INRIA, Evaluation of Theme
Observation, Modeling, and Control for Life Sciences

Project-team BIGS
2013

Project-team title: BIGS : Biology, genetics and statistics.
Scientific leader: Samy Tindel
Research center: INRIA Nancy Grand Est

1 Personnel

Personnel (2011)

<table>
<thead>
<tr>
<th></th>
<th>Misc.</th>
<th>INRIA</th>
<th>CNRS</th>
<th>University</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR / Professors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>CR / Assistant Professor</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>PhD Students</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Post-Doc.</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2</td>
<td>14</td>
<td></td>
<td></td>
<td>16</td>
</tr>
</tbody>
</table>

(1) “Senior Research Scientist (Directeur de Recherche)”
(2) “Junior Research Scientist (Chargé de Recherche)”
(3) “Civil servant (CNRS, INRIA, ...)”
(4) “Associated with a contract (Ingénieur Expert or Ingénieur Associé)”

Personnel (2013)

<table>
<thead>
<tr>
<th></th>
<th>Misc.</th>
<th>INRIA</th>
<th>CNRS</th>
<th>University</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR / Professors</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>CR / Assistant Professor</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>PhD Students</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Post-Doc.</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2</td>
<td>1</td>
<td>11</td>
<td></td>
<td>14</td>
</tr>
</tbody>
</table>

Changes in staff

<table>
<thead>
<tr>
<th>DR / Professors</th>
<th>Misc.</th>
<th>INRIA</th>
<th>CNRS</th>
<th>University</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR / Assistant Professors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inria Starting Positions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arrival</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td><strong>Leaving</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Note: During the evaluation period, T. Bastogne obtained Professorship in his institution. His activity in our team has been one of the elements which lead to this decision.
Current composition of the project-team (09/13):

Faculty Members
- Samy Tindel (Team leader, Professor, Université de Lorraine, Hdr)
- Thierry Bastogne (Professor, Université de Lorraine, Hdr)
- Sandie Ferrigno (Associate Professor, Université de Lorraine)
- Céline Lacaux (Associate Professor, Université de Lorraine, Hdr)
- Jean-Marie Monnez (Professor, Université de Lorraine, Hdr)
- Aurélie Muller-Gueudin (Associate Professor, Université de Lorraine)
- Pierre Vallois (Professor, Université de Lorraine, Hdr)
- Sophie Wantz-Mézières (Associate Professor, Université de Lorraine)

Administrative Assistant
- Sophie Drouot (shared with Score, Tosca, Véridis), until 2012
- Isabelle Blanchard, since 2012

Others
- Romain Bar (PhD student, IECL)
- Benoît Lalloué (PhD student, EHESP-IECL)
- Jean-Louis Marchand (Postdoc Fellowship, IECL)
- Laura Vinckenbosch (Postdoc Fellowship, INRIA-IECL)
- Juan Víquez (Postdoc Fellowship, INRIA-IECL)

Current position of former project-team members (including PhD students during the 2011-2013 period):

Provide a list of former project-team members including name, current position, name and location of the employer. TBR.

- Alexandra Chronopoulou, Visiting Assistant Professor, University of California at Santa Barbara.
- Mireia Besalú, Assistant Professor, Universitat Pompeu Fabra, Barcelona.
- Rémi Bonidal, Statistical Engineer, Advanced Perfusion Diagnostic (Lyon).

2 Work progress

2.1 Keywords

2.2 Context and overall goal of the project

BIGS is a team labeled by INRIA, by CNRS and by Université de Lorraine, via the Institut Élie Cartan of Nancy (UMR 7502 CNRS-INRIA-UL). Our research is mainly focused on statistics and stochastic processes techniques aiming at a better understanding of biological systems. A special attention is devoted to online data analysis, local regression techniques and identification of complex biological systems. Our investigations encompass both theoretical aspects and concrete applications of the issues alluded to above. To be more specific, we focus on the following topics:

Analysis of high dimensional data. We deal here with statistical methods in relation with biomedical issues. Our two main activities in this domain are:

- **Online Factorial Analysis**: High dimensional data are often obtained online, and cannot be stored integrally in a computer memory. One of the recent challenges in data analysis is then to be able to perform an accurate classification or clustering by taking advantage of the possibility of updating the information. This has to be done, of course, in a rather simple and efficient way, allowing real time analysis. To this aim, we use techniques based on some sophisticated tools coming from stochastic approximation.

- **Local Regression Techniques**: The main issue here is the construction of a procedure allowing to assess, in quite a general framework, whether a given model fits a data set regarding most assumptions made in elaborating the model. This is based on a generalization of the Cramer-Von Mises statistics and involves a non parametric estimate of the conditional distribution of the response variable. A detailed analysis of the procedure, including rate of convergence and asymptotic properties, is being performed. The strategy is then implemented for a study concerning fetal biometry.

Estimation for Complex and Biological Systems Photodynamic therapy is the main application we focus on for these methods based on sophisticated stochastic processes tools. Other topics of interest include bacteriophage systems and biochemical reactions.

- **Photodynamic therapy**: Since 1988, some control system scientists and biologists at the Centre de Recherche en Automatique de Nancy (CRAN in short, Health-Biology-Signal department) have worked together to develop the photodynamic therapy (PDT), an alternative treatment for cancer, by means of a model-based approach. The global aim in this direction is to use statistical as well as mechanistic models in order to (i) improve the response reproducibility, (ii) help biologists and chemists in the design of new photosensitizing agents and (iii) provide insight into complex phenomena associated with oncogenesis, tumor angiogenesis and interactions with the treatment. (iv) assist clinicians in their therapeutic decisions. This heavily relies on the production of accurate and simple enough models involving various type of stochastic processes, such as Markov chains, branching processes and stochastic differential equations. The main questions here concern generally identification or estimation properties, but simulation issues can be important too.

- **Other complex biological systems**: Numerous biological systems are accurately described by multidimensional noisy differential equations driven by Markovian processes. We have focused recently on two challenging models: (i) Bacteriophage systems, in which one tries to model the competition between a population of bacteria and their related bacteriophage virus. Under certain experimental assumptions (discussed with a Microbiology group at Universitat Autònoma de Barcelona), one can model these interaction by a system of delayed stochastic differential equations. (ii) Biochemical reactions, for which we focus on a special regime with rare species (like enzymes) and common species (proteins for instance). We have made relevant progresses on the analysis of these systems.
Inference for Gaussian systems. Some members of our team are very active in the analysis of two challenging probabilistic objects: (i) Multidimensional noisy differential equations driven by Gaussian processes, beyond the realm of Brownian motion. (ii) Fractional fields with anisotropic features, and in particular operator scaling properties. These systems are very recently introduced, and a lot of theoretical work is required in order to understand them properly. We have undertaken a long term analysis in this demanding context, but we also plan to apply these techniques in two concrete biomedical situations: random mechanical fluctuations of nanoparticles and automatic detection of osteoporosis.

