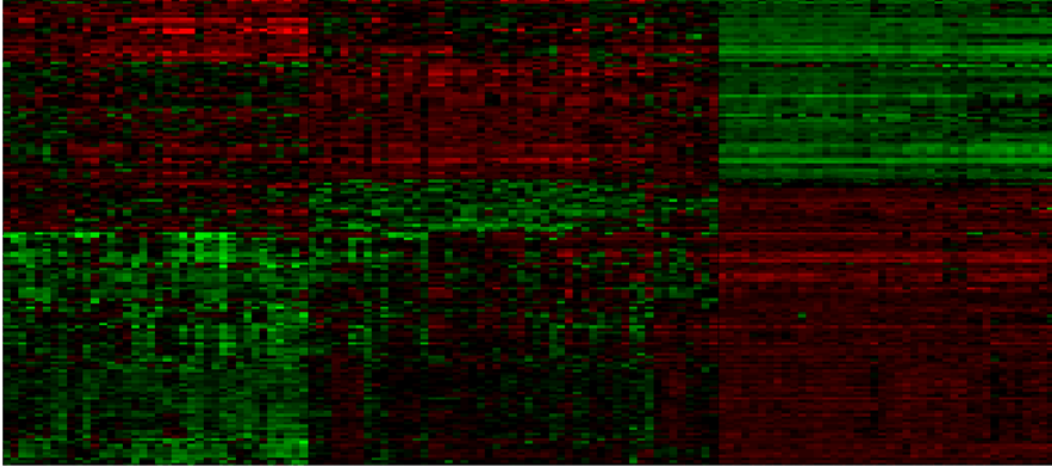


Figure 2: Transcriptomic profile (expression alterations) for 311 genes of colorectal tumoral cells for 120 patients

### 3 Scientific and technical programme, project management

#### 3.1 Detailed description of the work

Let us give below a detailed description of the project, task by task. Most of these tasks are specific to the model of marker-dependent counting process (see Section 2.1), which is directly related to the applications we have in view (see Section 2.2). However, some tasks are described in other models, such as the regression model. The reason is that there is still a lot of work to do, even in the more standard models, for the procedures we need in this project (in particular in the high dimension setting, see for instance Tasks 4, 6 and 11). For each task below, we give the members obviously related to it, but, of course, other members may participate, hence the question marks.

##### 3.1.1 Task 1: Model selection for counting processes

In Comte et al. (2008), we introduce a new empirical risk for the estimation of the intensity  $\alpha$  in the Aalen model, see Equations (1) and (2) above. This empirical risk writes

$$R_n(\alpha) := \frac{1}{n} \sum_{i=1}^n \int_0^1 \alpha(t, X_i)^2 Y^i(t) dt - \frac{2}{n} \sum_{i=1}^n \int_0^1 \alpha(t, X_i) dN^i(t). \quad (6)$$

It measures the goodness-of-fit of  $\alpha$  to the data from the sample

$$D_n := \{(X_i, N^i(t), Y^i(t)) : t \in [0, 1], 1 \leq i \leq n\}. \quad (7)$$

The idea is the following: the expected risk

$$R(\alpha) := E[R_n(\alpha)] = E \left[ \int_0^1 \alpha(t, X)^2 Y(t) dt - 2 \int_0^1 \alpha(t, X) dN(t) \right]$$

satisfies, in view of (1):

$$\begin{aligned} R(\alpha) &= E \left[ \int_0^1 (\alpha(t, X)^2 - 2\alpha(t, X)\alpha_0(t, X)) Y(t) dt \right] \\ &= \|\alpha - \alpha_0\|^2 - \|\alpha_0\|^2, \end{aligned} \quad (8)$$

so that the excess risk  $R(\alpha) - R(\alpha_0)$  is equal to  $\|\alpha - \alpha_0\|^2$ , where  $\|\cdot\|$  stands for the weighted norm

$$\|\alpha\|^2 := \int_{\mathbb{R}^d} \int_0^1 \alpha(t, x)^2 E[Y(t)|X = x] dt P_X(dx). \quad (9)$$

As a consequence,  $\alpha_0$  minimizes  $R(\cdot)$ , and a natural way to recover  $\alpha_0$  is to minimize  $R_n(\cdot)$ , or a penalized version of  $R_n(\cdot)$ .

We introduce a collection  $\{S_m, m \in \mathcal{M}_n\}$  of projection spaces:  $S_m$  is called a model and  $\mathcal{M}_n$  is a set of multi-indexes. We define

$$\hat{\alpha}_m = \operatorname{argmin}_{h \in S_m} R_n(h),$$

with a slight modification. Model selection leads to select the relevant space  $S_m$  via the penalized criterion:

$$\hat{m} = \operatorname{argmin}_{m \in \mathcal{M}_n} \{R_n(\hat{\alpha}_m) + \operatorname{pen}(m)\},$$

where  $\operatorname{pen}(m)$  relies on the dimension of  $S_m$ . Then, we prove that this procedure achieves the optimal rate of convergence over anisotropic Besov spaces, a lower bound in this model is given over such spaces.

**Subtask 1: Semiparametric models** (A. GUILLOUX, ?). The intensity  $\alpha_0$  is often supposed to belong to a semiparametric model. There are several examples of such models, intensively used in practice. We restrict our presentation to the two most famous ones:

- the Cox model (see Cox (1972b)), where there exists an unknown function  $\beta_0$  such that:

$$\alpha_0(t, x) = \beta_0(t) \exp(v_0^\top x),$$

- the Aalen model (see Aalen (1980)), which can be written as:

$$\alpha_0(t, x) = \beta_0(t) + v_0^\top x,$$

where in both models,  $v_0 \in \mathbb{R}^d$  is an unknown parameter and  $\beta_0(\cdot)$  is an unknown function. The convergence in probability and the  $\sqrt{n}$ -convergence in distribution of estimators  $\hat{v}$  in these two models has been established, see Andersen et al. (1993) for a detailed review, and estimators of  $\int \beta_0$  or  $\beta_0$  by kernel methods have been studied.

We aim to propose new estimators of  $\beta_0(\cdot)$  in this models. Towards that end, we adapt our empirical risk to the semiparametric setup considered, for example for the Cox model, we can set:

$$R_n^{\text{Cox}}(\beta, v) := \frac{1}{n} \sum_{i=1}^n \int_0^1 \beta(t)^2 \exp(2v^\top X) Y^i(t) dt - \frac{2}{n} \sum_{i=1}^n \int_0^1 \beta(t) \exp(v^\top X) dN^i(t).$$

Estimators of  $\alpha_0$  and  $v_0$  can be defined as minimizers of this modified empirical risk. We hope that model selection will lead us to state optimal upper bounds for adaptive estimators of  $\alpha_0$  constructed in the Cox, Aalen and other semiparametric models.

### 3.1.2 Task 2: Learning for counting processes

**Current work (almost completed)** (S. GAÏFFAS, A. GUILLOUX). As mentioned in the previous paragraph, we introduced in Comte et al. (2008) a new empirical risk, see Equation (6) for the estimation of the intensity  $\alpha$  in the Aalen model, see Equations (1) and (2) above. A current work, which is almost completed, is to use it in order to construct estimators based on empirical risk minimization (ERM) and aggregation, two procedures which are of importance in statistical learning theory: see for instance Vapnik (2000), Nemirovski (2000), Catoni (2004). As detailed above,  $\alpha_0$  minimizes  $R(\cdot)$ , and a natural way to recover  $\alpha_0$  is to minimize

$R_n(\cdot)$ , or a penalized version of  $R_n(\cdot)$ . Thus, considering a class of functions  $A$  satisfying some complexity assumption (in terms of entropy or entropy with bracketing, see for instance [Barron et al. \(1999\)](#), [Birgé and Massart \(1998\)](#), [Massart \(2007\)](#)), we consider an ERM

$$\bar{\alpha}_n \in \operatorname{argmin}_{\alpha \in A} R_n(\alpha),$$

and we prove risk bounds for this ERM using arguments coming from empirical process theory, as in [Massart and Nédélec \(2006\)](#), [van de Geer \(2007, 2000\)](#) or [Massart \(2007\)](#), among others. One of the challenges in this work is that for the empirical process related of this model, namely

$$Z_n(\alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^1 \alpha(t, X_i) dM^i(t), \quad (10)$$

(where  $M^i$  are the martingale innovations coming from the Doob-Meyer decomposition, see (1)) the Talagrand's inequality (see [Talagrand \(1996\)](#), [Bousquet \(2002\)](#)) is not directly helpful, since this empirical process does not write as the sum of bounded functions of independent variables. Indeed,  $\int \alpha(t, X_i) dM^i(t)$  is not, in general, almost surely bounded, even if  $\alpha$  is bounded.

The second problem that we consider is the agnostic learning, which was introduced by [Haussler \(1992\)](#), [Kearns et al. \(1994\)](#). In this problem, we fix a set of arbitrary functions  $A$  (usually called *dictionary*) and, without any assumption on  $\alpha$  (to be estimated), we want to construct (from the data  $D_n$ ) a procedure  $\hat{\alpha}_n$  which has a risk as close as possible to the smallest risk over  $A$ . Namely, we want to obtain *oracle inequalities*, of the form

$$E \|\hat{\alpha}_n - \alpha\|^2 \leq C \min_{a \in A} \|a - \alpha\|^2 + \phi(n, A),$$

where  $C \geq 1$  and  $\phi(n, A)$  is called the *residue*, which is the quantity that we want to be small as  $n$  increases. When  $A$  is of finite cardinality  $M$ , the agnostic problem is called *aggregation problem* and the residue  $\phi(n, A) = \phi(n, M)$  is called *rate of aggregation*. In this work we prove such an oracle inequality, which shows that the usual rate of aggregation  $\phi(n, M) = (\log M)/n$  (see for instance [Tsybakov \(2003\)](#)) can be achieved using aggregation with exponential weights, which is a popular aggregation algorithm: see for instance [Catoni \(2004\)](#), [Leung and Barron \(2006\)](#), [Juditsky et al. \(2005a\)](#), [Juditsky et al. \(2005b\)](#), [Yang \(2000\)](#), [Yang \(2004\)](#). If  $A = \{\alpha_1, \dots, \alpha_n\}$ , this aggregate is given by

$$\hat{\alpha}_n := \sum_{j=1}^M \theta(\alpha_j) \alpha_j, \quad (11)$$

where the weight of  $\alpha \in A$  is given by

$$\theta(\alpha) := \frac{\exp(-R_n(\alpha)/T)}{\sum_{j=1}^M \exp(-R_n(\alpha_j)/T)}, \quad (12)$$

where  $T > 0$  is the so-called *temperature* parameter and where  $R_n$  is the empirical risk given by (6). Then, we can use ERM and the aggregation procedures to construct adaptive estimator. Such estimator are proved to be adaptive to the smoothness of  $\alpha$ , but also to the structure of  $\alpha$ , whenever  $\alpha$  satisfies some semiparametric model, such as single-index, when the dictionary  $A$  is chosen using appropriate weak estimators.

**Subtask 2: Implementation** (S. GAÏFFAS, M2 STUDENT, ?). The subtasks proposed above and below contains new statistical techniques that are not present in statistical softwares. These can be quite challenging to code. This would require, for instance, the use of minimization algorithms for convex and non-convex problems, linear algebra in high dimension and quadratic programming, among others techniques (see Task 4 for instance). Many tools are present in standard statistical softwares, but not for model (2), excepted for some very particular cases. For instance in the Cox model (see (4)), the LASSO is available (see [Tibshirani \(1997\)](#) for instance). We think that this subtask should provide a nice research subject for a M2 student. The finality of this subtask is to provide tools that could be used by practioners, for medical applications for instance, see Section 2.2.

**Subtask 3: Go beyond the Cox model, learning** (S. GAÏFFAS, A. GUILLOUX, ?). Let us assume that in Task 1, we obtain finite sample results (such as concentration inequalities, convergence rates for the mean squared error) of several estimators, in particular the estimator of the parameter  $v_0$  in the Cox model, see Equation (4). Then, we can compute several estimators containing the ones working in semiparametric models, see Equations (3), (4) and (5) and other estimators working in the purely nonparametric setting. This should constitute a whole dictionary of so-called *weak estimators*, each of them being adapted to a particular model, that we can mix using an aggregation algorithm, for instance the one given in Equations (11) and (12). This could very well provide an answer to the first problematic of the project (see Section 1.1), which is *investigate beyond the Cox model*. Indeed, if data is not perfectly explained by a Cox model (as it is the case in Section 1.1), but if, on the other hand, we don't want to throw away this model, the aggregation approach seems to be a natural answer. This idea of "structure" adaptation by mixing has been studied in density and regression model, see Yang (2001), Yang (2000) and Yang (2004). As far as we know, this has not been investigated in models such as the one given in Equations (1) and (2), despite the fact that this is important in such models, even for applied problems (see Section 2.2).

**Subtask 4: Sparse aggregation with exponential weights** (S. GAÏFFAS, G. LECUÉ, ?). Assume that we have  $n$  observations  $D_n$ , that we write for short  $(Z_i)_{1 \leq i \leq n}$  (this could be data from a sample (7), or data for a regression of a density model). Let us consider again the aggregation problem: given is a finite class of functions  $F = \{f_1, \dots, f_M\}$ , we want to construct a procedure  $\tilde{f}_n$  which has a risk as close as possible to the minimal risk over  $F$ . We define the risk of a function  $f$  and the risk of an estimator  $\tilde{f}_n$  respectively by

$$R(f) := EQ(Z, f) \text{ and } R(\tilde{f}_n) := E[Q(Z, \tilde{f}_n) | D_n].$$

A classical approach is the aggregation with exponential weights algorithm defined by

$$\tilde{f}_n = \sum_{f \in F} \theta(f) f \text{ where } \theta(f) = \frac{\exp(-nR_n(f)/T)}{\sum_{g \in F} \exp(-nR_n(g)/T)},$$

where  $R_n(f) = \frac{1}{n} \sum_{i=1}^n Q(Z_i, f)$  is the empirical risk and  $T > 0$  the temperature parameter. Two important questions remain unsolved concerning this procedure:

1. Is  $\tilde{f}_n$  an optimal aggregation procedure?
2. How has to be chosen the temperature parameter?

We believe (see Section 3.1.6 below) that there is hope to solve these problems. Nevertheless, this procedure has a main flaw: the weights cannot be equal to zero. Thus, even if some of the elements in  $F$  are not relevant to mimic the oracle, they still appear in the final aggregate. We would like to construct a procedure which can "adapt" to the real complexity of  $F$  (in its form and in its residue). We know that the weights  $\theta = (\theta(f) : f \in F)$  of  $\tilde{f}_n$  are solution of the minimization problem

$$\theta \in \operatorname{argmin}_{\theta \in \Theta_M} \left\{ \bar{R}_n(\theta) + \frac{T}{n} \operatorname{ent}(w, u) \right\} \quad (13)$$

where  $\Theta_M = \{(\theta_1, \dots, \theta_M) : \theta_j \geq 0 \text{ and } \sum_{j=1}^M \theta_j = 1\}$ ,  $\bar{R}_n(\theta) = \sum_{j=1}^M \theta_j R_n(f_j)$  is the linearized version of the empirical risk  $R_n(\cdot)$  and  $\operatorname{ent}(w, u) = \sum_{j=1}^M w_j \log(w_j/u_j)$  (where  $u = (1/M, \dots, 1/M)$  are the uniform weights) is the entropy of  $w$  w.r.t.  $u$ .