2.3 Objectives for the evaluation period

We shall divide this section according to our 4 main topics of interest, namely: analysis of high dimensional data, local regression techniques, photodynamic therapy, estimation for complex biological systems.

2.3.1 Analysis of high dimensional data

Our projects in this direction fit into the global framework of online analysis (applied to rather classical data analysis tools), support vector machines theory and non parametric statistics. These three popular areas in statistics received a lot of attention during the last decades, and we shall try to summarize here our specific concerns.

Data analysis. Generally speaking, there exists an overwhelming amount of articles dealing with the analysis of high dimensional data, and it would be too long to try to give a global overview on the topic. The setting we have chosen for this project is the statistical formulation of the factorial analysis, established by Benzécri. For sake of simplicity, we shall explain this mathematical modeling of the situation we are faced with in the case of a Principal Component Analysis (PCA in short), with a metric which will simply be the identity (other kind of metrics can be useful for statistical purposes). In this context, we are given a sequence of observations \( \{ x_j; j \geq 1 \} \), which are supposed to be the realization of a sample \( \{ X_j; j \geq 1 \} \) of independent centered \( \mathbb{R}^p \)-valued random variables (generalization to non-independent or non identically distributed random variables are possible). Setting \( X = X_1 \), the aim of the PCA is to find the best \( r \)-dimensional approximation of \( X \) (with \( r < p \)), in the following sense: let \( F_r \) be a \( r \)-dimensional subspace of \( \mathbb{R}^p \), from which we can extract an orthonormal basis \( (u_1, \ldots, u_r) \). Denote by \( \pi \) the orthogonal projection on \( F_r \). Since \( E[\|X\|^2] \) can be decomposed into:

\[
E[\|X\|^2] = E[\|X - \pi(X)\|^2] + E[\|\pi(X)\|^2],
\]

our aim is to maximize the second term \( E[\|\pi(X)\|^2] \) in the identity above. Denoting by \( \text{Cov}(X) \) the covariance matrix of \( X \), this can be reduced to a maximization of the quadratic function in \( u \):

\[
E[\|\pi(X)\|^2] = \sum_{k=1}^{r} u_k^2 \text{Cov}(X) u_k,
\]

under some constraints taking into account the fact that \( (u_1, \ldots, u_r) \) has to be an orthonormal basis.

However, data analysts are often faced with the problem of dealing with a continuous, rapid and infinite flow of data. Examples include web, telecommunications, process control or financial data. Let us recall that a data stream is an ordered sequence of instances \( \{ x_j; j \geq 1 \} \), which can be read only once or a small number of times, using limited computing and storage capabilities. In this context, one still has to find a sequential way to determine the vectors
defined above. However, in the simple (and unrealistic) case where \( \text{Cov}(X) \) is known, the first factor can be determined in a recursive way by considering the sequence defined by 

\[
X_{j+1} = (\text{Id} + a_j \text{Cov}(X))X_j.
\]

Indeed, for a well chosen family \( \{a_j; 1 \leq j \leq n\} \), it can be shown that a normalized version of \( X_j \) converges almost surely to the first coordinate of the PCA. This basic procedure has to be generalized then to the case of a real online situation, when \( \text{Cov}(X) \) is unknown, and to other factorial analysis situations. Notice also that further information, apart from the almost sure convergence, should be obtained in order to validate the methodology: central limit theorems, quadratic rate of convergence, and other ways to measure the convergence of our procedure. This ongoing project had already been started in the series of papers (Monnez 1994, 2006).

With these preliminary results in hand, we are now faced with the following challenges:

(1) **Probabilistic analysis of the convergence:** The results mentioned above are always stated in the so-called *almost sure* sense, which is of course fundamental for applications, since it yields convergence of the approximate component analysis to the exact one for any real world experience. However, other kind of convergences are also highly important to probabilists: the central limit theorem gives an important information about the rate of convergence of the method, while large deviations or concentration techniques tell us about the decay of the probability to be far away from the exact solution as the number of data increases. It will thus be a priority for us to work out this kind of convergence in our online data analysis context.

(2) **Non stationary cases:** In many cases, the assumption that the data are generated at random according to a stationary distribution does not hold true. Indeed, for large time periods, one should really expect some changes in the distribution of the instances, and a nice recent analysis of this phenomenon is given in Gomeni and al., 1988. An important goal for us is then to produce and analyze some suitable adaptive algorithms for component analysis in this new and challenging context.

**Support Vector Machines.** The Support Vector Machine classification problem can be divided in two main parts:

1. Given a deterministic sample \( s = \{(x_i, y_i); i \leq n\} \), with \( x_i \in \mathcal{X} \) and \( y_i \in \{-1; 1\} \), try to separate the points \( s^\pm = \{(x_i, y_i); y_i = \pm 1\} \) by a boundary \( B \), according to a suitable criterion involving in general the minimal distance of the points to the boundary. This task is obtained, in most of the cases, in two steps: first one tries to be reduced to the case of a hyperplane in the role of the separating boundary. This is generally achieved by a change of variables, and enlarging the dimension of the space \( \mathcal{X} \) we are working on, allowing it to become a Hilbert space \( \mathcal{H} \). Then, in the case of a separating hyperplane, the problem is solved thanks to the optimization of a quadratic function subject to a series of linear constraints.

2. The first step produces a certain boundary \( B = B(s) \), and one can associate an error \( \mathcal{E}_n(B,s) \) to this boundary, for instance by counting the number of misclassified points. A probabilistic analysis of this error bound is then fundamental, and a lot of work is available in this direction for 2-class situation thanks to concentration of measure techniques. Notice however that generalizations to the multiclass case are highly nontrivial.

This being recalled, our aim for this part of the project is twofold:

1. Delve into the regularization path for multiclass support vector machines, where we recall that the regularization path is a classical technique in order to optimize a machine. Indeed, one can play with a given regularization parameter \( \lambda \geq 0 \) which goes in front of the penalization \( \|\xi\|_1 \) or \( \|\xi\|_2 \) of the errors in the quadratic programming problem. It is then natural to ask the machine to adjust the parameter \( \lambda \) in order to get the best possible classification performance. This also means that one has to be able to compute
a family of machines indexed by $\lambda$ without resolving the optimization problem for each value of the parameter. We planned to carry out this program for different multiclass machines and different kinds of penalization.