An empirical study (see Gaïffas and Lecué (2007)) proved that the weights associated to non-relevant elements of  $F$  have very small weights. We would like to use a kind of thresholding method to get an exact zero coefficient for non-relevant elements in  $F$ . For that we introduce a kind of thresholded version of the aggregate with exponential weights. We propose two different approach:

1. Use the pre-selection step on the dictionary introduced in [Lecué and Mendelson \(2008\)](#). Assume that we have  $2n$  observations. Consider two subsamples  $D_n^1 = \{Z_1, \dots, Z_n\}$  and  $D_n^2 = \{Z_{n+1}, \dots, Z_{2n}\}$ . We define the following set of almost empirical minimizers:

$$\hat{F}_1 := \{f \in F : R_n(f) \leq R_n(\hat{f}_n^*) + C \min(\|\hat{f}_n^* - f\|_n \alpha, \alpha^2)\},$$

where  $\alpha = \sqrt{(x + \log M)/n}$ ,  $\hat{f}_n^* \in \operatorname{argmin}_{f \in F} R_n(f)$ ,  $R_n(\cdot)$  is the empirical risk over  $D_n^1$  and  $\|\cdot\|_n$  is the empirical  $L^2$ -norm. Now, introduce the empirical risk over  $D_n^2$ :  $R_n^{(2)}(f) = \frac{1}{n} \sum_{i=1}^{2n} Q(Z_i, f)$  and we construct the exponential weights using the set  $D_n^2$  over the new dictionary  $\hat{F}_1$ :

$$\bar{f}_n = \sum_{f \in \hat{F}_1} \bar{\theta}^{(2)}(f) f \text{ where } \bar{\theta}^{(2)}(f) = \frac{\exp(-nR_n^{(2)}(f)/T)}{\sum_{g \in \hat{F}_1} \exp(-nR_n^{(2)}(g)/T)}.$$

Remark that, thanks to the pre-selection step, non-relevant element of  $F$  are likely to be associated with a zero coefficient in the final aggregate. The second step algorithm (minimization of the linearized empirical risk penalized by the entropy) may be analyzed the same way we want to treat the aggregation with exponential weights algorithm (cf. tools and technics from Section 3.1.6). Moreover, using some empirical  $L_\infty$  technics (cf. [S. Mendelson \(2008\)](#) or [Mendelson et al. \(2007\)](#)), we believe to be able to consider a non-bounded dictionary.

2. We can add a  $\ell_1$ -penalization term in Equation (13). Since we want to profit of the non-differentiability of  $\theta \mapsto |\theta|_1$  at 0, we have to change the set of minimization. We thus consider the solution

$$\theta^S \in \operatorname{argmin}_{\theta \in B_1} \left\{ \bar{R}_n(\theta) + \frac{T}{n} (\lambda \operatorname{ent}(|w|, u) + (1 - \lambda)|w|_1) \right\}$$

where  $B_1$  is the unit  $\ell_1$ -ball of  $\mathbb{R}^M$ , where  $|\theta| = (|\theta_1|, \dots, |\theta_M|)$  and where  $\lambda \in (0, 1)$ . We can see the entropy term  $\operatorname{ent}(|w|, u)$  as a regularization term (in the same spirit as the  $\ell_2$ -penalization in the elastic net procedure, compare with Equation (19)).

### 3.1.3 Task 3: Dimension reduction, the single-index/multi-index approach

**Single-index for censored data (Done work).** For medical applications, it is essential to develop techniques for explaining or predicting a process related to the disease evolution from a set of covariates  $X \in \mathbb{R}^d$ . In many practical situations, the dimension  $d$  can be high. In such a framework, nonparametric regression approaches are known to behave poorly, which is often called the "curse of dimensionality". Therefore it is important to introduce dimension reduction techniques by studying semiparametric models designed for finding a good compromise between the flexibility of the model and the need to circumvent the curse of dimensionality. The single index and multi index models have been widely studied in the regression setup, see [Delecroix et al. \(2003\)](#), [Delecroix et al. \(2006\)](#), or [Hristache et al. \(2001\)](#), for instance. O. BOUAZIZ and O. LOPEZ have studied the SIM (single-index model) for censored data (see Section 2.1), see [Lopez \(2008\)](#) and [Bouaziz and Lopez \(2008\)](#). In particular, they studied estimation in the following model:

$$\exists v_0 \in V \subset \mathbb{R}^d \text{ s.a. } f(t, x) = f_{v_0}(t, v_0^\top x), \iff \exists v_0 \in V \subset \mathbb{R}^d \text{ s.a. } \alpha(t, x) = \alpha_{v_0}(t, v_0^\top x), \quad (14)$$

where  $f_v(y, u)$  denotes the conditional density of  $T$  given  $v^\top X = u$  evaluated at  $t$ . In comparison to Cox proportional hazards model (see Equation (4)), this model is more general, since it only assumes that the law of  $T$  given  $X$  depends on an unknown linear combination of the covariates, without imposing additional conditions on the conditional hazard rate.

The consistency and asymptotic normality of the estimator of the index parameter by proving its asymptotic equivalence with the (uncomputable) maximum likelihood estimator, using martingales results for counting processes and arguments of empirical processes theory.

**Subtask 5: Change-point single-index model** (O. LOPEZ, O. BOUAZIZ, S. GEFFRAY, A. GUILLOUX, ?). As it has been mentioned in the paragraph above, an interesting feature of the SIM model consists of avoiding problems arising in nonparametric inference in an high-dimensional space. On the other hand, generally the single-index assumption does not hold, and this model can only be seen as an approximation of the true purely nonparametric model. Therefore, it is important to propose semiparametric models for dimension reduction which are rich enough to correctly approximate a larger number of regression function.

This is particularly true in the context of modelling the process  $N(t)$ . In this context, the single-index assumption seems too strong for the general case, since it does not cover processes with regression structure evolving with time. A way to avoid this drawback is to replace the single-index assumption of Equation (14) by the following one:

$$\alpha(t, x) = \alpha_{v_0}(t, v_0^\top x)$$

where the index (that is the relevant direction for explaining the data) evolves with time. Then assumptions have to be made on function  $v_0(t)$ . The classical single-index assumption consider the case  $v_0(t) \equiv v_0$ . A more general setting consists of assuming  $v_0(t) = v_0 \mathbf{1}_{t \in [t_0, t_1]} + \dots + v_k \mathbf{1}_{t \in [t_k, t_{k+1}]}$ , where  $(t_0, \dots, t_k)$  are times corresponding to a change in the parameter.

Estimation of  $(\theta_0, \dots, \theta_k)$  can be easily be performed in the case where the times  $t_i$  are known, by simply studying a regular single-index model on each interval separately. But for practical applications, knowing exactly the times  $t_i$  is generally impossible. Therefore it is necessary to develop procedures that allow to detect from the data such change-points in the parameter. O. BOUAZIZ, O. LOPEZ, S. GEFFRAY and A. GUILLOUX plan to apply change-point techniques (see e.g. Csörgő (1996) and Spokoiny (2008)) to this new time-dependent single-index model.

As O. BOUAZIZ, O. LOPEZ derive the asymptotic law of an estimator of the (constant) single-index model. This knowledge of the asymptotic law of the estimator of  $\theta_0$  (under the assumption that there is no change-point) can be used to adapt asymptotic techniques that allow to test and detect the times of change. In Spokoiny (2008), non-asymptotic results were derived for detecting a change-point. These results could probably be adapted to the framework of time-dependent single-index after obtaining non-asymptotic results on the estimator of  $\theta_0$ .

**Subtask 6: Single-index and aggregation** (S. GAÏFFAS, G. LECUÉ, ?). Another approach in the single-index model (in the counting process or the regression setting) is based on aggregation. Let us consider the single-index regression model:

$$Y = f(X) + \epsilon, \text{ where } f(x) = g(v^\top x)$$

where  $Y$  is a real-valued random variable,  $g : \mathbb{R} \rightarrow \mathbb{R}$  is a measurable function,  $X$  is a random variable with values in  $\mathbb{R}^d$ ,  $v$  is a vector of  $\mathbb{R}^d$  and  $\epsilon$  is a sub-Gaussian noise. We want to recover  $f$  based on  $n$  i.i.d observations  $D_n = [(Y_i, X_i) : 1 \leq i \leq n]$  with same law as  $(Y, X)$ . In Gaïffas and Lecué (2007), in order to estimate the regression, we aggregate univariate estimators (local polynomials) based on "projected" training samples ( $m < n$  is fixed)

$$D_m(v) = [(Y_i, v^\top X_i) : 1 \leq i \leq m]$$

for every  $v \in \bar{S}_n$ , where  $\bar{S}_n$  is a grid of  $S_+^{d-1}$  (the half-unit sphere of  $\mathbb{R}^d$  for the Euclidean norm). This gives a family  $(\bar{f}_v : v \in \bar{S}_n)$  of "weak" estimators. Then, we aggregate them: we compute

$$\hat{f}_n = \sum_{v \in \bar{S}_n} \theta(v) \bar{f}_v,$$

where  $\theta(v)$  is the exponential weight computed based on the learning sample  $D_{(m)} = [(Y_i, X_i) : m+1 \leq i \leq n]$ , which is associated to the weak estimator  $\bar{f}_v$ . We show that the procedure  $\hat{f}_n$  is adaptive to the smoothness of  $g$  and to the index, and that it achieves the minimax rate of convergence  $n^{-\beta/(2\beta+1)}$  when  $g$  is assumed to be  $\beta$ -Hölderian.

The main drawback of this procedure is that we have  $\text{Card}(\bar{S}_n) \sim (n \log n)^{d/2}$  weak estimators to compute. This requires serious computational efforts when  $d$  is large (say, when  $d$  is of order 10). To solve this problem, we think of two directions:

1. The first technique is to use some convex optimization to minimize  $\theta \mapsto \sum_{i=m+1}^n (Y_i - \bar{f}_v(v^\top X_i))^2$ . Starting with the classical gradient descent algorithm to some more advanced minimization algorithms.
2. We can construct an iterative version of the estimator proposed in [Gaïffas and Lecué \(2007\)](#). For that, we start with a large  $\ell_2$ -net of  $S_+^{d-1}$  denoted by  $S_n^{(1)}$  with radius  $r_n^{(1)}$ . In the same spirit as [Lecué and Mendelson \(2008\)](#), we consider a set of almost empirical risk minimizers in  $S_n^{(1)}$ . This step provides a set  $\hat{S}_n^{(1)} \subset S_n^{(1)}$ . Now, we consider the convex hull in  $S_+^{d-1}$  of  $\hat{S}_n^{(1)}$ . We take a  $\ell_2$ -net of this convex hull of radius  $r_n^{(2)}$ , and continue with the same iterations, until  $r_n^{(k)}$  is small enough.

Depending on the number of iteration and of the size of the net (for the second algorithm), we believe that the number of weak estimators we have to compute can be much smaller than  $(n \log n)^{d/2}$ , hopefully of order  $(\log n)^{d/2}$ . This task, while described here for the regression model, is of importance for the practical implementation of the procedure proposed in Task 3 above, where we will need to compute a large number of weak estimators.

### 3.1.4 Task 4: High dimension

We are currently studying the following high dimensional problem: let be  $Y \sim \mathcal{N}(\mu, \Sigma)$  a Gaussian vector with a large dimensionality  $m$ , where the mean  $\mu = (\mu_i)_i \in \mathbb{R}^m$  is unknown and where  $\Sigma$  is different than the identity, with potentially large non-diagonal entries. How could we approach the non-zero components of  $\mu$ ? Or equivalently, what is  $\mathcal{H}_1 = \{i : \mu_i \neq 0\}$ ? In many cases, a limit central theorem gives an approximation of the matrix  $\Sigma$ , so that the latter can be assumed to be known. In this framework, we currently investigate two different methods to approach  $\mathcal{H}_1$ : multiple testing and model-based estimation methods. In both cases, the originality of the work is to integrate in our procedures the matrix  $\Sigma$ . This work is motivated by several applications in Biology (see sections below).

**Subtask 7: Multiple testing approach** (E. ROQUAIN, F. VILLERS, ?). Since the paper of [Benjamini and Hochberg \(1995\)](#), many methods have been developed for controlling the FDR (see for instance [Blanchard and Roquain \(2008\)](#) for a review). However, only few of them deal with a FDR control when the tests statistics have arbitrary dependencies. Moreover, such methods, like for instance the [Benjamini and Yekutieli \(2001\)](#) correction, are generally very conservative because they do not take into account the particular dependencies contained in the data (e.g.  $\Sigma$ ).

Considering the FWER and in the case of a Markov model, [Roquain \(2007\)](#) proposed to integrate  $\Sigma$  in a multiple testing procedure by simply generating the distribution of the supremum of all the tests statistics. Our aim is now to generalize this result to the case of the FDR control. For this, we would like to use the recent work of [Blanchard and Roquain \(2008\)](#), which put forward two sufficient conditions for FDR control. A major point will be to adapt to our setting their "dependency control condition", by linking precisely the distribution of the final number of rejections to the matrix  $\Sigma$ .

**Subtask 8: Estimation approach** (E. ROQUAIN, F. VILLERS, ?). The research field of the model selection has known an important development in the last decades, specially in Gaussian regression framework or for models with non-Gaussian noise under some condition on moments. However, most of papers studied the case of vector whose components are independent with a common known or unknown variance ([Birgé and Massart \(2001\)](#), [Huet \(2006\)](#), [Baraud et al. \(2007\)](#)). The case of heteroscedastic framework was studied in few papers only ([Comte and Rozenholc \(2002\)](#), [Gendre \(2008\)](#)). All the papers cited above deal with the case of a Gaussian vector whose components are independent. In a paper not published yet, Baraud and Gendre consider a gaussian vector with some known covariance matrix (known up to a scalar). They propose a penalty function to take into account the dependance assumption and provide upper-bound for the

quadratic risk of the selected estimator. To our knowledge, this is the only work dealing with dependencies between the components of the Gaussian vector.

From a methodological point of view, our aim is, following the work of Baraud and Gendre, to design a model selection procedure to estimate the mean of a vector whose components are dependent, and to provide a control of some risk of the selected estimator. From a more practical point of view, by exploiting dependencies, we aim at designing procedures giving better performances in practice than the procedure which does not take dependencies into account.

**Subtask 9: Detection of non-zero coefficients in the Cox model** (A. GUILLOUX, G. BIAU, PHD STUDENT). We restrict our presentation to the Cox model defined by:

$$\alpha_0(t, x) = \beta_0(t) \exp(v_0^\top x)$$

where  $v_0 \in \mathbb{R}^d$  is an unknown parameter and  $\beta_0(\cdot)$  is an unknown function. The convergence in probability and the  $\sqrt{n}$ -convergence in distribution of estimators  $\hat{v}$  in the Cox has been established, see Andersen et al. (1993) for a details. More precisely, the usual estimator  $\hat{v}$  is known to have to have a Gaussian limit distribution with a variance matrix which is not identity nor diagonal. In the Cox and Aalen models, see Equations (4) and (5) and Section 3.1.1, there exists estimators of the parameter  $v_0$ . The  $\sqrt{n}$ -convergence in distribution of these estimators in these two models has been established, see Andersen et al. (1993) for a detailed review, and estimators of  $\int \beta_0$  or  $\beta_0$  by kernel methods have been studied. Therefore, testing the null-hypothesis on the coordinates of  $v_0$  is close of the current work of E. ROQUAIN and F. VILLERS described in Subtasks 7 and 8, for the detection of zeros of a Gaussian vector with non-identity matrix. More generally, we believe that this should lead to a nice subject for a Ph.D thesis, with supervision of G. BIAU and A. GUILLOUX.

**Subtask 10: High dimensionality with correlations of covariates: the elastic-net approach** (S. GAÏFFAS, G. LECUÉ, ?). Another approach to handle the problem of high dimensional covariates with correlations is the following. Let us describe the idea in the following model to simplify the presentation, with in mind a linear version of the model described in Section 2.1 above. Consider the linear regression model with Gaussian error and deterministic design:

$$Y = X\beta^* + \epsilon \tag{15}$$

where  $Y \in \mathbb{R}^n$  is the output vector,  $X \in \mathcal{M}_{n,p}$  is a deterministic matrix,  $\epsilon$  is a standard Gaussian vector of dimension  $n$  (with a known variance matrix  $\sigma^2 I_n$ ) and  $\beta^*$  is a  $p$ -dimensional vector we wish to estimate. The classical approach OLS (ordinary least square) fails in the high dimensional setup:  $p$  much larger than  $n$ . In this context, we can hope that most of the coordinates of  $\beta^*$  are equal to zero. That is why it is natural to impose a constraint on the number of non-zero coefficients while performing the OLS. Thus, one can think of the following the algorithm, called " $L_0$ -regularization" or "subset selection":

$$\hat{\beta}^{(0)} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{ \|Y - X\beta\|_n^2 : \text{s.t. } \|\beta\|_0 \leq k \} \tag{16}$$

where  $\|\cdot\|_n$  is the euclidean norm of  $\mathbb{R}^n$  and  $\|\beta\|_0 = \sum_{j=1}^p I(\beta_j \neq 0)$ , where  $I$  is the indicator function. Unfortunately, the optimization problem (16) is NP-hard in the dimension  $p$ . Finding approximative algorithms which can be efficiently computed is an important problem. Up to now, mainly two procedures have been proved (theoretically and in practice) to tackle this high dimensional problem:

- The LASSO estimator:

$$\hat{\beta}^L \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{ \|Y - X\beta\|_n^2 + \lambda \|\beta\|_1 \}. \tag{17}$$

- The Dantzig selector:

$$\hat{\beta}^D \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{ \|\beta\|_1 \text{ s.t. } \|X^\top(Y - X\beta)\|_\infty \leq \lambda \}.$$

From a theoretical point of view, it is interesting to study the behavior of a procedure  $\hat{\beta}$ , in this context, for the three following problems:

Pb1: the rate of convergence of  $\hat{\beta}$  to  $\beta$ ;

Pb2: the rate of convergence of  $X\hat{\beta}$  to  $X\beta$ ;

Pb3: the probability that the set  $\{j : \hat{\beta}_j = 0\}$  equals to the set  $\{j : \beta_j = 0\}$ .

For the second problems (Pb1) and (Pb2), the rate of convergence of the LASSO and Dantzig procedures (w.r.t. the risk  $\|\cdot\|_n^2$ ) is

$$\frac{\sigma^2 s \log p}{n} \quad (18)$$

where  $s$  is the number of non-zero coefficient of  $\beta$  (cf. [Bickel et al. \(2008\)](#) and [Candes and Tao](#)). We can see that, thanks to this penalization technique, the classical residue of the OLS:  $(\sigma^2 p)/n$  has been replaced by  $(\sigma^2 s)/n$  (up to the  $\log p$  factor which is the price to pay for "adaptation to the sparsity"), where the dimension  $p$  of the input variable has been replaced by the true dimension  $s$  of the problem.

There are other procedures which can exploit this sparsity assumption from another point of view. Contrarily to the Lasso and Dantzig estimators, which use the  $\|\cdot\|_1$ -norm as a convex surrogate for the penalty  $\|\cdot\|_0$ , these procedures are based on some algorithmic considerations. They provide some approximation algorithm to the subset selection procedure. For instance, one of these procedure performs a recursive gradient descent algorithm (or "boosting" in learning theory) to find the best directions which minimize the risk up to a certain confidence bound. This step is usually called "Forward greedy algorithm". The main problem of this procedure is that it does not correct mistakes made during previous iterations. This is the reason why, some authors propose to include, inside the loops of the forward algorithm, a "backward loop" which can remove some directions which have been selected by error.

As the dimension  $p$  grows it is likely that the *covariates* (these are the  $p$  columns of  $X$ ) become correlated. All the results dealing with the LASSO or Dantzig require some assumption on the correlation matrix of the covariates (for instance, the uniform uncertainty principle, the irrepresentable condition, the restricted eigenvalue assumption, etc.). Unfortunately, these assumptions are unrealistic. Moreover, the LASSO estimator is known to have serious flaws (cf. [Zou and Hastie \(2005\)](#)):

1. when  $p > n$ , the LASSO selects at most  $n$  variables before it saturates;
2. if there is a group of variables among which correlations are very high, then the LASSO tends to select one of the group and doesn't care which one is selected;
3. when  $n > p$ , if there are correlations between predictors, performance of the LASSO is dominated by the ridge.

Therefore, it is important to provide a procedure which can handle some correlation structures between the covariates. For instance, we believe that the following procedure, known as *elastic-net*, may take into account these correlation structures:

$$\beta_{\lambda, \theta}^{(1)} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|_n^2 + \lambda(\theta|\beta|_1 + (1 - \theta)|\beta|_2^2) \}, \quad (19)$$

Remark that in [Zou and Hastie \(2005\)](#), a rescaling version of the elastic-net is suggested:  $\beta_{\lambda, \theta}^{(1)'} = (1 + \lambda(1 - \theta))\beta_{\lambda, \theta}^{(1)}$  to avoid a double shrinkage. From a theoretical point of view, the correct penalization may be  $\theta|\beta|_1 + (1 - \theta)|\beta|_2$  (which is a norm). But, from a practical point of view, the minimization of (19) is easier. The dual form of (19) is:

$$\beta_{t, \theta}^{(2)} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|_n^2 : \text{s.t. } \theta|\beta|_1 + (1 - \theta)|\beta|_2^2 \leq t \}, \quad (20)$$

or, more generally

$$\beta_\lambda^{(3)} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|_n^2 + \lambda \operatorname{pen}(\beta) \}, \quad (21)$$

where  $\operatorname{pen}(\cdot)$  is a penalty that shall have some properties (non-differentiability in 0 and strictly convexity).

### 3.1.5 Task 5: Other estimation settings

**Subtask 11: Recurrent events** (S. GEFFRAY, A. GUILLOUX, ?). S. Geffray and A. Guilloux are currently working on the analysis of recurrent events with few recurrences in the presence of dependent death and independent right-censoring. Examples of such recurrent events include repeated ischaemic attacks in atherosclerotic patients and tumor recurrences in cancer patients. The recurrence of such serious events is often associated with a high risk of death with the consequence that subjects may die during the study. On the other hand, during a trial, the further observation of both recurrent events and death may be definitively precluded by independent right-censoring. Possible reasons for independent right-censoring include loss-to-follow-up, end-of-study or independent death (i.e. death that is related neither to the disease nor to the treatment under study).

A classical analysis for this kind of data consists in skipping the times to recurrent events and in considering the survival times. We are then back to a usual independent right-censoring situation since, for a given individual, we do not observe both death and the censoring event but only the first event that occurs as time runs.

As in [Lin et al. \(1999\)](#), interest is centered on the successive inter-event gap times. We focus on the gap time between the  $(k - 1)^{\text{th}}$  recurrent event and the  $k^{\text{th}}$  recurrent event rather than on the total time from treatment start to the  $k^{\text{th}}$  recurrent event even if considering the total time from treatment start to the  $k^{\text{th}}$  recurrent event would be interesting as a complementary approach. Note that, unlike [Lin et al. \(1999\)](#), we do not make the restrictive assumption that each individual experiences a fixed number of recurrent events.

The methodology of [Lin et al. \(1999\)](#) is adapted to take dependent death into consideration by treating death as a dependent competing risk acting at each disease recurrence as in [Li and Lagakos \(1997\)](#). To refine our analysis, we consider separately recurrent event and death occurrence. The present work focuses on the cause-specific distribution function of two successive gap times with first and second outcomes being non-fatal and on the cause-specific distribution function of two successive gap times with first outcome being non-fatal and second outcome being fatal. Nonparametric estimators of these functions are proposed. We show that these estimators are strongly consistent and have a Gaussian asymptotic behavior. We intend to use these estimators to derive a two-sample testing procedure. We anticipate that a bootstrap step will be helpful to get a variance estimation of better quality. Then, simulations will be carried out to investigate the finite-sample behavior of the proposed estimators and tests. We also plan to discuss covariates incorporation techniques. See [Geffray \(2006\)](#) for more details.

**Subtask 12: SIM for recurrent events** (O. BOUAZIZ, S. GEFFRAY, O. LOPEZ). O. Bouaziz, S. Geffray and O. Lopez are currently focusing on adapting dimension reduction techniques to regression models for the recurrent event process. This process denoted by  $N(t)$  counts the number of recurrent events occurring over the time interval  $[0, t]$  and reflects the disease evolution through time.

Practical examples of recurrent events include asthma or epileptic seizures. In these examples, studying  $N(t)$  gives valuable informations on the patient state of health since the more steps the process  $N(t)$  records the badder the patient is.

S. GEFFRAY worked on recurrent event problems in her Ph.D. thesis and is currently pursuing this work in a collaboration with A. GUILLOUX (see Task 11 above) while [Lopez \(2008\)](#) and [Bouaziz and Lopez \(2008\)](#) studied dimension reduction in a censored regression model (see Section 3.1.3).

The aim of the current work of O. BOUAZIZ, O. LOPEZ, S. GEFFRAY is to combine tools coming from the classical theory of recurrent events and from the single-index regression model. This constitutes a working paper entitled "Semi-parametric inference for the recurrent event process by means of a single-index model". The originality of this working paper lies in the fact that the period  $T$  of observation of a patient is censored, but the model stands on  $N(t)$  instead of standing on the censored variable directly.

When a covariate vector  $X$  is available, many regression models for  $N(t)$  have been developed in the literature according to three approaches: conditional approach (intensity models of the Cox-type conditional on the medical history of a patient), marginal approach (models focusing on the cumulative mean function  $M(t|X) = \mathbb{E}[N(t)|X]$ ) and frailty approach (an unobserved covariate reflects the general state of health of a given patient). O. BOUAZIZ, O. LOPEZ, S. GEFFRAY focus on the marginal approach which provides a better interpretation in terms of risk factors comparison and treatment effects identification.

Consider a single-index regression model on  $N(t)$ , that is assume the existence of a parameter vector  $v_0$  such that  $M(t|X) = E [N(t)|v_0^\top X = v_0^\top x]$ . This new model for recurrent events relies on a dimension reduction assumption :  $M(t|X)$  only depends on a unknown linear combination of the covariates. This can be easily generalized to multi-index regression where  $M(t|X)$  is assumed to depend on  $v_0^\top X, v_1^\top X, \dots, v_k^\top X$ . In addition, even in the case where  $v_0$  is exactly known, the function  $M$  is not specified and belongs to a nonparametric family. In practice, the dimension reduction assumption may not be fulfilled. However, as it is implicit in the single-index literature, the model can correctly approximate the true purely nonparametric model provided that the number  $k$  of indexes is sufficiently high.

This semi-parametric approach relies on weaker assumptions than those proposed so far in the literature while maintaining interesting properties. An drawback of existing approaches for modeling recurrent events in presence of covariates is that these approaches often rely on making some assumptions about the regression model of the lifetime  $T$  of the patient. An interesting feature of our approach is that we avoid this drawback by directly modeling  $N(t)$ . Therefore, no dimension reduction assumption on the lifetime is required.

A  $n^{1/2}$ -consistent estimator of the parameter vector  $v_0$ , is constructed and used to estimate  $m_{v_0}(\cdot) = E [N(t)|v_0^\top X = \cdot]$ . Rates of convergence for the final estimator of  $M(t|X)$  are provided. We plan to use these estimators to define a two-sample testing procedure. We also plan to discuss model checking techniques.

**Subtask 13: Time-dependent covariates** Since longitudinal data occur over time, important covariates we wish to consider may also change over time. These covariates are referred as time-dependent covariates. Examples of such covariates are:

- treatments, blood pressure, smoking status or cumulative exposure to some risk factor,
- the number of recurrence of an event,
- sojourn time, that is, the duration in a given state (e.g. disease),
- artificial covariates (e.g. for testing proportional hazards assumption).

Therefore, it is of great interest that the estimation methods developed in this project are able to take into account time-dependent covariates.

Consider, as an example, the classical Cox proportionnal hazards model. Several methods have been proposed to take into account time-dependent covariates (Therneau and Grambsch (2000)). Among them, the counting process approach allows in a simple way to handle the problem of time-dependent covariates. Each subject is simply represent by a set of observations over its interval at risk, with time intervals delimited by the time of change in the time-dependent covariates status. The partial likelihood looks almost identical to one derived for time-independent covariates. The only difference is that at time  $t$ , the values of the time-dependent covariates were used both for individual who dies at that time, as well as the individuals who are at risk sets at that time. The same approach based on counting process could be studied in order to deal with time-dependent covariates in more complex models.

The use of time-dependent covariates is much more complicated in practice than the fixed (time-independent) covariates (Fisher and Lin (1999)). Indeed, the modeling of such covariates involves the choice of a function over time which is far from obvious in practice. For instance, consider a sample of workers in a factory expose to a toxic product: we may use the cumulative exposure, the average exposure up to time  $t$  or the maximum exposure up to time  $t$ . Moreover, at each event time we need to know the exact value of the covariate at that event for all individuals under study (a natural choice seems to used the convention that the values of the covariates for a patient under observation at some given time  $t$  are the values recorded at the last

follow-up examination before time  $t$ , nevertheless other approaches would be possible). Another difficulty is that with time-dependent covariates the ability to predict is usually lost because the model depends on the value of a changing quantity (the time-dependent covariate), at a future time the future values are usually unknown until they are actually observed. Care must also be exercised in interpretation when modeling time-dependent covariates like exposure or treatment, especially if the change in exposure or treatment is related to the subject's health status.

**Subtask 14: Semi-Markov process** In the study of transition intensities of nonhomogeneous time-continuous Markov processes (see example 3 of section 2.1.1), the r.v.  $N_{jl}^i(t)$  which counts the number of observed direct transitions from state  $h$  to  $l$  before time  $t$  for individual  $i$  is introduced. These counting processes satisfies the Aalen multiplicative intensity model. Such property allows to derive asymptotic results in the case of Markov process (see example IV.1.5 Andersen et al. (1993)).