2. Assess a comparison between usual statistical classification techniques (discriminant analysis, logistic regression) and support vector machines, in the restricted framework of Gaussian mixtures (but without recurring to parametric techniques). This study can be lead at both theoretical and applied level, and answers a very natural question in learning theory.

Notice that, due to the fact that our PhD student Rémi Bonidal had to defend his dissertation quite early for professional reasons, only the first part of the program above has been achieved.

**Local regression techniques.** In the context where a response variable $Y$ is to be related to a set of regressors $X$, one of the general goals of Statistics is to provide the end user with a model that has some usefulness in predicting $Y$ for various values of $X$. Except for the simplest situations, the determination of a good model involves many steps. For example, for the task of predicting the value of $Y$ as a function of the covariate $X$, statisticians have elaborated many models such as the regression model with random regressors (D’Agostino and Stephens, 1986):

$$Y = g(X, \theta) + \sigma(X)\epsilon.$$  
Many assumptions must be made to reach it as a possible model. Some requires much thinking, as for example, those related to the functional form of $g(\cdot, \theta)$. Some are made more casually, as often those related to the functional form of $\sigma(\cdot)$ or those concerning the distribution of the random error term $\epsilon$. Finally, some assumptions are made for commodity. Thus the need for methods that can assess if a model is concordant with the data it is supposed to adjust. The methods fall under the banner of goodness of fit tests. Most existing tests are *directional*, in the sense that they can detect departures from only one or a few aspects of a null model. For example, many tests have been proposed in the literature to assess the validity of a given structural part $g(\cdot, \theta)$ (see Alcalá and al., 1999). Some authors have also proposed tests about the variance term $\sigma(\cdot)$ (cf. Liero, 2003). Procedures testing the normality of the $\epsilon_i$ are given, but for other assumptions much less work has been done (see however D’Agostino and al.). Therefore the need of a global test which can evaluate the validity of a global structure emerges quite naturally.

One quantity which embodies all the information about the joint behavior of $(X, Y)$ is the cumulative conditional distribution function, defined by

$$F(y|x) = P(Y \leq y|X = x).$$

Thus the need for a test that can detect any departures from a model for $F(y|x)$. If this model is $F_0(y|x)$, the problem is that of testing

$$H_0 : F(y|x) = F_0(y|x).$$

In order implement this test, the idea is to compare a nonparametric estimator of $F(y|x)$ with $F_0(y|x)$. Here, the nonparametric estimator $\hat{F}_n(y|x)$ is based on the so-called local polynomial approach. As for measuring the distance between the estimator $\hat{F}_n(y|x)$ and $F_0$, we use a version of the generalized Cramer-von Mises statistic

$$n\sqrt{h} \int \int \left(\hat{F}_n(y|x) - F(y|x)\right)^2 w(x,y)F(dy|x)dx,$$
with a weight \( w \) which has to be chosen accurately. We have obtained the asymptotic normality of the test statistic under \( H_0 \) in both cases where \( F_0 \) is fully specified and depends on unknown parameters. We state that the asymptotic normality is not affected by the introduction of estimators in the test statistic. We also give the local power properties of the test, and we use some simulation methods in order to evaluate the power of the test. Our study further includes a discussion about the choice of parameters used in local polynomial estimation. In particular, we focus on methods for the choice of the bandwidth parameter.

**Cohorts analysis.** Let us mention some important applications of our data analysis methods in a biomedical context: the analysis of cohorts of data. Indeed, some medical teams in Nancy are faced with an overwhelming amount of data, for which a serious statistical assessment is needed. Among those let us mention the INSERM team of Pr. Jean-Louis Guéant and the INRIA team Orpailleur (particularly with Marie-Dominique Desvignes and Malika Smail). The goal of the collaboration we have recently established with them is to extract biological markers for different diseases (cognitive decline; inflammatory intestinal diseases; liver cancer). To this aim, the INSERM team provides us with several data cohorts with a high number of variables and subjects. As in many instances in Biostatistics, one is then faced with a very high dimensional data, from which we hope to extract a reduced number of significant variables allowing to predict the cardiovascular risk accurately. Moreover, these characters should be meaningful to practitioners. The objective for us is thus to design an appropriate variable selection, plus a classification procedure in this demanding context. Finally, let us highlight an original feature of this collaboration: it combines our own data analysis techniques with those developed by the Orpailleur team, based on symbolic tools. We hope that this experience will enrich both points of view and give raise to new methods of data analysis.

### 2.3.2 Estimation for Complex and Biological Systems

The application which federates our activity on stochastic processes is photodynamic therapy. We shall describe this cancer treatment as well as related statistical problems below, and then move to the other project which fall within the banner of estimation problems for biological systems: estimation for equations driven by Gaussian processes and application to anomalous protein fluctuations, study of a bacteriophage system and introduction of a system describing molecular dynamics.

**Model-based development of nanoparticle-based photodynamic therapy.** Since 2010, cancer has become the leading cause of death worldwide. Recent developments on multifunctional nano-systems [ANR PCV Nano-VTP, ANR P2N PDTX, ANR EMED Target-PDT] have opened new perspectives for tumor control by proposing new nano-actuators and nano-sensors in *in vivo* anti-cancer treatments. A direct consequence was the emergence of nanoparticle-based therapies such as the microwave hyperthermia therapy, nanoparticle-enhanced radiation therapy and the targeted photodynamic therapy (PDT): [ANR PCV Nano-VTP (2009-2011), ANR P2N PDTX (2011-2013), ANR EMED Target-PDT (2010-2012)]. All these treatments rely on the selective uptake of a nanoparticle embedding therapeutic agents and markers by the cancer cells. This administration phase is then followed by exposure to the appropriate wavelength of light or X-ray to activate the therapeutic agent (radio- or photosensitizers). When activated by irradiation, the agent interacts with biological molecules to produce cytotoxic species, that elicit both apoptotic and necrotic responses within treated tumors.
The development of these innovative treatments requires **new treatment planning systems based on numeric models** that shall help the clinicians to select the suited treatment modalities for two targeted applications: nanoparticle-based photodynamic therapy and nanoparticle-enhanced radiotherapy. Models are required to **predict** the distribution of light and/or radiation in biological tissues with or without nanoparticles, to **optimize** the positioning of optical fibers into the brain tumor, to **understand** the interactions between X-rays and nanoparticles and to **design** model-based control strategies able to adapt the light and radiation dose in real-time during the treatment application.

This project gathers clinicians, biologist and system control researchers of the CRAN (UL-CNRS UMR 7039, Santé-Biologie-Signal department) with the mathematicians of the IECL (UL-CNRS UMR 7502) within the framework of the INRIA team BIGS. This collaboration has been initiated since 2007, by the development of innovative models for the pharmacokinetic characterization of photosensitisers (Bastogne et al. 2008), the phenomenological modeling of tumor diameter growth based on a mixed effects model [8] and the development of model-based treatment planing systems [6, 46]. More recently, some members of the two laboratories have been involved in a PEPS CNRS-INSERM-INRIA OPTIQUE project (2012-2014) based on the modeling and optimization of irradiance in heterogenous biological tissues treated during the interstitial photodynamic therapy.

**Modeling and statistical estimation with stochastic processes.** As mentioned above, estimation procedures for diffusion type processes is an important ingredient of our PDT modeling. Indeed, let us recall first that the mathematical description of photodynamic therapy can be split up into four parametric models: the uptake model (pharmacokinetics of the photosensitizing drug into cancer cells and tumors), the model of light propagation (interstitial application of PDT), the photoreaction model and the model of tumor response.

(i) Several papers have been reported for the application of system identification techniques to pharmacokinetics modeling problems (see Gomeni and al., 1988, Cobelli and al., 2000, Delforge and al., 2000, Audoly and al., 2001 and Evans and al., 2004). But two issues were ignored in these previous works: presence of timing noise and identification from longitudinal data. In Bastogne and al. (2007), we have proposed a bounded-error estimation algorithm based on interval analysis to solve the parameter estimation problem while taking into consideration uncertainty on observation time instants.

(ii) According to what we know so far, no parameter estimation study has been reported about the photoreaction model in photodynamic therapy. A photoreaction model, composed of six stochastic differential equations, is proposed in Dobre and al. (2007). The main open problem is to access to data. We currently build on an experimental platform which aims at overcoming this technical issue.

(iii) An identifiability study coupled to a global sensitivity analysis of the photoreaction model has recently been proposed (see [34]). The photoreaction model has also been used to design an original and efficient model-based control strategy of the cytotoxicity produced during the treatment. An innovative technical platform was build up by the CRAN which has already been validated by in vitro assays. In vivo tests are ongoing.

(iv) When tumors are deep inside biological tissues, such as glioblastoma (high grade brain tumors) an interstitial application of PDT is needed. This treatment first consists in positioning at appropriate coordinates optical fibers whithin the tissues to be treated. Clinicians need assistance to help them to select this optimal positioning. Indeed, the therapeutic efficiency of the interstitial PDT depends on the suited position of the optical fibers into the brain. Their optimal positioning generally consists to inverse a mathematical model of the light propagation in biological tissues. The radiative transfer equations are too complex to be used
in this purpose. Secondly, one of the major drawbacks of this equation is the high number of photon simulations it requires, yielding a huge demand in computer means. Consequently, we wish to adapt the path integral formulation of the light transport equation to the context of photon transport into tissues. Specifically, this strategy hinges on a representation of the solution to the light transport problem as an integral over an infinite dimensional domain, and then approximates this integral by means of Metropolis type techniques. This powerful method induces nontrivial technical elaborations in the context of inhomogeneous tissues, but is expected to be very interesting in terms of computational costs. Note that the same kind of equation has been initially used in image synthesis but has never been employed to describe the light transport in real biological tissues. This scientific challenge is the central objective of the PEPS CNRS-INSERM-INRIA OPTIQUE project (2012-2014). First results have been presented in [66] and an APP (Agence pour la Protection des Programmes) application is planned, and further details about the research program are detailed at Section 5.2.3.

Bacteriophage and other biological systems. In the last years Bacteriophage therapies are attracting the attention of several scientific studies. They can be a new and powerful tool to treat bacterial infections or to prevent them, being successful e.g for animals such as poultry or swine. Very roughly speaking, they consist in inoculating a (benign) virus in order to kill the bacteria known to be responsible of a certain disease. This kind of treatment is known since the beginning of the 20th century, but has been in disuse in the Western world, erased by antibiotic therapies. However, a small activity in this domain has survived in the USSR, and it is now re-emerging (at least at an experimental level). Among the reasons of this re-emersion we can find the progressive slowdown in antibiotic efficiency (antibiotic resistance). Reported recent experiments include animal diseases like hemorrhagic septicemia in cattle or atrophic rhinitis in swine, and a need for suitable mathematical models is now expressed by the community.

Our team has been in contact with Montserrat Llagostera, a researcher from the Universitat Autònoma de Barcelona specialized in bacteriophages for 30 years. She had a demand for rigorous probabilistic models describing the bacteriophage dynamics, and the following is an attempt to answer her needs.

A simplified version of the kind of system one is faced with when dealing with the competition problem between bacteria and bacteriophage can be written as:

\[
\begin{align*}
    dS_t &= [\alpha - kQ_t]S_t \, dt + \sigma_1 S_t \, dW^1_t \\
    dQ_t &= -mQ_t - kQ_tS_t + \ell Q_{t-\tau}S_{t-\tau} + \sigma_2 Q_t \, dW^2_t.
\end{align*}
\]  

(1)

In this equation, \(S\) stands for the bacteria concentration, \(Q\) for the bacteriophage concentration, \(\alpha, k, d, m, \sigma_1, \sigma_2\) and \(\ell\) are positive coefficients, \(W^1, W^2\) are two independent standard Wiener processes, and \(\tau\) is a certain delay, interpreted as a latency time in biological terms. The deterministic part of the system expresses the fact that the bacteria growth is stopped by the inoculation of phages, which are introduced with rate \(d\) and reproduce themselves by killing bacteria, after latency time \(\tau\) and with rate \(\ell\). The stochastic differential is interpreted in the Stratonovich sense for physical reasons, and we usually think of \(\sigma_1, \sigma_2\) as small perturbation coefficients.

A first fundamental issue for practical purposes is to find possible equilibriums \((\bar{s}, \bar{q})\) of the system, and the convergence to those equilibriums. Indeed, the existence of these equilibrium would help to quantify the minimal amount \(d\) of viruses to be inoculated to animals in order to get rid of bacteria. It should be mentioned that some similar problems have been treated (in a rather informal way, invoking a linearization procedure) by Carletti in 2007, and more rigorously in Dalal and al (2008) and Mao and al (2005) by means of Lyapounov
functions techniques. None of these tools can be applied directly to our system, and we wish to use concentration and large deviations methods (on which we have already an expertise, see Márquez-Rovira-Tindel (2006), or Tindel (2003)) in order to combine convergence to equilibrium for the deterministic system and deviations of the stochastic system.