The same results can be obtained for a semi-Markov process in the illness-death model. The states 0,1 and 2 denote "healthy", "diseased" and "dead" and transitions  $0 \rightarrow 1$ ,  $0 \rightarrow 2$  and  $1 \rightarrow 2$  are possible. The transitions intensities  $\alpha_{01}(t)$ ,  $\alpha_{02}(t)$  depend on time and  $\alpha_{12}(t, d)$  depends on time and duration in state 1. The trivariate counting process  $N(t) = (N_{01}(t), N_{02}(t), N_{12}(t))$  counts the number of transitions in  $[0, t]$ . The intensity of the process  $N_{0h}(t)$ ,  $h = 1, 2$ , is  $\alpha_{0h}(t)Y_0(t)$  with  $Y_0(t) = 1 - N_{01}(t^-) - N_{02}(t^-)$  whereas  $N_{12}(t)$ ,  $h = 1, 2$  has intensity process  $\alpha_{12}(t, t-T)Y_1(t)$  with  $Y_1(t) = N_{01}(t^-) - N_{02}(t^-)$  and  $T = \inf\{t : N_{01}(t) = 1\}$ . In the situation where  $\alpha_{12}(t, d)$  depends only on  $t$  and not on  $d = t - T$ , the illness-death process is a Markov process. On the other hand, assume that  $\alpha_{12}(t, d)$  depends only on duration  $d$ . Since time and duration in state 0 coincide for individuals still in 0, the process is then a semi-Markov process. The intensity of  $N_{12}(t)$  is  $\alpha_{12}(t-T)Y_1(t)$  which does not factor into a product of a deterministic function and an observable process; hence, the model is not an Aalen multiplicative intensity model. However, considering duration  $d = t - T$  as the basic time variable, that is, defining  $K(d) = N_{12}(d + T)$ ,  $U(d) = Y_1(d + T)$ . The counting process  $K$  has an intensity process  $\alpha_{12}(d)U_1(d)$  with respect to the filtration  $G_\tau \vee K_d$  where  $G_\tau = \sigma\{(N_{01}(t), N_{02}(t)) : 0 < t < \tau\}$  and  $K_d = \sigma\{K(d) : 0 < d < \infty\}$ . As a consequence, the multiplicative intensity model is satisfied in this time scale and then the same kind of asymptotic results can be derived.

The above time transformation will not work in semi-Markov processes allowing transitions back and forth, that is, it is possible to return in at least one state once it has been left. Consider a semi-Markov process with  $k$  states specified by transition  $\alpha_{hj}(t)$ ,  $h, j = 1, \dots, k, h \neq j$ , depending on duration in state  $h$ ,  $h = 1, \dots, k$ . Define  $J_0, J_1, \dots$  the state occupied before jump 1, 2, ... and let  $N_{hj}(t)$  be the number of transitions from  $h$  to  $j$  in  $[0, t]$  and  $N(t) = \sum_h \sum_j N_{hj}(t)$  be the total number of transitions in  $[0, t]$ . The state currently occupied  $Z(t) = J_{N(t)}$  a semi-Markov process. Let  $X_i$  be the  $i$ -th sojourn time (or duration),  $S_0 = 0$ ,  $S_i = X_1 + \dots + X_i$  and  $L(t) = t - S_{N(t^-)}$ . The counting process  $N_{hj}(t)$  has intensity process given by  $\alpha_{hj}(L(t))Y_h(t)$  where  $Y_h(t) = \mathbf{1}_{Z(t^-)=h}$ . However, as mentioned above, this is not a multiplicative intensity model. As above, it is possible to use the intrinsic time scale of this process: duration rather than "calendar" time. Define  $K_{hj}(d)$  = the number of durations in  $h$  observed to take on a value  $\leq d$  and to be followed by a jump to  $j$ ,  $U_h(d)$  = the number of durations in  $h$  observed to take on a value  $\geq d$ . Let us defined

$$H_{hj}(d) = K_{hj}(d) - \int_0^d U_h(z)\alpha_{hj}(z)dz,$$

but, unlike the semi-Markov in an illness-death model, there is no filtration making  $H_{hj}$  a martingale (Gill (1980)). Therefore, the results on martingale like the central limit theory are not applicable. It is no longer possible to use the time transformation tricks in order to obtain asymptotic results for a general semi-Markov process. However, it has been shown (Gill (1980)) that  $H_{hj}$  have the same properties as the counting process martingales  $M_{hj}$  generated by a Markov process (Example 3 of section 2.1.1). Indeed, it is possible to show that  $H_{hj}(\cdot)$  has uncorrelated increments. Then identical asymptotic results, as for Markov processes, can be derived using alternative techniques to the martingale central limit theory (Gill (1980)).

### 3.1.6 Task 6: Some parallel developments

Some parallel developments can be useful for the project. These are related to the research of G. LECUÉ on statistical learning theory and empirical process theory. Some of these shall be in collaboration with S. MENDELSON (Pr. at Australian National University), who could eventually be invited professor with the project, see Section 5 below.

**Subtask 15: SVM and sparsity** (G. LECUÉ, ?). Consider the classification setup, where one is given a set of  $n$  observations  $(X_i, Y_i)_{1 \leq i \leq n}$  where  $X_i \in \mathcal{X}$  and  $Y_i \in \{-1, 1\}$ . Many classification procedures have the following form

$$\hat{f}(x) = \text{Sign}(\hat{\eta}(x)) \text{ where } \hat{\eta}(x) = \sum_{j=1}^d \hat{\alpha}_j \psi_j(x), \forall x \in \mathcal{X}, \quad (22)$$

where  $(\psi_1, \dots, \psi_d)$  is a dictionary which has to be chosen wisely for the classification problem. We would like to take into account some sparse underlying structure. Consider the class of functions

$$F = \left\{ f_\alpha := \sum_{j=1}^d \alpha_j \psi_j : \alpha_j \in \mathbb{R} \right\}$$

and for any vector  $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$ , we consider the function

$$J(\alpha) = \text{Card}\{j \in \{1, \dots, d\} : \alpha_j \neq 0\}.$$

Consider  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  a loss function for the classification problem. Many different losses have been discussed in the literature along the last decade for instance:

$\phi(x) = \mathbb{I}_{(x \leq 0)}$	classical loss or 0 – 1 loss
$\phi(x) = \max(0, 1 - x)$	hinge loss (SVM loss)
$\phi(x) = \log_2(1 + \exp(-x))$	logit-boosting loss
$\phi(x) = \exp(-x)$	exponential boosting loss
$\phi(x) = (1 - x)^2$	squared loss
$\phi(x) = \max(0, 1 - x)^2$	2-norm soft margin loss

and consider the risk and its empirical version associated with  $\phi$ :

$$R^\phi(f) = E\phi(Yf(X)) \text{ and } R_n^\phi(f) = \frac{1}{n} \sum_{i=1}^n \phi(Y_i f(X_i)).$$

When  $\phi$  is the 0 – 1 loss, the risk is denoted by  $R$  (without the exponent  $\phi$ ). The minimum  $\phi$ -risk is defined by

$$R^{\phi*} = \inf_{f \text{ measurable}} R^{\phi*}(f).$$

We want to prove some SOI (Sparse Oracle Inequalities) w.r.t. the  $\phi$ -risk and the 0 – 1 risk:

$$E[R^\phi(\hat{f}) - R^{\phi*}] \leq C \min_{s=1, \dots, d} \min_{f_\alpha : J(\alpha) \leq s} \left( R^\phi(f_\alpha) - R^{\phi*} + \frac{s \log n}{n} \right). \quad (23)$$

and SOI w.r.t. the 0 – 1 loss:

$$E[R(\hat{f}) - R^*] \leq C \min_{s=1, \dots, d} \min_{f_\alpha : J(\alpha) \leq s} \left( R(f_\alpha) - R^* + \left( \frac{s \log n}{n} \right)^{\frac{\kappa}{2\kappa-1}} \right). \quad (24)$$

We hope that some  $\ell_1$ -penalization technic would satisfied this kind of SOI:

$$\hat{f} \in \underset{f_\alpha : \alpha \in \mathbb{R}^d}{\text{argmin}} \left( R_n^\phi(f_\alpha) + \lambda |\alpha|_1 \right). \quad (25)$$

A nice dictionary may be found in [Wainwright et al. \(2006\)](#) and an outline of the proof in [van de Geer \(2008\)](#) and [Tarigan and van de Geer \(2006\)](#). The residual term should depend on some margin parameters:

$$E(\phi(Yf(X)) - \phi(Yf^*(X)))^2 \leq c(R^\phi(f) - R^{\phi*})^{1/(2\kappa)}, \forall f$$

or

$$E|f(X) - f^*(X)| \leq c(R(f) - R^*)^{1/(2\kappa)}, \forall f : \mathcal{X} \mapsto \{-1, 1\}.$$

Maybe, to obtain SOI w.r.t. the 0 – 1 loss for the procedure defined in (25), we will have to prove some sharp inequalities:

$$R(\text{sign}(f)) - R^* \leq c(R^\phi(f) - R^{\phi*}(f))^?, \forall f_\alpha.$$

**Subtask 16: Form of coordinate projections** (G. LECUÉ, ?). This concept is at the heart of many learning problems. It has been first introduced in many different works of S. Mendelson. Assume that we have  $n$  observations  $D^{(n)} = (Z_i)_{1 \leq i \leq n}$ . Given is a class of function  $F$ , we want to construct a procedure  $\hat{f}^{(n)}$  which has a risk as close as possible to the minimal risk over  $F$ . We define the risk of a function  $f$  and of an estimator  $\hat{f}^{(n)}$  by

$$R(f) = E[Q(Z, f)] \text{ and } R(\hat{f}^{(n)}) = E[Q(Z, \hat{f}^{(n)}) | D^{(n)}].$$

A classical approach is the empirical risk minimization algorithm defined by

$$\hat{f}^{(n)} \in \underset{f \in F}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n Q(Z_i, f).$$

We assume that there exists  $f^* \in F$  such that  $\inf_f R(f) = R(f^*)$ , where the infimum is taken over all measurable functions. Classical tools (the Giné-Zinn symmetrization argument, a majoration of the expectation of supremum of Rademacher processes by Gaussian processes and the Dudley's entropy integral) provide the following upper bound

$$\begin{aligned} E[R(\hat{f}^{(n)})] - \inf_{f \in F} R(f) &\lesssim E_{D^{(n)}} \left[ E_g \left[ \sup_{t \in P_\sigma \mathcal{F}} \frac{1}{n} \sum_{i=1}^n g_i t_i \right] \right] \\ &\lesssim E_{D^{(n)}} \left[ \frac{1}{\sqrt{n}} \int_0^{\text{Diam}(P_\sigma \mathcal{F}, \|\cdot\|_{2,n})} \sqrt{\mathcal{N}(P_\sigma \mathcal{F}, \epsilon, \|\cdot\|_{2,n})} d\epsilon \right], \end{aligned} \quad (26)$$

where  $g_1, \dots, g_n$  are i.i.d. standard Gaussian variables,  $\mathcal{F} = \{Q(\cdot, f) - Q(\cdot, f^*) : f \in F\}$ ,  $P_\sigma \mathcal{F} = \{(f(Z_1), \dots, f(Z_n)) : f \in \mathcal{F}\}$ ,  $\|x\|_{2,n} = \left(\frac{1}{n} \sum_{i=1}^n x_i^2\right)^{1/2}$  is the normalized  $\ell_2$ -norm,  $\text{Diam}(P_\sigma \mathcal{F}, \|\cdot\|_{2,n})$  is the diameter of  $P_\sigma \mathcal{F}$  w.r.t. the norm  $\|\cdot\|_{2,n}$  and  $\mathcal{N}(P_\sigma \mathcal{F}, \epsilon, \|\cdot\|_{2,n})$  is the entropy of  $P_\sigma \mathcal{F}$  w.r.t.  $\|\cdot\|_{2,n}$ .

The important quantity of (26) that we would like to focus on is the form of the set  $P_\sigma \mathcal{F}$  that we call *the coordinate projections*. Let us formalize this problem: let  $\mathcal{F}$  be a class of function on  $\mathcal{Z}$  (resp.  $V$  a class of vectors of  $\mathbb{R}^d$  with  $d \gg n$ ), let  $(Z_i)_i$  be a sequence of i.i.d. random variables with values in  $\mathcal{Z}$  (resp. with values in  $\{1, \dots, d\}$ ). What is the typical form of the random set of  $\mathbb{R}^n$ :

$$P_\sigma \mathcal{F} = \{(f(Z_1), \dots, f(Z_n)) : f \in \mathcal{F}\} \text{ (resp. } P_\sigma V = \{(v_{Z_1}, \dots, v_{Z_n}) : v \in V\}.$$

We would like a kind of result (for the class of coordinates projection) in the same spirit as the probabilistic version of Dvoretzki's theorem by Milman: let  $n$  be an integer and  $\epsilon > 0$ . There exists an integer  $N(n, \epsilon) \lesssim \exp(C(\epsilon)n)$  such that the following holds. Let  $K$  be a convex symmetric body of a Banach space  $B$  of dimension  $d$ , such that  $d \geq N(n, \epsilon)$ . With probability very close to 1 (w.r.t. the measure endowed by  $n \times k$  Gaussian matrices with independent entries and the subspaces are the images of  $\mathbb{R}^k$  in  $\mathbb{R}^n$  using such matrices), a random  $n$ -dimensional subspace  $E = \text{Span}(x_1, \dots, x_n)$  satisfies  $|\beta|_2 \leq \left\| \sum_{i=1}^n \beta_i x_i \right\|_K \leq (1 + \epsilon)|\beta|_2, \forall \beta \in \mathbb{R}^n$ , where  $\|b\|_K = \inf\{t \geq 0 : x \in tK\}, \forall b \in B$  defined the norm associated with  $K$ . This means that, with large probability, when we project a convex symmetric body on a  $n$ -dimensional subspace, this projection behaves (up to  $\epsilon$ ) as the Euclidean ball of  $\mathbb{R}^n$ .

Dvoretzky's Theorem cannot be directly applied in the case of coordinate projections, because the set of coordinates is of null measure inside the set of all vectorial subspaces of dimension  $n$ . Nevertheless, we believe that we can adapt Milman's proof to our setup (maybe the Shannon basis should be useful), starting with the following question: *Let  $(\psi_j)_j$  be an orthonormal basis of  $L_2(\lambda)$ . Let  $(a_j)_j \in \ell_2(\mathbb{N})$  be a decreasing sequence of positive numbers. Consider the Sobolev ellipsoid:*

$$\mathcal{F} = \left\{ f = \sum_i \alpha_i \psi_i \in L_2(\lambda) : \sum_i (\alpha_i/a_i)^2 \leq 1 \right\}.$$

*Let  $(Z_i)_i$  a sequence of i.i.d. random variables uniformly distributed on  $[0, 1]$ . What is the form of the coordinate projections*

$$P_\sigma \mathcal{F} = \{(f(Z_i), \dots, f(Z_n)) : f \in \mathcal{F}\} ?$$

The idea of the proof is that we would like to construct a  $L_2$  basis (in the same spirit as the Shannon basis) such that  $(f(X_1), \dots, f(X_n))$  would be the beginning of the expansion of  $f$  in this basis. This random basis would provide a measure on all the subspaces of dimension  $n$  of  $L_2$  which can then be plug in Milman's version of the proof of Dvoretzky's theorem.