A second important problem consists in an accurate parameter estimation for the coefficients of equation (1), in concordance with the data provided by our colleagues from the Genetics and Microbiology department of the Universitat Autònoma de Barcelona. Here again, one should adapt the tools recalled previously, and we will pay a special attention to those based on ergodic properties of our differential system. These ergodic properties will first have to be established.

Notice that A. Muller-Gueudin is also working with A. Debussche and O. Radulescu on a related topic, namely the convergence of a model of cellular biochemical reactions. We don’t describe this model precisely here for sake of conciseness, and postpone its introduction to the section describing our main results. Also notice that C. Lacaux is working with H. Bierné on parameter estimation of anisotropic random fields, with potential applications to model bones and assessment in the diagnostic of the osteoporosis. This project is described in Section 5 and is linked to previous works on self-similar anisotropic random fields (see Section 2.6).

### 2.3.3 Inference for Gaussian systems

**Statistical inference for Gaussian random equations.** As mentioned above, the main application we have in mind for this statistical inference problem is a very nice model of anomalous fluctuations within proteins introduced by Samuel Kou, and related to very recent single molecule experiment. However, before one can move towards applications, a lot of fundamental work on theoretical estimation procedures is still to be performed. We focus on these aspects in what follows.

The basic theoretical ingredients of differential equations driven by a fractional Brownian motion (fBm) are now well understood, mainly thanks to the so-called *rough paths* method, see Friz and Victoir (2010), Gubinelli (2004), Lyons and Qian (2002). The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in Deya and Tindel (2008). On the other hand, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. Two good references on the question are Tudor and Viens (2009), and Papavasiliou and Ladroue (2009), dealing with some very particular families of equations, which do not cover the cases of interest for us.

In order to understand better what we might aim at for parameter estimation for equations driven by a Gaussian process and by fractional Brownian motion in particular, let us put a special emphasis on some basic facts about parameter estimation for diffusion processes. Indeed, this topic lies at the heart of our projects on bacteriophages and subdiffusion within molecules, and is also present in photodynamic therapy.

First of all, the simplest form of differential system one is faced with in this context can be written as:

\[
Y_t = a + \int_0^t \mu(Y_s; \theta) \, ds + \int_0^t \sigma(Y_s; \theta) \, dB_s, \quad t \in [0, T],
\]

where \(a\) is an initial condition lying in \(\mathbb{R}^k\), \(\{\mu(\cdot; \theta), \sigma(\cdot; \theta) ; \theta \in \Theta\} \) are two families of smooth functions from \(\mathbb{R}^k\), respectively to \(\mathbb{R}^k\) and \(\mathbb{R}^{d,k}\), and where the parameter \(\theta\) varies in a subset \(\Theta \subset \mathbb{R}^m\). The driving noise \(B\) of equation (2) is a \(d\)-dimensional Brownian motion. The challenge is then to estimate the parameter \(\theta\) from a single observation of a path of \(Y\) at some discrete instants \(\{t_i; 1 \leq i \leq n\}\), for \(n\) large enough.
Almost all the methods which can be found in the literature in order to solve the problem mentioned above are based on (or are closely linked to) the maximum likelihood principle. Indeed, if $B$ is a Brownian motion and $Y$ is observed at some equally distant instants $t_i = iT/n$ for $i = 0, \ldots, n$, then the log-likelihood of a sample $(Y_{t_1}, \ldots, Y_{t_n})$ can be expressed as

$$
\ell_n(\theta) = \sum_{i=1}^{n} \ln \left( p \left( \frac{T}{n}, Y_{t_i-1}, Y_{t_i}; \theta \right) \right),
$$

where $p$ stands for the transition semi-group of the diffusion $Y$. Theoretically, one can then estimate $\theta$ from expression (3) by standard differentiation arguments, though numerical approximations are often crucial.

Our aim is to transpose the considerations above to the case of a generic equation (2) when the driving noise $B$ is a fractional Brownian motion. Recall that this family of processes is indexed by a parameter $H \in (0, 1)$ (called Hurst parameter), and is a natural generalization of Brownian motion. Indeed, its covariance function is given by $E[(B_t - B_s)^2] = |t - s|^{2H}$. This means that for $H = 1/2$ one retrieves the usual Brownian motion, and otherwise one gets a family of centered Gaussian processes with any prescribed Hölder regularity, plus nice scaling properties.

One of the main difficulty when one faces studies differential equations driven by a fractional Brownian motion is that the Markovian structure disappears. This means in particular that the likelihood of a sample cannot be represented anymore by a simple function like (3), and one has to think about completely distinct procedures. The two pioneering works we wish to complete in this direction are thus the following:

1. Implement numerically a pseudo maximum likelihood procedure thanks to integration by parts techniques in the Malliavin calculus sense, and observe their performances on numerical data.
2. Get a consistent estimator of the drift term in fairly general situations, by means of ergodic type properties.

Gaussian or stable random fields. Recently, an important class of anisotropic processes called operator scaling random fields has been introduced by Biermé, Meerschaert and Scheffer (2007). It accommodates for self similarity encoded by matrices, which allows privileged directions for the field. To give a quick definition of these processes, we will say that $X = \{X(x); x \in \mathbb{R}^d\}$ is an $E$-operator scaled field if the following holds true:

$$
\forall c > 0, \quad \left( X(c^E x) \right)_{x \in \mathbb{R}^d} \overset{(d)}{=} c \left( X(x) \right)_{x \in \mathbb{R}^d},
$$

where the expression $c^E$ has to be understood in the matrix sense. Many basic properties of this kind of objects, such as regularity of dimensions of the level sets, still have to be established. We shall also use them for a project of automatic detection of osteoporosis on the basis of image analysis. For sake of conciseness, we will go back to a more detailed description of the results concerning operator scaled fields at Section 2.6.

2.4 Analysis of high dimensional data objective: executive summary

2.4.1 Personnel

R. Bar (PhD), S. Ferrigno (MdC), B. Lalloué (PhD), J-M. Monnez (PR), A. Muller-Gueudin (MdC), S. Tindel (PR).
2.4.2 Project-team positioning

The statistical modeling of data is an increasing area, which is obviously represented in the INRIA projects panel. Among the teams whose scope can be related to ours, let us mention the SELECT project, run by Pascal Massart and Gilles Celeux, which focuses on selection of models by means of probabilistic tools. Our project is quite different in its methodology and aims: while we appeal to a common background in probability and statistics, we will use a wider spectrum of techniques in order to analyze high dimensional data. This is mainly because the data analysis part of our project has a more application oriented point of view, always focused on biologic systems. This forces us to have a grasp on many different techniques. We claim however that the composition of our team, made of competent probabilist and statisticians coming from different areas, allows us to face this challenging situation.

The online adaptation of PCA and other dimension reduction statistical algorithms is mostly developed by our team, and especially by J-M. Monnez. However, the online data analysis in general is a fast growing area, and let us mention here some relevant groups or individuals of the field, whose concerns are close enough to ours:

• M. Maloof (Georgetown University, USA) is a leading expert in online analysis and concept drift, from a computer science point of view. His work embraces theoretical and empirical aspects of the problem.

• An important group at Princeton University (USA) is working, under the direction of V. Poor, on statistical change detection. This might be an important element for our future investigations.

• Among the pioneers of stochastic approximation, the works of Hanfu Chen (Beijin University, China), Harro Walk (Stuttgart University, Germany) and Lennart Ljung (Linkhöping University, Sweden) are an important source of inspiration for us.

• Within the French community, the new developments of Jérôme Lelong (ENSTA, Paris) on stochastic algorithms can also be related to our main concerns.

• The Théorie de l’information, Apprentissage statistique, Analyse de données massives group at CMLA (ENS Cachan). This group is lead by A. Trouvé and N. Vayatis, and is focused on several aspects of statistical learning. In particular, the ranking problem and the adaptation of statistical learning to online data is one of its main current topics.

• Some people at the Laboratoire de Probabilités et Modèles Aléatoires (Universities of Paris 6 and 7) are also working on several interesting aspects of statistical learning. Among them, let us quote P. Alquier, G. Biau and S. Boucheron.

• Let us also mention the name of two top experts in probabilistic aspects of classification theory, whose current work provides us with an important source of inspiration: G. Lugosi (Universitat Pompeu Fabra, Barcelona, Spain) and P.L. Bartlett (Berkeley University, USA).

One of the major challenges in local regression techniques is the choice of an accurate bandwidth for the estimation kernel. This crucial technical point is also dealt with by other research groups in statistics:

• P. Hall (Australian National University, Canberra) and B. Park (Seoul University, Korea) have treated many aspects of bandwidth selection in an interesting way, among
other relevant contributions to nonparametric statistics. L.S. Huang (University of Rochester, USA) is also a leading expert in nonparametric curve estimation, whose recent contributions have to be kept in mind.

- Let us mention the wide number of contributions of J. Fan (Princeton University, USA) for the development of robust estimation methods of nonparametric type. W. González-Manteiga (University of Santiago de Compostela, Spain) has also worked on challenging aspects of polynomial regression and related goodness of fit results.

2.4.3 Scientific achievements

Our contributions to data analysis in a Biological context can be decomposed in 5 subprojects:

(1) At a theoretical level, we have kept on working on the so-called online data analysis alluded to at Section 2.3.1. Specifically, we have carried out in [17] the construction of a fast and recursive algorithm for clustering large data sets with the $k$-medians methods, which is an important challenge in computational statistics. Borrowing ideas from MacQueen, who introduced a sequential version of the $k$-means algorithm, a new class of recursive stochastic gradient algorithms designed for the $k$-medians loss criterion has been proposed. By their recursive nature, these algorithms are very fast and well adapted to deal with large samples of data that are allowed to arrive sequentially. It is proved that the stochastic gradient algorithm converges almost surely to the set of stationary points of the underlying criterion. A particular attention is paid to the averaged versions, which are known to have better performances, and a data-driven procedure that allows automatic selection of the value of the descent step is proposed. The performance of the averaged sequential estimator is compared on a simulation study, both in terms of computation speed and accuracy of the estimations, with more classical partitioning techniques such as $k$-means, trimmed $k$-means and PAM (partitioning around medoids). Finally, this new on-line clustering technique is illustrated on determining television audience profiles with a sample of more than 5000 individual television audience measured every minute over a period of 24 hours.

Moreover, we have handled (see [1]) the analysis of data whose characteristics such as mathematical expectation or covariance matrix may vary with time, a problem which arises very naturally in this context. In order to save computation time and thus take into account more data, a method considering several data at each step (we talk about data blocks) is proposed. This technique can also be useful if data are sent and received block-wise. In parallel, a R package performing most of the methods of factorial analysis in an online way is under development.

(2) At a practical level, our efforts have focused first on an interesting study concerning peanuts allergy [26], for which our expertise in data analysis allows for a good prediction of allergy severity by means of rigorous methods. We developed an algorithm for simultaneously clustering eliciting doses and selecting discriminant variables. Our main conclusion is that antibody measurements offer information on the allergy severity, especially those directed against $rAra-h1$ and $rAra-h3$. Further independent validation of these results and the use of new predictors will help extend this study to clinical practices.

(3) At a practical level, our efforts have focused in a second time on an interesting study concerning the construction of a socio-economic neighborhood index which might quantify health inequalities, see [53]. While several socio-economic indices already exist in this application field, most of them are very simple both in term of methodological construction and of number of variables taken into account, and only a few use data mining techniques. In order to exploit the large data sets of socio-economic variables provided by censuses and create neighborhood
socio-economic indices yielding a better highlight of social health inequalities, a procedure
was set in order to automatically select the best indicators in a set of socio-economic variables
and synthesize them in a quantitative index. Application to three French metropolitan areas
allowed testing the procedure and confirming both its reproducibility on various urban areas
and the quality of the neighborhood socio-economic indices we had created (according to field
experts and study partners). In this context, our expertise in data analysis allows for a good
prediction by means of rigorous methods. Eventually, in order to simplify the application
of the creation procedure of a socio-economic index for non-statisticians, a R package called
SesIndexCreatoR was created to implement it.

(4) As mentioned above (see Section 2.3.1), the analysis of cohorts of data have been a promin-
ent feature of our applied activity since our creation. It includes the statistical assessment
at INSERM with Pr. Guéant, and a long term study of heart disease patients lead with Pr.
Albuisson (CHU Nancy).

(5) We have completed our regularization path program for SVM, focusing mainly on the
biclass case with a penalization involving the \(\ell^2\)-norm of the errors. We have improved many
computational aspects and compared our results to numerous benchmarks.

(6) As far as local regression techniques are concerned, we have focused our attention on
the development of a local linear estimator of the conditional distribution function. Indeed,
consider \((X, Y)\), a random vector defined in \(\mathbb{R} \times \mathbb{R}\). Here \(Y\) is the variable of interest and \(X\n\) the concomitant variable. As usual in the statistics literature, we work under the assumption
that a sample \(\{(X_i, Y_i)_{1 \leq i \leq n}\}\) of independent and identical replica of \((X, Y)\) is available. We
have studied the conditional distribution function \(F(y|X = x) = P(Y \leq y|X = x)\) and
a nonparametric estimator associated to this quantity. The distribution function has the
advantage of completely characterizing the law of the random variable at stake, allowing to
obtain the regression function, the density function, the moments and the quantile function.
It should also be noticed that conditional distribution functions are used for the estimation
of references curves in medical applications.