**Subtask 17:  $\ell_1$ -penalization and Convex Asymptotic Geometry** (G. LECUÉ; ?). High dimensional problems are at the heart of the field of Convex Asymptotic Geometry (CAG). We would like to introduce some ideas and technics coming from the CAG community to solve some high dimensional problems in statistics. Following the lines of [S. Mendelson \(2008\)](#), we consider the quantity

$$E \left[ \sup_{v \in B(r) \cap S_1(t)} \frac{1}{n} \sum_{i=1}^n \langle \Psi(X_i), v \rangle \right]$$

where  $S_1(t)$  is the  $\ell_1^{d-1}$  sphere of radius  $t$ ,  $\Psi(x) = (f_1(x), \dots, f_M(x))^t$  (for  $\{f_1, \dots, f_M\}$  a given dictionary) and  $B(r)$  can be some  $\ell_p^d$  balls of radius  $r$  or the set  $\{v : \text{ent}(v, \pi) \leq r\}$  (where  $\pi$  can be an a priori probability measure on the dictionary). These quantities are at the heart of the analysis of  $\ell_1$  penalized algorithm. At the same time, using these technics we would like to obtain better lower bound for the ERM over the convex hull of a finite class of functions. We believe that these technics may also lead us to prove that the classical aggregation with exponential weights (cf. for instance [Lecué \(2007\)](#)) is an optimal aggregation procedure (and hopefully, a nice choice of the Temperature parameter may come out of the proof). Some useful references may be found in [Guédon et al. \(2007\)](#), [Gordon et al. \(2007\)](#) and [Klartag and Mendelson \(2005\)](#). This part of the project research may provide some results to the Subtasks 4, 10 and 15.

**Subtask 18: General study of the ERM procedure in the agnostic setup** (G. LECUÉ, ?). The aim of this section is to study the Empirical Risk Minimization (ERM) algorithm in the functional learning setup. Many results have already been obtained (cf., for instance, [Bartlett and Mendelson \(2006\)](#), [Mendelson \(2008\)](#) or [Mendelson and Lecué \(2008\)](#)), but there are still some room for improvement.

Let  $D^{(n)} = (X_i, T(X_i))_{1 \leq i \leq n}$  be a set of  $n$  observations. The function  $T$  is called the target function. Given is a set  $F$  of functions, we want to construct a procedure  $\hat{f}$  (using the observations  $D^{(n)}$ ) which has a risk as close as possible to the minimum risk  $\min_{f \in F} R(f)$ . We focus on the  $L_2$ -risk (we can also consider any other risk having some convexity properties). We define the risk of a real-valued function  $f$  and of a statistic  $\hat{f}$  by

$$R(f) = E(f(X) - T(X))^2 \text{ and } R(\hat{f}) = E[(\hat{f}(X) - T(X))^2 | D^{(n)}].$$

The Empirical risk minimization algorithm (ERM) is defined by

$$\hat{f} \in \underset{f \in F}{\text{argmin}} R_n(f) \text{ where } R_n(f) = \frac{1}{n} \sum_{i=1}^n (f(X_i) - T(X_i))^2.$$

We would like to obtain a sharp result of the following form. There exists three absolute constants  $0 < c_0 \leq c_1, c > 0$  such that,

1. with high probability,

$$R(\hat{f}) - \min_{f \in F} R(f) \leq c_1 \phi(F, n).$$

2. with probability greater than  $c$ ,

$$c_0 \phi(F, n) \leq R(\hat{f}) - \min_{f \in F} R(f).$$

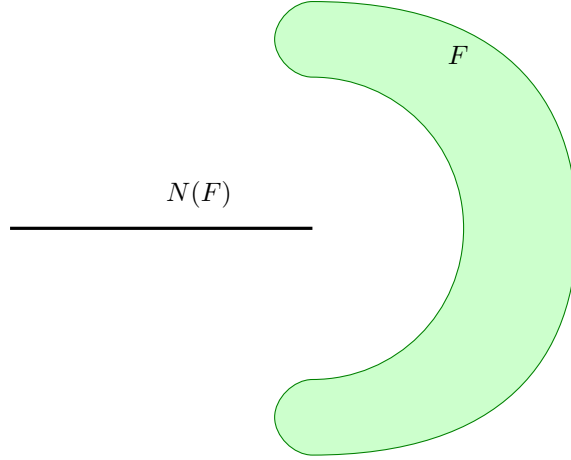
For this problem the geometry of the set  $F$  is of first importance. We introduce two quantity which are at the heart of this analysis:

1. We say that the couple  $(T, F)$  (where  $T$  is the target and  $F$  the class of candidates) satisfies the *Bernstein's condition with parameter*  $0 < \beta \leq 1$  when

$$\forall f \in F, E[(f(X) - T(X))^2 - (f_T^*(X) - T(X))^2] \leq B(R(f) - R(f_T^*))^\beta,$$

where  $R(f_T^*) = \min_{f \in F} R(f)$  and  $f_T^* \in F$ .

2. the set of oracles  $N(T, F) = \{f \in F : R(f) = \min_{f \in F} R(f)\}$  and  $N(F) = \{T \in L_2 : \text{Card}(N(T, F)) \geq 2\}$ .



We summarize some results from [Mendelson \(2008\)](#) or [Mendelson and Lecué \(2008\)](#).

1. when the target  $T$  is far from  $N(F)$  then the Bernstein's parameter is  $\beta = 1$  (best case of the Bernstein's condition) and so we can obtain an upper bound of the order of  $1/n$ .
2. when  $T$  belongs to  $N(F)$ , no more Bernstein's condition holds and we can obtain a lower bound of the order of  $1/\sqrt{n}$ .
3. there is a degradation of the constant  $B$  when the target comes closer to the set  $N(F)$ . This degradation of the Bernstein's condition implies a degradation of the rate of convergence (from  $1/n$  to  $1/\sqrt{n}$ ).

We want first to focus on the case  $T \in N(F)$  (that is  $T$  has at least two oracles). We consider the set  $F_a = \{f \in F : R(f) \leq \min_{f \in F} R(f) + a/\sqrt{n}\}$ . It is easy to prove that, there exists some  $a < b$ ,

$$c_0 \frac{H(F_a)}{\sqrt{n}} \leq R(\hat{f}) - \min_{f \in F} R(f) \leq c_1 \frac{H(F_b)}{\sqrt{n}} \quad (27)$$

where the first inequality holds w.p.g. than  $c$  and the second w.h.p. and  $H(Q) = E \sup_{f \in Q} G_f$  where  $(G_f)_{f \in Q}$  is the canonical Gaussian process associated with  $Q$ . We would like to make  $a$  and  $b$  in (27) as

close as possible. This leads to treat carefully the "complexity" behavior of  $F$  in the corona  $F_b - F_a$ . This can be done by studying the function

$$\zeta_n(r) = E \left[ \sup_{f \in F: R(f) = \min_{f \in F} R(f) + r} R(f) - R_n(f) \right], \forall r > 0. \quad (28)$$

We believe that the function  $r > 0 \mapsto \zeta_n(r)$  contains all the information about the complexity of  $F$  and the geometry between  $F$  and  $T$ .

### 3.1.7 Task 7: Applications

**Subtask 19: Applications in Biology** (E. ROQUAIN, F. VILLERS, ?). Taking into account the dependencies existing within the data is a problem of interest in different applications. One part of the thesis of F. Villers was in particular devoted to detect proteins whose abundance differs according to the experimental condition. In statistical terms, this comes to detect the non zero components of a Gaussian vector. When one compares simultaneously a large number of experimental conditions (at least 3), the components of the Gaussian vector are not independent. To this aim, Villers has in particular worked with a first procedure proposed by Baraud (procedure that Baraud and Gendre are currently improving) to take into account dependencies.

Currently, in continuity with this work, a collaboration between S. Schbath and S. Huet (INRA Jouy-en-Josas), X. Gendre (Université de Nice), E. Roquain and F. Villers started around the problem of detection of exceptional words in a DNA sequence. A major problem in biological sequence analysis is indeed to find significantly under- or over- represented words (i.e. small DNA sequences), as they may have a particular biological function. The significance of each word can be computed by counting the number of occurrences of the word in the observed sequence and by comparing it to what is expected. This problem was studied, for instance [Robin et al. \(2003\)](#) and [Roquain \(2007\)](#), when the sequence is modelised by a Markov chain. To each word is thus associated some score of significance measuring the exceptionality of its count. The problem at hand is to detect simultaneously exceptional words among all the words of a given size in a DNA sequence. Since the number of occurrences of two different words are always correlated, (especially highly positively correlated when they overlap on large set of bases), the scores of significance are in this case correlated. According to the Gaussian approximation of [Prum et al. \(1995\)](#), the vector of the scores of significance can be modelised by a Gaussian vector. Moreover, the covariance matrix can be easily estimated and can therefore be assumed known. The detection of exceptional words leads thus to detect the non zero components of a Gaussian vector whose covariance matrix is known.

**Subtask 20: Applications in Cancerology** (A. DUVAL, A. GUILLOUX, ?). As mentioned in Section 2.2, one of the main objectives of this project is to develop a toolbox to apply the methods developed in the theoretical part of the project. We would like to analyze a dataset provided by A. DUVAL. The data consist of a large panel of tumor samples, for which they conducted an analysis of genomic (see Section ) and expression alterations. This project is supported by the Tumor ID-card program (CIT) from LNCC (**Ligue Nationale contre le Cancer**) and INCa (**Institut National du Cancer**). The figure below shows the transcriptomic profile (expression alterations) for around 2000 genes of colorectal tumoral cells for 120 patients, for which disease-free survival times (and status) and other clinical or biological variables are available. We aim to provide a method and informatic tools to investigate the influence of the expression alteration of the genes on the risk of a disease recurrence.

### 3.2 Planning of tasks

In Figure 2 below, we give the Gantt diagram of the project. It contains an approximative planning of each of the subtasks described in Section 3.1 above.

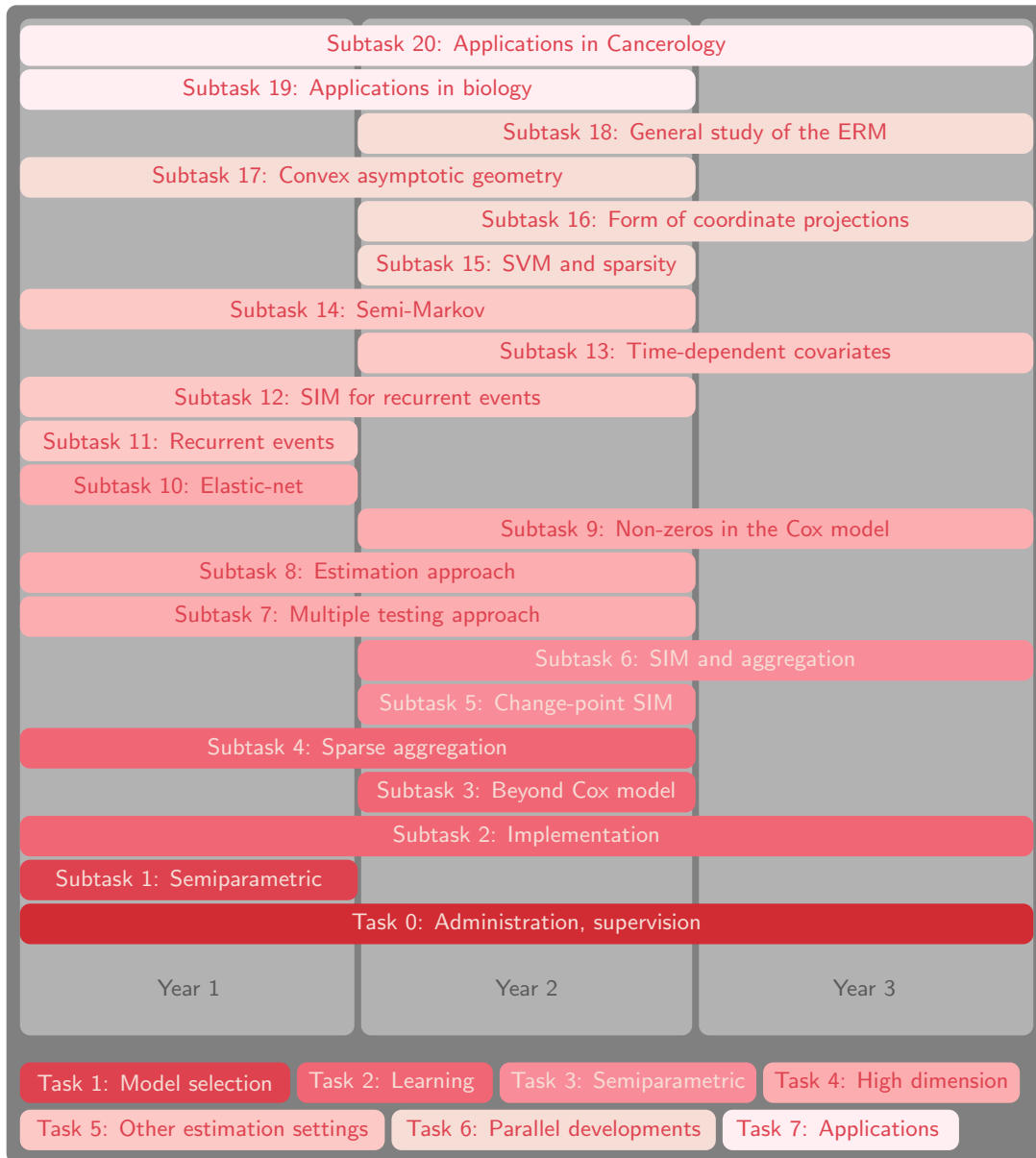


Figure 3: Gantt diagram of the project

## 4 Consortium organisation and description

### 4.1 Qualification of the principal investigator

AGATHE GUILLOUX is 30 years old and holds since 2005 an Assistant professor position at UPMC (Université Pierre et Marie Curie-Paris 6). She is member of two laboratories: the LSTA (Theoretical and Applied Statistics Lab., director Prof. P. Deheuvels) and the medical Research Center of Saint-Antoine hospital (Unit 13: “Instability of micro-satellites and cancer”, director A. Duval). She graduates from her PhD in 2004 under the supervision of Prof. M. Delecroix and Prof. J-Y. Dauxois and from two masters before that: one in Mathematics (major in Statistics) and one in Biostatistics and Epidemiology. As a member of UPMC, she is also involved in several administrative activities: former vice-presidency of the “Commission des Spécialistes” for Applied Mathematics, elected member for the ISUP board (the statistical school of UPMC), etc.

Her research activities may be divided in two parts: Theoretical Statistics and Biostatistics. Her research interest in Theoretical Statistics is mainly in survival analysis and counting processes, for which she works on non-parametric inference, limit theorems, tests, concentration inequalities and bootstrap. She published 10 papers in international journals, one chapter of a book and gave approx. 25 conferences or seminars. She currently collaborates with 8 researchers from various universities, 3 of them are participants in this project. These collaborations give rise to 8 submitted or working papers. She was invited at the University of Northern Iowa (USA - 6 months) and the Vigo University (Spain - 1 week).