At a more technical level, our study is based on a local linear nonparametric estimator
of the conditional distribution function instead of the widely spread Nadaraya-Watson esti-
mator. Indeed, it is a well-known fact that the asymptotic bias of the Nadaraya-Watson
estimator behaves somehow badly. Observe however that local polynomial techniques are
good alternatives. Based on these techniques, we have focused on the following steps:

- Our main result is the uniform law of the logarithm concerning the local linear esti-
mator of the conditional distribution function (see \([37]\)). We investigate convergence in
probability and almost sure convergence results.

- The uniform law of the logarithm has then been used to construct uniform asymptotic
certainty bands for the conditional distribution function (see \([?]\)).

- The certainty bands alluded to above have been applied to simulated data.

- A variant of the test has been introduced in \([36]\).

Let us also mention that applications of these theoretical results to survival analysis are
currently the object of active research. An ongoing work concerning local regression techniques
is related to Fetal Biometry, an investigation line suggested by a collaboration between our
team and the Centre de Placentologie et Foetopathologie de la Maternité Régionale de Nancy,
under the direction of Professor Bernard Foliguet. The methods involved in Fetal Biometry
are usually based on the comparison of some measured values with the predicted values derived
from reference charts or equations in a normal population. However, it happens that maternal

14
and pregnancy characteristics have a significant influence on in-utero Fetal Biometry. We will thus produce some models allowing to construct customized fetal biometric size charts. In order to evaluate them, classical and polynomial regression can be used, but they are not the most appropriate to the kind data we have to handle. Hence, we plan to use local regression estimation in order to perform such an evaluation.

2.4.4 Collaborations

- Members of the Statistics team in Dijon (Cardot, Cénac) for the \( k \)-medians project, and in Strasbourg (Maumy-Bertrand) for local regression techniques.
- Genclis laboratory (an INSERM unit specialized in genetics studies) for the peanut allergy investigation. Guéant (also from INSERM) is our partner for cohorts analysis.
- Zmirou-Navier from a Public Health Department at Ecole des Hautes Études en Santé Publique for our social inequality index.
- Y. Guermeur (LORIA) for our SVM project.

2.4.5 External support

PhD thesis of B. Lalloué funded by Ecole des Hautes Études en Santé Publique.

2.4.6 Self assessment

**Strong points.** Our leadership in online data analysis is quite strong, and this research line is obviously worth pursuing. Moreover, it is strongly linked to interesting applied studies. We also believe that the studies concerning local regression techniques are promising, both in terms of theoretical studies and applications.

**Dead ends.** Our relations with the Genclis laboratory have deteriorated a couple of years ago. Though the topics handled by this entity are of primary interest for us, our level of collaboration has severely decreased. We have also discovered that our common interests with Y. Guermeur (LORIA) are not as strong as we expected. We might postpone further collaborations.

2.5 Modeling for complex and biological systems objective: executive summary

During the evaluation period we have tried to focus on our mainstream PDT application, but we have also turned to other projects involving several research group outside Nancy. Let us detail our scientific achievements in this direction.

2.5.1 Personnel

T. Bastogne (PR), R. Keju (PhD), J-L. Marchand (Postdoc), S. Mézières (MdC), A. Muller-Gueudin (MdC), S. Tindel (PR), P. Vallois (PR).

2.5.2 Project-team positioning

Several mathematical teams at INRIA are concerned with biological applications (most of them present at this evaluation). Among them let us highlight the BANG project, highly focused on cancer therapy, while the REO (www-rocq.inria.fr/REO/) project handles the
modeling of biologic (e.g. blood, air) flows. Both projects essentially rely on the development of mechanistic models of tumor growth and tumor angiogenesis. Complementariness between physical and statistical point of views of this subject would probably be profitable for the PDT project by identifying possible synergies between the modeling approaches.

The system identification and estimation for drug kinetics and other related complex biological processes constitutes a very lively area of research. The following teams are representative of the recent relevant contributions in this field:

- The group lead by N. D. Evans, K. R. Godfrey and M. J. Chappell (School of Engineering, University of Warwick, UK) has important contributions on systems modeling, identifiability and control of drug kinetics, epidemiological and biomedical processes.

- R. Middleton (Hamilton Institute, National University of Ireland) works on applications of systems control and dynamics in biology. His research puts a special emphasis in understanding the dynamics of networks of biochemical reactions, and also in modeling and control of HIV infection dynamics and Parkinson’s disease.

- From a mathematical point of view, the kind of problem we are dealing with is well represented by the production of two French researchers. Indeed, the PDE aspect is handled in an interesting way by B. Laroche (Laboratoire des Signaux et Systèmes, Université de Paris 11), in order to model and control biological systems including a clock process. On a statistical ground, A. Leclerc-Samson (University of Paris 5) develops some nice elements of parameter estimation for stochastic differential equations with applications to tumor growth, VIH dynamics and agronomy. The stimulating works of X. Mao (University of Strathclyde, Glasgow, UK), M. Carletti (University of Urbino, Italy) and S. Kou (Harvard University, USA) on complex stochastic systems in biology have been influential for the bacteriophage project.

2.5.3 Scientific achievements

Here again we can decompose our objective in several sub-projects.

(1) Model-based treatment planning system for ionizing and non-ionizing radiation therapies. This research direction fits into the model production for photodynamic therapy. Indeed, up to now the treatment planning systems used in radiotherapy only use mathematical models to describe the delivery of physical doses of radiation within biological tissues but cannot accurately predict the biological damages caused by such treatments. One important bottleneck is to account for the cell damage heterogeneity in the treated tumor. To this aim we firstly introduced in [6] a stochastic model based on multi-state Markov chains able to describe both treatment damage and cell reparation process.

More recently, we have proposed another model describing the lifespan of heterogenous tumors treated by radiotherapy. It is a bi-scale model in which the cell and tumor lifespans are represented by random variables. First and second-order moments, as well as the cumulative distribution functions and confidence intervals are expressed for the two lifespans with respect to the model parameters. One interesting result is that the mean value of the tumor lifespan can be approached by a logarithmic function of the initial cancer cell number. Moreover, we show that TCP (Tumor Control Probability) and NTCP (Normal Tissue Complication Probability), used in radiotherapy to evaluate, optimize and compare treatment plans, can be derived from the tumor lifespan and the surrounding healthy tissue respectively. Finally, we propose a ROC curve, entitled ECT (Efficiency-Complication Trade-off), suited to the selection by clinicians of the appropriate treatment planning (see [16]).
One difference between photodynamic therapy (PDT) and radiotherapy (RT) is the irradiation signal (X-ray in RT and light beam in PDT). Another one is the treatment planning: 10 to 30 daily sessions of treatment in RT against only one for PDT. To adapt the previous model to PDT, a continuous-time version was developed and proposed in [47]. The model has been implemented into Matlab and numerical simulations have emphasized the effects of the model parameters on the model output.