In Biostatistics, she formerly worked at the Department of Descriptive Epidemiology at the International Agency for Cancer (WHO, Lyon, France) and is currently working at the Research Center of Saint-Antoine hospital. She has 3 published papers in international journals and one submitted paper.

She also supervised the PhD thesis of SÉGOLEN GEFFRAY, in collaboration with Prof. P. Deheuvels, and a graduate dissertation. As a teacher, she published a book on Applied Statistics.

### 4.2 Contribution and qualification of each project participant

The team consists of 12 young researchers from different laboratories and universities. In Table 1, we give, for each member of the team, his involvement in the project in percentage and her/his affiliation. Most of them hold the position “Maître de Conférences” (MdC). We recall that the principal investigator is A. GUILLOUX.

NAME	SURNAME	Position (since)	Affiliation	Person.month
GUILLOUX	AGATHE	MdC (2005)	LSTA and CRSA E13 (UPMC, INSERM)	28.8
GAÏFFAS	STÉPHANE	MdC (2007)	LSTA (UPMC)	21.6
LECUÉ	GUILLAUME	CR2 (2007)	LATP (UdP)	21.6
GEFFRAY	SÉGOLEN	MdC (2007)	(UdN)	21.6
SAINT-PIERRE	Philippe	MdC (2006)	LSTA (UPMC)	18
LOPEZ	OLIVIER	MdC (2008)	LSTA (UPMC)	18
VILLERS	FANNY	MdC (2008)	LPMA (UPMC)	9
ROQUAIN	ÉTIENNE	MdC (2008)	LPMA (UPMC)	7.2
BOUAZIZ	OLIVIER	PhD Student	LSTA (UPMC)	7.2
BIAU	GÉRARD	Professor (2004)	LSTA (UPMC)	3.6
DUVAL	ALEX	DR2 INSERM (2007)	CRSA E13 (INSERM, UPMC)	3.6

Table 1: Who’s who

We used the following acronyms: UPMC=Université Pierre et Marie Curie ([www.upmc.fr](http://www.upmc.fr)), LSTA=Laboratoire de Statistique Théorique et Appliquée ([www.lsta.upmc.fr](http://www.lsta.upmc.fr)), LPMA = Laboratoire de Probabilités et Modèles

Aléatoires ([www.proba.jussieu.fr](http://www.proba.jussieu.fr)), LATP = Laboratoire d'Analyse, Topologie et Probabilités (<http://www.latp.univ-mrs.fr/>), UdP = Université de Provence ([www.univ-provence.fr](http://www.univ-provence.fr)), CRSA E13=Centre de Recherche Saint-Antoine, Equipe 13 Instabilité des microsattellites et cancers ([http://www.upmc.fr/fr/recherche/pole\\_4/pole\\_vie\\_et\\_sante/cdr\\_saint\\_antoine\\_umr\\_s\\_893.html](http://www.upmc.fr/fr/recherche/pole_4/pole_vie_et_sante/cdr_saint_antoine_umr_s_893.html)).

## 5 Scientific justification of requested budget

The total amount of the requested budget is 142 268 euros, which are spread among different posts as described below.

- **Equipment:** 15 000 euros are expected to buy a computation server. This computation server will be useful in particular to perform tasks 2 and 4. Indeed, the high dimensionnal data encountered in genetics (task 4) requires important computational resources for handling and storage. Moreover, the aim of Subtask 2 is the implementation of the developed methodology. These tasks require an efficient computer in order to run the implemented software on a high dimensional dataset.
- **Staff:** 3 600 euros are needed to remunerate the internship of two postgraduate (Master 2) students.
  - A postgraduate student will work on the implementation problems related to the developed methodology (see Subtask 2).
  - A postgraduate student will work on the problem of high dimensional data (see Subtask 9).
- **Missions:** 59 600 euros for three years are required for different missions fees.
  - 19 800 euros will be used for paying the trip fees for the members of the team to visit each other.
  - 10 800 euros will be used for invited professors in Paris. It is scheduled to invite two professors for 2 weeks each year (1800 euros is required for each guest).
  - 29 000 euros for the three years are expected to pay the conferences fees of the project members.
- **Internal expenses:** 2 000 euros are scheduled for paying internal expenses (paper, photocopies, printing or telephone fees).
- **Other expenses:** 27 000 euros will be used for other external expenses:
  - 24 000 euros for computers, laptops, printers and consumables,
  - 3 000 euros for scientific books.
- **Compensation of education services:** For each year, 10 000 euros are necessary to compensate the education services (96 HETD hours each year).

## 6 Annexes

### 6.1 Bibliography

- AALEN, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.*, **6** 701–726.
- AALEN, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wisła, 1978)*, vol. 2 of *Lecture Notes in Statist.* Springer, New York, 1–25.
- ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics, Springer-Verlag, New York.

- ANTONIADIS, A., GREGOIRE, G. and NASON (1999). Density and hazard rate estimation for right censored data using wavelet methods. *JRSS*, **61** 63–84.
- BARAUD, Y. and BIRGE, L. (2008). Estimating the intensity of a random measure by histogram type estimators. To appear in PTRF, URL <http://arxiv.org/abs/math/0608663>.
- BARAUD, Y., GIRAUD, C. and HUET, S. (2007). Gaussian model selection with unknown variance. *Annals of Statistics*.
- BARRON, A., L.BIRGÉ and P.MASSART (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113** 301–413.
- BARTLETT, P. L. and MENDELSON, S. (2006). Empirical minimization. *Probab. Theory Related Fields*, **135** 311–334.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57** 289–300.
- BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29** 1165–1188.
- BERAN, R. (1981). Nonparametric regression with randomly censored survival data. Tech. rep., University of California, Berkeley.
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2008). Simultaneous analysis of lasso and dantzig selector. To appear in AoS.
- BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, **4** 329–375.
- BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, **3** 203–268.
- BLANCHARD, G. and ROQUAIN, E. (2008). Two simple sufficient conditions for fdr control. *Electron. J. Stat.*, **2** 963–992.
- BOUAZIZ, O. and LOPEZ, O. (2008). Conditional density estimation in a censored single-index regression model. Available at <http://hal.archives-ouvertes.fr/hal-00305495/fr/>.
- BOUSQUET, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, **334** 495–500.
- BRUNEL, E. and COMTE, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, **67** 441–475.
- BRUNEL, E., COMTE, F. and LACOUR, C. (2008). Adaptive estimation of the conditional density in presence of censoring. *Sankhyā*. To appear.
- CANDES, E. and TAO, T. (????). The dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . To appear in AoS.
- CANDES, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Ann. Statist.*, **35** 2313–2351.
- CATONI, O. (2004). *Statistical learning theory and stochastic optimization*, vol. 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.
- COMTE, F., GAÏFFAS, S. and GUILLOUX, A. (2008). Adaptive estimation of the conditional intensity of marker-dependent counting processes. Available at <http://arxiv.org/abs/0810.4263>.
- COMTE, F. and ROZENHOLC, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Processes and their Applications*, **97** 111–145.

- COX, D. R. (1972a). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, **34** 187–220. With discussion.
- COX, D. R. (1972b). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, **34** 187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.
- CSÖRGŐ, S. (1996). Universal Gaussian approximations under random censorship. *Ann. Statist.*, **24** 2744–2778.
- CUCKER, F. and SMALE, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, **39** 1–49 (electronic).
- DABROWSKA, D. M. (1987). Nonparametric regression with censored survival time data. *Scand. J. Statist.*, **14** 181–197.
- DEL RIO, M., MOLINA, F., BASCOUL-MOLLEVI, C., COPOIS, V., BIBEAU, F., CHALBOS, P., BAREIL, C., KRAMAR, A., SALVETAT, N., FRASLON, C., CONSEILLER, E., GRANCI, V., LEBLANC, B., PAU, B., MARTINEAU, P. and YCHOU, M. (2007). Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. *J Clin Oncol.*, **25** 773–780.
- DELECROIX, M., HÄRDLE, W. and HRISTACHE, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.*, **86** 213–226.
- DELECROIX, M., HRISTACHE, M. and PATILEA, V. (2006). On semiparametric  $M$ -estimation in single-index regression. *J. Statist. Plann. Inference*, **136** 730–769.
- EL-BCHIRI, J., GUILLOUX, A., DARTIGUES, P., LOIRE, E., MERCIER, D., BUHARD, O., SOBHANI, I., DE LA GRANGE, P., AUBOEUF, D., PRAZ, F., FLÉJOU, J. and DUVAL, A. (2008). Nonsense-mediated mrna decay impacts msi-driven carcinogenesis and anti-tumor immunity in colorectal cancers. *PLoS ONE*, **3**.
- FISHER, L. and LIN, D. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Z. Wahrsch. Verw. Gebiete*, **20** 145–157.
- GAÏFFAS, S. and GUILLOUX, A. (2008). Learning for marker-dependent counting processes. Working paper.
- GAÏFFAS, S. and LECUÉ, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, **1** 538–573.
- GAÏFFAS, S. and LECUÉ, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electron. J. Stat.*, **1** 538–573.
- GAÏFFAS, S. and LECUÉ, G. (2008). Aggregation of penalized empirical risk minimizers in regression. Available at <http://arxiv.org/abs/0810.5288>.
- GEENENS, G. and DELECROIX, M. (2005). A survey about single-index models theory. URL <http://www.stat.ucl.ac.be/ISpub/dp/2005/dp0508.pdf>.
- GEFFRAY, S. (2006). *Estimation non-paramétrique de données censurées dans un cadre multi-états*. Ph.D. thesis, Université Pierre et Marie Curie, Paris 6.
- GEFFRAY, S. and GUILLOUX, A. (2005). Estimation in a generalized koziol-green model. *C. R. Math. Acad. Sci.*
- GEFFRAY, S. and GUILLOUX, A. (2008a). Inference for proportional cumulative incidence functions. Working paper.
- GEFFRAY, S. and GUILLOUX, A. (2008b). Inference for proportional cumulative incidence functions. Working paper.
- GENDRE, X. (2008). Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression. *Electronic Journal of Statistics*.
- GILL, R. D. (1980). Nonparametric estimation based on censored observations of a Markov renewal process. *Z. Wahrsch. Verw. Gebiete*, **53** 97–116.

- GORDON, Y., LITVAK, A. E., MENDELSON, S. and PAJOR, A. (2007). Gaussian averages of interpolated bodies and applications to approximate reconstruction. *J. Approx. Theory*, **149** 59–73.
- GRÉGOIRE, G. (1993). Least squares cross-validation for counting process intensities. *Scand. J. Statist.*, **20** 343–360.
- GUÉDON, O., MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2007). Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Positivity*, **11** 269–283.
- GUILLOUX, A. and SAINT-PIERRE, P. (2008). Estimateur de la fonction de répartition bivariée avec censures à droite et à gauche. *Annales de l'ISUP*.
- HAUSSLER, D. (1992). Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. and Comput.*, **100** 78–150.
- HRISTACHE, M., JUDITSKY, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, **29** 595–623.
- HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.*, **27** 1536–1563.
- HUET, S. (2006). Model selection for estimating the non zero components of a Gaussian vector. *ESAIM Probab. Stat.*, **10** 164–183.
- JACOBSEN, M. (1982). *Statistical analysis of counting processes*, vol. 12 of *Lecture Notes in Statistics*. Springer-Verlag, New York.
- JACOD, J. and SHIRYAEV, A. N. (1987). *Limit theorems for stochastic processes*, vol. 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.
- JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A. (2005a). Learning by mirror averaging. URL <http://arxiv.org/abs/math/0511468>.
- JUDITSKY, A. B., NAZIN, A. V., TSYBAKOV, A. B. and VAYATIS, N. (2005b). Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, **41** 78–96.
- KEARNS, M. J., SCHAPIRE, R. E., SELLIE, L. M. and HELLERSTEIN, L. (1994). Toward efficient agnostic learning. In *Machine Learning*. ACM Press, 341–352.
- KLARTAG, B. and MENDELSON, S. (2005). Empirical processes and random projections. *J. Funct. Anal.*, **225** 229–245.
- LECUÉ, G. (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, **13** 1000–1022.
- LECUÉ, G. and MENDELSON, S. (2008). Aggregation via empirical risk minimization. To appear in *Probability theory and related fields*.
- LEDoux, M. (2001). *The concentration of measure phenomenon*, vol. 89 of *Mathematical Surveys and Monographs*. American Mathematical Society.
- LEDoux, M. and TALAGRAND, M. (1991). *Probability in Banach spaces*, vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.
- LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, **52** 3396–3410.
- LI, G. and DOSS, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.*, **23** 787–823.
- LI, Q. and LAGAKOS, S. (1997). Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Statist. Med.*, **16** 925–940.
- LIN, D., SUN, W. and YING, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, **86** 59–70.

- LINTON, O. B., NIELSEN, J. P. and VAN DE GEER, S. (2003). Estimating multiplicative and additive hazard functions by kernel methods. *Ann. Statist.*, **31** 464–492. Dedicated to the memory of Herbert E. Robbins.
- LIPTSER, R. S. and SHIRYAYEV, A. N. (1989). *Theory of martingales*, vol. 49 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht. Translated from the Russian by K. Dzjaparidze [Kacha Dzjaparidze].
- LOPEZ, O. (2008). Single-index regression model with censored responses. *J. Statist. Plann. Inference* (In press) doi:10.1016/j.jspi.2008.06.012.
- MALLOWS, C. (1973). Some comments on  $c_p$ . *Technometrics*, **15** 661–675.
- MASSART, P. (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- MASSART, P. and NÉDÉLEC, É. (2006). Risk bounds for statistical learning. *Ann. Statist.*, **34** 2326–2366.
- MCKEAGUE, I. W. and UTIKAL, K. J. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.*, **18** 1172–1187.
- MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70** 53–71.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34** 1436–1462.
- MENDELSON, S. (2008). Obtaining fast error rates in nonconvex situations. *J. Complexity*, **24** 380–397.
- MENDELSON, S. and LECUÉ, G. (2008). Sharper lower bound for the. Sharper lower bounds on the performance of the Empirical Risk Minimization Algorithm.
- MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, **17** 1248–1282.
- NEMIROVSKI, A. (2000). *Topics in Non-Parametric Statistics*. Ecole d’été de probabilités de Saint-Flour XXVIII - 1998. Lecture Notes in Mathematics, no. 1738, Springer, New York.
- PRUM, B., RODOLPHE, F. and TURCKHEIM, E. (1995). Finding words with unexpected frequencies in dna sequences. *J. R. Statist. Soc. B*, **57** 205–220.
- RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11** 453–466.
- REYNAUD-BOURET, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, **126** 103–153.
- REYNAUD-BOURET, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, **12** 633–661.
- ROBIN, S., RODOLPHE, F. and SCHBATH, S. (2003). *DNA, Words and models*. Cambridge University Press.
- ROQUAIN, E. (2007). *Motifs exceptionnels dans des séquences hétérogènes. Contributions à la théorie et à la méthodologie des tests multiples*. Ph.D. thesis, Université Paris XI.
- S. MENDELSON, J. N. (2008). Regularization in kernel learning.
- SPOKOINY, V. (2008). Multiscale local change point detection with application to value-at-risk. Available at <http://www.e-publications.org/ims/submission/index.php/AOS/user/submissionFile/2986?confirm=7c329402>.
- STEINWART, I. and SCOVEL, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, **35** 575–607.

- STUTE, W. (1994). Strong and weak representations of cumulative hazard function and Kaplan-Meier estimators on increasing sets. *J. Statist. Plann. Inference*, **42** 315–329.
- STUTE, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, **23** 461–471.
- TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, **126** 505–563.
- TARIGAN, B. and VAN DE GEER, S. A. (2006). Classifiers of support vector machine type with  $l_1$  complexity regularization. *Bernoulli*, **12** 1045–1076.
- THERNEAU, T. M. and GRAMBSCH, P. M. (2000). *Modeling survival data: extending the Cox model*. Statistics for Biology and Health, Springer-Verlag, New York.
- TIBSHIRANI, R. (1997). The lasso method for variable selection in the cox model. *Statist. in Med.*, **16** 385–395.
- TSYBAKOV, A. B. (2003). Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, **2777** 303–313. Springer, Heidelberg.
- VAN DE GEER, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, **23** 1779–1801.
- VAN DE GEER, S. (2007). Oracle inequalities and regularization. In *Lectures on empirical processes*. EMS Ser. Lect. Math., Eur. Math. Soc., Zürich, 191–252.
- VAN DE GEER, S. A. (2000). *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
- VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.*, **36** 614–645.
- VAPNIK, V. N. (1998). *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control, John Wiley & Sons Inc., New York. , A Wiley-Interscience Publication.
- VAPNIK, V. N. (2000). *The nature of statistical learning theory*. 2nd ed. Statistics for Engineering and Information Science, Springer-Verlag, New York.
- VILLERS, F. (2007). *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. Ph.D. thesis, Université Paris XI.
- WAINWRIGHT, M., RAVIKUMAR, P. and LAFFERTY, J. (2006). High-dimensional graphical model selection using  $l_1$ -regularized logistic regression. **19**. Advances in Neural Information Processing Systems (NIPS).
- XIA, Y. and HÄRDLE, W. (2006). Semi-parametric estimation of partially linear single-index models 1162–1184.
- YANG, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.*, **28** 75–87.
- YANG, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, **96** 574–588.
- YANG, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, **10** 25–47.
- ZAAANAN, A., CUILLIERE-DARTIGUES, P., GUILLOUX, A., PARC, Y., LOUVET, C., DE GRAMONT, A., TIRET, E., DUMONT, S., GAYET, B., VALIDIRE, P., FLÉJOU, J.-F., DUVAL, A. and PRAZ, F. (2008). Impact of p53 expression and microsatellite instability on stage iii colon cancer disease-free survival in patients treated by fluorouracil and leucovorin with or without oxaliplatin. *Submitted*.
- ZOU, H. and HASTIE, T. (2005). Addendum: “Regularization and variable selection via the elastic net” [J. R. Stat. Soc. Ser. B Stat. Methodol. **67** (2005), no. 2, 301–320; mr2137327]. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67** 768.

## 6.2 CV of the members

## BIAU Gérard

35 ans

Full Professor

Université Pierre et Marie Curie - Paris 6

email : [gerard.biau@upmc.fr](mailto:gerard.biau@upmc.fr)

www : <http://www.lsta.upmc.fr/biau.html>

### Work address

LSTA, Équipe d'Accueil 3124

Université Pierre et Marie Curie - Paris 6

8ème étage, plateau A, 175 rue du Chevaleret

75013 Paris

Tel : 33 (0) 1 44 27 85 63

## Education

---

- Civil engineer, 1997, Ecole des Mines, Paris, France;
- Agrégation de mathématiques, 2000;
- Ph.D., 2001, Statistics, Université Montpellier II, France.

## Research Interests

---

Nonparametric statistics; Statistical learning; Supervised and unsupervised classification; clustering; High-dimensional statistics.

## Appointments

---

- 2007- : Professeur of Mathematics, Université Pierre et Marie Curie — Paris VI.
- 2004-2007 : Professeur of Mathematics, Université Montpellier II.
- 2001-2004 : Maître de Conférences, Université Pierre et Marie-Curie — Paris VI.

## Memberships

---

- Société Française de Statistique (SFdS, I serve as the Vice-President of the association since 2006).

## List of 5 publications

---

- BIAU, G., BUNEA, F., AND WEGKAMP, M.H. (2005), Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, **Vol. 51**, pp. 2163-2172.
- BIAU, G., AND GYÖRFI, L. (2005), On the asymptotic properties of a nonparametric  $L_1$ -test statistic of homogeneity. *IEEE Transactions on Information Theory*, **Vol. 51**, pp. 3965-3973.
- BLEAKLEY, K., BIAU, G., AND VERT, J.-P. (2007), Supervised reconstruction of biological networks with local models, *Bioinformatics*, **Vol. 23**, pp. i57-165.
- BIAU, G., DEVROYE, L., AND LUGOSI, G. (2008), On the performance of clustering in Hilbert spaces, *IEEE Transactions on Information Theory*, **Vol. 54**, pp. 781-790.
- BIAU, G., DEVROYE, L., AND LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers, *Journal of Machine Learning Research*, In press.

**BOUAZIZ Olivier**  
PhD student  
Université Pierre et Marie Curie - Paris 6

email : [olivier.bouaziz@upmc.fr](mailto:olivier.bouaziz@upmc.fr)  
www : <http://www.lsta.upmc.fr>

**Address**  
LSTA, Équipe d'Accueil 3124  
Université Pierre et Marie Curie - Paris 6  
8ème étage, plateau A, 175 rue du Chevaleret  
75013 Paris

tel : 33 (0) 1 44 27 85 66

## Education

---

- Since 2006 : PhD student LSTA-University of Paris 6.  
Censored regression models : an alternative to the Cox model. Under the direction of Pr. Michel Delecroix.
- 2006 : Master in Mathematics and their Applications, Speciality Statistics, University Paris 6.

## Publications

---

- BOUAZIZ O., LOPEZ O. (2008). Conditional density estimation in a censored single-index regression model. *Submitted*.
- BOUAZIZ O., GEFFRAY S., LOPEZ O. (2008). Semiparametric inference for the recurrent event process through a single-index model. *Work in progress*.

## Workshop and conference

---

- Séminaire de Probabilités et Statistique, Université Montpellier II, France, October 6th 2008.
- International Workshop on Applied Probability, Compiègne, France, July 7-10 2008.
- 36th conference of the SSC and 40th conference of the SFDS, Ottawa, Canada, May 25-29 2008.
- Groupe de Travail des Thésards (GTT) du LSTA, Paris, France, 2006 and 2008.

## Teaching experience

---

“Mathematics for economists” for 1st year students (L1) in economics and management discipline, Paris 12 (2006-2008).

## Duval Alex

41 ans

Directeur de Recherche INSERM, MD, PhD  
Directeur de l'équipe 13 - " Instabilité des Microsatellites et Cancers " Centre de Recherche Saint-Antoine (UMRS 893)  
Université Pierre et Marie Curie - Paris 6

email : [alex.duval@inserm.fr](mailto:alex.duval@inserm.fr)

### Working adress

Centre de Recherche Saint-Antoine, équipe 13  
Hôpital Saint Antoine - Bat. Kourilsky  
Université Pierre et Marie Curie - Paris 6  
184 rue du faubourg Saint Antoine  
75571 Paris Cedex 12

Tel : 33 (0) 1 53 72 51 20

---

## Academics, positions

- 2008 : Director of the 13rd team "Instabilité des Microsatellites et Cancers" of UMR S 893.
- 2006-2007: Director of the INSERM U762 "Instabilité des Microsatellites et Cancers".
- 2005: Habilitation à Diriger des Recherches (HDR).
- 2002-2005: MD, service d'hématologie - cytogénétique (CHU de Bicêtre).
- 2002-2005: Research assistant (Chargé de Recherches (CR1)) INSERM à l'U434 (Director: Gilles THOMAS).
- 1999-2002: PhD and post-doctorate in Sciences (U 434).

---

## List of 5 publications

- EL-BCHIRI J, GUILLOUX A, DARTIGUES P, LOIRE E, MERCIER D, BUHARD O, SOBHANI I, DE LA GRANGE P, AUBOEUF D, PRAZ F, FLÉJOU JF, DUVAL A. (2008). Nonsense-mediated mRNA decay impacts MSI-driven carcinogenesis and antitumor immunity in colorectal cancers. *PLoS One*, 9(3) : e2583.
- SVRCEK M, EL-BCHIRI J, CHALASTANIS A, CAPEL E, DUMONT S, BUHARD O, OLIVEIRA C, SERUCA R, BOSSARD C, MOSNIER JF, BERGER F, LETEURTRE E, LAVERGNE-SLOVE A, CHENARD MP, HAMELIN R, COSNES J, BEAUGERIE L, TIRET E, DUVAL A AND FLEJOU JF. (2007). Specific clinical and biological features characterize inflammatory bowel disease associated colorectal cancers showing microsatellite instability. *J Clin Oncol*. 25: 4231-8.
- DUVAL A, RAPHAEL M, BRENNETOT C, POIREL H, BUHARD O ET AL. (2004). The mutator pathway is a feature of immunodeficiency-related lymphomas. *Proc. Nat. Acad. Sci. USA* 101(14) : 5002-07.
- BRENNETOT C, PINTO M, OLIVEIRA C, SCHWARTZ JR S, SERUCA R, DUVAL A, HAMELIN R. (2003). Frequent Ki-ras mutations in gastric tumors of the MSI phenotype. *Gastroenterology* 125 : 1282-83.
- DUVAL A, IACOPETTA B, THORSTENSEN L, MELING GI, LOTHE RA, THUILLE B, SURAWEEERA N, THOMAS G, HAMELIN R. (2001). Gender difference for mismatch repair deficiency in human colorectal cancer. *Gastroenterology* 121: 1026-1027.

---

## Contracts

- 2008 : Extension of contract d'Interface with the Assistance Publique des Hôpitaux de Paris (2 years).
- 2008 : Institut Clinique de la Souris, Strasbourg (20 KE) : establishment of a transgenic mouse model for Upf1 gene.
- 2008 : Ligue Nationale Contre le Cancer and Institut National du Cancer - Tumors Identity Card Project (CIT3): a coupled analysis of the genome and transcriptome for colorectal cancer based on their genetic heterogeneity and the tumor micro-environment. (> 200 KE).
- 2007 : Association pour la Recherche contre le Cancer (50 KE) : Role of NMD in carcinogenesis MSI tumors.
- 2006 : Prix Jean and Madeleine Schaefferberke, Fondation de France (12 KE).
- 2005-2008 : Contract d'Interface with the Assistance Publique des Hôpitaux de Paris (3 ans).

## GAÏFFAS Stéphane

29 ans

Maître de conférences

University Pierre et Marie Curie - Paris 6

email : [stephane.gaiffas@upmc.fr](mailto:stephane.gaiffas@upmc.fr)

www : <http://www.lsta.upmc.fr/gaiffas.php>

### Working adress

LSTA, Équipe d'Accueil 3124  
University Pierre et Marie Curie - Paris 6  
8ème étage, plateau A, 175 rue du Chevaleret  
75013 Paris, France

Tel : 33 (0) 1 44 27 70 48

## Positions

---

Since fev. 2007	<b>Maître de Conférences</b> at University Pierre et Marie Curie – Paris 6.
2006 – 2007	<b>ATER</b> at University Paris X – Nanterre.
2005 – 2006	<b>ATER</b> University Paris VII

## Education

---

2002 – 2005	<b>Ph.D.</b> , Statistics, University Paris VII.
-------------	--

## List of 5 publications

---

- S. GAÏFFAS (2009), **Global estimation of a signal based on inhomogeneous data**, *Statistica Sinica*, in press.
- S. GAÏFFAS, G. LECUÉ (2007), **Optimal rates and adaptation in the single-index model using aggregation**, *Electronic Journal of Statistics*, Volume 1 (2007), p. 538-573.
- S. GAÏFFAS (2006), **Sharp estimation in sup norm with random design**, *Statistics and Probability Letters*, Volume 77 (2006), issue 8, p. 782-794.
- S. GAÏFFAS (2006), **On pointwise adaptive curve estimation based on inhomogeneous data**, *ESAIM P&S*, Volume 11 (2007), p. 344-364.
- S. GAÏFFAS (2005), **Convergence rates for pointwise curve estimation with a degenerate design**, publié dans *Mathematical Methods of Statistics*, Volume 14 (2005), No. 1, 1–27.

## Reviewer activity

---

Referee for *Annals of Statistics*, *Bernoulli*, *Metrika*, *Electronic Journal of Statistics*, *Journal of Applied Statistics*, *Journal of Multivariate Analysis*.

# GEFFRAY Ségolen

29 ans

Maître de conférences  
Université de Nantes

email : [segolen.geffray@univ-nantes.fr](mailto:segolen.geffray@univ-nantes.fr)

## Work adress

Laboratoire de Mathématiques Jean Leray  
Université de Nantes  
1 rue Gaston Veil  
44035 Nantes cedex 01

## Current position

---

- Since sept. 2007: Assistant Professor (Maître de Conférences), University of Nantes.
- Since sept. 2007: Member of Laboratoire de Mathématiques Jean Leray, University of Nantes.

## Formation

---

- 2007: D.U. Cell and gene therapy, University Montpellier I
- 2007: Ph.D., Statistics, University Paris VI, France.

## Research Interests

---

Survival analysis, recurrent events, competitive risks, semi-parametric and non-parametric statistics, Cellular and Molecular Medicine.

## List of publications

---

- GEFFRAY, S. (2008). Strong approximations for dependent competing risks with independent censoring with statistical applications, *TEST*, DOI: 10.1007/s11749-008-0113-y.
- GEFFRAY, S. (2007). Estimation dans un modèle de risques concurrents éventuellement dépendants en présence de censure, *C. R. Acad. Sci. Paris, Ser. I*, 344: 457-460.
- GEFFRAY, S., GUILLOUX, A. (2005). Estimation dans un modèle de Koziol-Green généralisé, *C. R. Acad. Sci. Paris, Ser. I*, 341: 49-52.

## Submitted papers

---

- GEFFRAY, S., GUILLOUX, A. Inference for proportional cumulative incidence functions.
- GEFFRAY, S. Comparison of confidence bands for cumulative incidence functions of possibly dependent competing risks with independent censoring.

**GUILLOUX Agathe**  
Maître de conférences  
Université Pierre et Marie Curie - Paris 6

email : [agathe.guilloux@upmc.fr](mailto:agathe.guilloux@upmc.fr)  
www : <http://www.lsta.upmc.fr/guilloux.php>

**Adresse professionnelle**  
LSTA, Équipe d'Accueil 3124  
Université Pierre et Marie Curie - Paris 6  
8ème étage, plateau A, 175 rue du Chevaleret  
75013 Paris

tél bureau : 33 (0) 1 44 27 70 48

---

## Position, affiliations and administrative activities

---

- Assistant professor (Maître de Conférences) at l'Université Pierre et Marie Curie – Paris 6 (since sep. 2005)
- Member of LSTA (Laboratoire de Statistiques Théoriques et Appliquées) (Director Pr. Deheuvels).
- Member of the Centre de Recherche Saint-Antoine (UMR S 893) Equipe 13 - Université Pierre et Marie Curie - Paris 6 (Director Alex Duval).
- Vice-president of the Commission des Spécialistes de la 26-ième Section de l'UPMC (2007-2008).
- Elected member of the Board of ISUP -UPMC.
- Member of the Comity of ISUP -UPMC.
- Membre de la Commission d'Avancement des Maîtres de Conférences de l'UPMC

---

## Education

---

- **2001-2004** PhD in Statistics at the University of Rennes 1
- **2001** Master in Biostatistics and Epidemiology and Master in Statistics

---

## List of 5 Publications

---

- BRUNEL E., COMTE F., GUILLOUX A. (200?). Nonparametric density estimation in presence of bias and censoring. To appear in *Test* and Prépublication du MAP5 2005-22.
- EL-BCHIRI J, GUILLOUX A, DARTIGUES P, LOIRE E, MERCIER D, BUHARD O, SOBHANI I, DE LA GRANGE P, AUBOEUF D, PRAZ F, FLÉJOU JF, DUVAL A. (2008). Free in PMC Nonsense-mediated mRNA decay impacts MSI-driven carcinogenesis and anti-tumor immunity in colorectal cancers. *PLoS ONE*.9 no.3(7):e2583.
- BRUNEL E., COMTE F., GUILLOUX A. (2008). Adaptive nonparametric strategies for censored lifetimes with unknown sampling bias. *Scandinavian Journal of Statistics* 35, no. 3, p 557–576.
- DAUXOIS J.-Y. AND GUILLOUX A. (2008). Nonparametric inference under competing risks and selection bias sampling. *Journal of Multivariate Analysis* 99, no. 4, 589–605.
- GUILLOUX A. (2007). Non-parametric estimation for censored lifetimes suffering from an unknown selection bias. *Mathematical Methods in Statistics* 16, no. 3, 202–216.

---

## Supervisions

---

- Co-supervision of Ségolen Geffray's PhD (Director: Paul Deheuvels, LSTA -UPMC) and member of the comity.
- Supervision of the TER of Noury Sin (graduate student at UPMC, 2008).

## LECUE Guillaume

29 ans  
CNRS researcher  
Laboratoire d'analyse, topologie et probabilité  
Marseille

email : [lecue@latp.univ-mrs.fr](mailto:lecue@latp.univ-mrs.fr)

**Work adress**  
Bureau 135, LAMP, CMI  
Technopôle Château-Gombert  
39, rue F. Joliot Curie  
13453 Marseille Cedex 13, France

Tel : 33 (0) 4 91 11 35 58

## Current position

---

CNRS researcher at the Laboratoire d'analyse, topologie et probabilité, LAMP, Marseille.

## Formation

---

- May 2007: CNRS researcher Position.
- Sep. 2004-May 2007: PhD student and teaching assistant. PhD supervisor: Pr. Tsybakov. Laboratoire de Probabilités et Modèles aléatoires. Univ. Paris 6.
- 2003-2004: Master in stochastic modeling and statistics, (rank: 3-rd).  
**Forth year at ENS Cachan.** *Université Paris-Sud, Centre Scientifique d'Orsay.*
- 2002-2003: Agrégation in Mathematics, rank: 43-th  
**Third year at ENS Cachan.** *Ker Lann, Université de Rennes 1.*

## Invitations

---

- Sept.-Nov. 2008: Seminar für Statistik, ETH, Zürich, Switzerland. Invited by Pr. van de Geer.
- May.-June. 2008: Department of mathematics of Technion, Haïfa, Israël. Invited by Pr. Mendelson.
- Feb.-March 2008: Departamento de Estadística, Universidad de Valparaiso, Chile. Invited by Dr. Bertin.
- Oct.-Dec. 2007: Department of mathematics of Technion, Haïfa, Israël. Invited by Pr. Mendelson.
- July-Oct. 2007: Mathematical Science Institute, Australian National University, Canberra, Australia. Invited by Pr. Mendelson.
- March 2006: Department of Statistics of University of California, UC Berkeley. Invited by Pr. Bickel.
- April 2006: Department of Mathematics of the Institute of Technology of Georgia, Atlanta, USA. Invited by Pr. Koltchinski.

## List of 5 publications

---

- G. LECUÉ. Simultaneous adaptation to the margin and to complexity in classification. *The Annals of Statistics*, 2007, Vol. 35, No. 4, 1698-1721.
- G. LECUÉ. Optimal oracle inequality for aggregation of classifiers under low noise condition. 19th Annual Conference on Learning Theory, COLT06. Proceedings. Gabor Lugosi, Hans Ulrich Simon (Eds.). Springer. LNAI 2006, 364-378. "Mark Fulk award" for "the best student paper".
- G. LECUÉ. Optimal rates of aggregation in classification. *Bernoulli*, 2007, Vol. 13, No. 4, 1000-1022.
- S. GAÏFFAS AND G. LECUÉ. Minimax rates and adaptivity in the Single-Index Model. *Electronic Journal of Statistics*, Vol. 1, pp. 538-573. 2007
- G. LECUÉ AND S. MENDELSON Aggregation Via Empirical Risk Minimization. To appear in *Probability Theory and related fields*.
- Total number of published paper: 5. Paper to appear: 2. Paper submitted: 3. Conference with lecture committee: 2.

**LOPEZ Olivier**  
28 ans  
Maître de conférences  
Université Pierre et Marie Curie - Paris 6

email : [olivier.lopez0@upmc.fr](mailto:olivier.lopez0@upmc.fr)

**Work adress**  
LSTA, Équipe d'Accueil 3124  
Université Pierre et Marie Curie - Paris 6  
8ème étage, plateau A, 175 rue du Chevaleret  
75013 Paris

Tel : 33 (0) 1 44 27 33 53

## Positions

---

since sept. 2008	<b>Maître de Conférences</b> at Université Pierre et Marie Curie – Paris 6
Feb. – aug 2008	<b>Post-doctoral Fellowship</b> of Weierstrass Institute für Angewandte Analysis und Stochastik – Berlin, Allemagne
sept. 2005 – janv. 2008	<b>PhD.</b> in Université Rennes I and Ensai
2001 – 2005	<b>Elève normalien</b> in ENS Cachan (Antenne de Bretagne)

## List of publications

---

- O. LOPEZ & V. PATILEA (2009), **Nonparametric lack-of-fit tests for parametric mean-regression models with censored data**, *Journal of Multiv. Analysis*, Volume 100, Issue 1, January 2009, pages 210–230.
- O. LOPEZ (2008), **Single-index regression model with censored responses**, *Journal of Statis. Plan. and Inference*, in press.
- M. DELECROIX, O. LOPEZ & V. PATILEA (2008), **Nonlinear Censored Regression Using Synthetic Data**, *Scand. Journal of Statist.*, Volume 35, Number 2, June 2008 , pages 248–265(18).
- O. LOPEZ (2007), **Réduction de dimension en présence de données censurées**, PhD Thesis under the direction of M. Delecroix and V. Patilea, jury composed by P. Bertail, B. Delyon, D. Picard, W. Stute. <http://tel.archives-ouvertes.fr/tel-00195261/fr/>

## Submitted papers

---

- O. BOUAZIZ & O. LOPEZ (2008), **Conditional density estimation in a censored single-index regression model** <http://hal.archives-ouvertes.fr/hal-00305495/fr/>
- O. LOPEZ (2007), **On the estimation of the joint distribution in a censored regression model**, <http://www.crest.fr/doctravail/document/11.pdf>

## ROQUAIN Etienne

27 ans

Maître de conférences

University Pierre et Marie Curie - Paris 6

email : [etienne.roquain@upmc.fr](mailto:etienne.roquain@upmc.fr)

www : <http://www.proba.jussieu.fr/pageperso/roquain/>

### Working adress

Laboratoire de Probabilités et Modèles Aléatoires (LPMA)

University Pierre et Marie Curie - Paris 6

175 rue du Chevaleret

75013 Paris

Tel: 33 (0) 1 44 27 70 44

## Current position

---

- Since sept. 2008: Assistant Professor (Maître de Conférences), University Pierre et Marie Curie, Paris 6.
- Since sept. 2008: Member of the laboratoire de Probabilités et Modèles Aléatoires (LPMA).

## Formation

---

- 2002-2003 : Agrégation de Mathématiques.
- 2007 : Ph.D., Statistics, University Paris Sud, France.
- 2007-2008 : Post-doctoral Fellowship, Vrije Universiteit, Amsterdam.

## List of 5 publications

---

- ROQUAIN, E., SCHBATH, S. (2007). Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov Chain. *Advances in Applied Probability*. Volume 39, pages 128-140.
- BLANCHARD, G., ROQUAIN, E. (2008). Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*. Volume 2, pages 963-992.
- ARLOT, S., BLANCHARD, G. ET ROQUAIN, E. (2008). Some non-asymptotic results on resampling in high dimension, I: Confidence regions. *Accepted for publication in the Annals of Statistics*.
- ARLOT, S., BLANCHARD, G. ET ROQUAIN, E. (2008). Some non-asymptotic results on resampling in high dimension, II: Multiple tests. *Accepted for publication in the Annals of Statistics*.
- BLANCHARD, G., ROQUAIN, E. (2008). Adaptive FDR control under independence and dependence. *Submitted*. ArXiv : 0707.0536

## Reviewer activity

---

- Referee for *Electronic Journal of Statistics*.
- Referee for *Journal of Machine Learning Research*.

**SAINT PIERRE Philippe**  
30 ans  
Maître de conférences  
Université Pierre et Marie Curie - Paris 6

email : [philippe.saint\\_pierre@upmc.fr](mailto:philippe.saint_pierre@upmc.fr)  
www : <http://www.lsta.upmc.fr/psp.html>

**Work adress**  
LSTA, Équipe d'Accueil 3124  
Université Pierre et Marie Curie - Paris 6  
8ème étage, plateau A, 175 rue du Chevaleret  
75013 Paris

Tel: 33 (0) 1 44 27 33 43

---

## Position

---

- Since sept. 2006: Assistant Professor (Maître de Conférences), University Pierre et Marie Curie – Paris 6.
- Since sept. 2006: Member of Laboratoire de Statistique Théorique et Appliquée (Director Pr. Dehevels).

---

## Education

---

- 2005: Ph.D., Statistics, Université Montpellier I, France.

---

## Administrative activities

---

- Member (alternate) of the specialist commission, 26th section, UPMC.
- Elected member of the administrative board of ISUP -UPMC.
- Member of the committee of ISUP -UPMC.

---

## Research Interests

---

Markov and semi-Markov multi-states models, Survival analysis, Informative censoring, Application to asthma.

---

## List of 5 publications

---

- GUILLOUX A., SAINT-PIERRE P.(2008). Estimateur de la fonction de répartition bivariée avec censures à droite et à gauche. *Annales de l'ISUP*, 19(2):157 - 164.
- FOUCHER Y., SAINT-PIERRE P., PUGLIESE P., DELLAMONICA P., DAURÈS J.P. (2006). A semi-Markov frailty model for multistate survival data : illustration on HIV disease. *Far East Journal of Theoretical Statistics*, 19(2):185 - 201.
- FOUCHER Y., MATHIEU E., SAINT-PIERRE P., DURAND J.F., DAURÈS J.P. (2006). A semi-Markov model based on generalized weibull distribution with an illustration for HIV disease. *Biometrical Journal*, 47(6):825-833.
- SAINT-PIERRE P., BOURDIN A., CHANEZ P., DAURÈS J.P., GODARD P. (2006). Are overweighted asthmatics more difficult to control? *Allergy*, 61(1):79-84.
- SAINT-PIERRE P., COMBESCURE C., DAURÈS J.P., GODARD P. (2003). The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine* 22(24):3755-70.

---

## Scientific award

---

- May 2005: Henri Philippart epidemiology prize awarded by InVS (Institut de Veille Sanitaire) and ADELFF (Association Des Epidémiologistes De Langue Française)

## VILLERS Fanny

27 ans

Maître de conférences

Université Pierre et Marie Curie - Paris 6

email : [fanny.villers@upmc.fr](mailto:fanny.villers@upmc.fr)

### Work adress

laboratoire de Probabilités et Modèles Aléatoires (LPMA)

UMR 7599

Université Pierre et Marie Curie - Paris 6

175 rue du Chevaleret

75013 Paris

## Current position

---

- Since sept. 2008: Assistant Professor (Maître de Conférences), University Pierre et Marie Curie, Paris 6.
- Since sept. 2008: Member of the laboratoire de Probabilités et Modèles Aléatoires (LPMA).

## Formation

---

- 2002-2003 : Agrégation de Mathématiques
- 2007 : Ph.D., Statistics, University Paris Sud, France.

## List of publications

---

- N. VERZELEN, F. VILLERS. (2008). Goodness-of-fit Tests for high-dimensional Gaussian linear models. *The Annals of Statistics*, to appear. arXiv:0711.2119v4.
- N. VERZELEN, F. VILLERS. (2008). Tests for Gaussian graphical models. *Computational Statistics and Data Analysis*, special issue : *Statistical Genetics and Statistical Genomics*, to appear, hal-00193268.
- F. VILLERS, B. SCHAEFFER, C. BERTIN, S. HUET. (2008). Assessing the validity domains of graphical gaussian models in order to infer relationships among components of complex biological systems. *Statistical Applications in Genetics and Molecular Biology*, volume 7, issue 2, Article 14.
- J.F. CHICH, O. DAVID, F. VILLERS, B. SCHAEFFER, D. LUTOMSKI, S. HUET. (2007). Statistics for proteomics: Experimental design and 2-DE differential analysis. *Journal of Chromatography B*, volume 849, pages 261-272.

### 6.3 Involvement of project participants to other grants

Name	Person.Month	Type of Grant	Name project	PI	Begin/End
G. BIAU	12	ANR Blanc	CLARA	B. Pelletier	2009/2013
A. DUVAL	24	PFF	(1)	A. DUVAL	2009/2013
A. DUVAL	9.6	APHP	(2)	A. DUVAL	2009/2013
A. DUVAL	9.6	LNCC	(3)	A. DUVAL	2009/2013
G. LECUÉ	12	ANR Blanc	PARCIMONIE	E. Le Pennec	2009/2013
F. VILLERS	10.8	ANR Jeune	DETECT	S. Arlot	2009/2012
	9.6	ANR Blanc	PARCIMONIE	E. Le Pennec	2009/2013
E. ROQUAIN	14.4	ANR Jeune	DETECT	S. Arlot	2009/2012
	14.4	ANR Blanc	PARCIMONIE	E. Le Pennec	2009/2013

Table 2: Involvement of project participants to other grants, where:

PFF = Projet Fondation de France,

(1) = "Déficiency en réparation des erreurs de réplication et oncogénèse : conséquences fonctionnelles de l'instabilité des répétitions microsatellites codantes",

APHP = Contrat d'Interface APHP,

(2) = "Immunosuppression, inflammation et oncogénèse : étude du rôle de l'instabilité des microsatellites",

LNCC = Ligue Nationale contre le Cancer,

(3) = projet "Carte d'identité des tumeurs"