In the framework of a new collaboration with S. Niclou (NorLux Neuro-Oncology Laboratory, Department of Oncology, Centre de Recherche Public de la Santé, Luxembourg), we have extended our stochastic model of cell damage to describe the phenotypic heterogeneity in brain tumors. Preliminary results have recently been presented in [5]. Cancer stem cell (CSC) hypothesis suggests that tumor progression and recurrence rely on a small subpopulation of cancer cells with stem-like properties. The unresolved question is whether cancer stem cells lead to organisation of intratumoral phenotypic heterogeneity by hierarchical differentiation events or whether they represent one of the transitory phenotypic states. This is crucial not only for our understanding of tumor progression, but also for the successful design of novel therapeutic strategies targeting CSCs.

(2) A stochastic model for bacteriophage therapies. Let us recall that whenever the mobility of the different biological actors is high enough, bacteriophage systems can be modeled by a kind of predator-prey equation. Namely, set $S_t$ (resp. $Q_t$) for the bacteria (resp. bacteriophages) concentration at time $t$. Then a model for the evolution of the couple $(S, Q)$, which is a slight modification of (3), is as follows:

$$\begin{align*}
\text{d} S_t &= [\alpha - k Q_t] S_t \text{d}t + \varepsilon S_t \text{d}W^1_t \\
\text{d} Q_t &= \left[ d - m Q_t - k Q_t S_t + k b e^{-\mu \zeta} Q_{t-\zeta} S_{t-\zeta} \right] \text{d}t + \varepsilon Q_t \text{d}W^2_t,
\end{align*}$$

(4)

where $\alpha$ is the reproducing rate of the bacteria and $k$ is the adsorption rate. In equation (4), $d$ also stands for the quantity of bacteriophages inoculated per unit of time, $m$ is their death rate, we denote by $b$ the number of bacteriophages which is released after replication within the bacteria cell, $\zeta$ is the delay necessary to the reproduction of bacteriophages (called latency time) and the coefficient $e^{-\mu \zeta}$ represents an attenuation in the release of bacteriophages (given by the expected number of bacteria cell’s deaths during the latency time, where $\mu$ is the bacteria’s death rate). A given initial condition $(S_0, Q_0)$ is also specified, and the term $\varepsilon \text{d}W_t$ takes into account a small external noise standing for both uncertainties on the measures and the experiment conditions. One should be aware of the fact that the latency time $\zeta$ (which can be seen as the reproduction time of the bacteriophages within the bacteria) cannot be neglected, and is generally of the same order (about 20mn) as the experiment length (about 60mn).

With this model in hand, our main results in this direction (see [3]) have been the following:

- Quantification of the exponential convergence to a bacteria-free equilibrium of equation (4) when $d$ is large enough.

- Use of the previous result plus concentration inequalities in order to study the convergence of the noisy system to equilibrium in a reasonable time range.

- Simulation of the stochastic processes at stake in order to observe the convergence to equilibrium.

We have now gone back to interact with our contact Montserrat Llagostera (Microbiology team at Universitat Autònoma de Barcelona), in order to get a feedback on our results. The outcome will be mentioned at Section 5.2.7.
(3) Convergence of stochastic gene networks. In [58, 27], we propose simplified models for the stochastic dynamics of gene network models arising in molecular biology. Those gene networks are classically modeled by Markov jump processes, which are extremely time consuming. To overcome this drawback, we study the asymptotic behavior of multiscale stochastic gene networks using weak limits of Markov jump processes.

We consider a set of chemical reactions \( R_r, r \in R \); \( R \) is supposed to be finite. These reactions involve species indexed by a set \( S = 1, \ldots, M \), the number of molecules of the species \( i \) is denoted by \( n_i \) and \( X \in \mathbb{N}^M \) is the vector consisting of the \( n_i \)'s. Each reaction \( R_r \) has a rate \( \lambda_r(X) \) which depends on the state of the system, described by \( X \) and corresponds to a change \( X \rightarrow X + \gamma_r, \gamma_r \in \mathbb{Z}^M \).

Mathematically, this evolution can be described by the following Markov jump process. It is based on a sequence \( (\tau_k)_{k \geq 1} \) of random waiting times with exponential distribution. Setting \( T_0 = 0, T_i = \tau_1 + \cdots + \tau_i, X \) is constant on \([T_{i-1}, T_i)\) and has a jump at \( T_i \). The parameter of \( \tau_i \) is given by \( \sum_{r \in R} \lambda_r(X(T_{i-1})) \):

\[
P(\tau_i > t) = \exp \left( -\sum_{r \in R} \lambda_r(X(T_{i-1})) t \right).
\]

At time \( T_i \), a reaction \( r \in R \) is chosen with probability \( \lambda_r(X(T_{i-1}))/\sum_{r \in R} \lambda_r(X(T_{i-1})) \) and the state changes according to \( X \rightarrow X + \gamma_r: X(T_i) = X(T_{i-1}) + \gamma_r \). This Markov process has the following generator:

\[
Af(X) = \sum_{r \in R} [f(X + \gamma_r) - f(X)] \lambda_r(X).
\]

In the applications we have in mind, the numbers of molecules have different scales. Some of the molecules are in small numbers (e.g enzymes or other catalysts) and some are in large numbers (proteins or other main reactants). Accordingly, we split the set of species into two sets \( C \) and \( D \) with cardinals \( M_C \) and \( M_D \). This induces the decomposition \( X = (X_C, X_D) \), \( \gamma_r = (\gamma_r^C, \gamma_r^D) \). For \( i \in D \), \( n_i \) is of order 1 while for \( i \in C \), \( n_i \) is proportional to \( N \) where \( N \) is a large number. For \( i \in C \), setting \( \tilde{n}_i = n_i/N, \tilde{n}_i \) is of order 1. We define \( x_C = X_C/N \) and \( x = (x_C, X_D) \).

For this kind of system, we are able to give in [27] some relevant information on the asymptotic regime \( N \to \infty \) when different type of reactions are involved. Depending on the time and concentration scales of the system we distinguish four types of limits:

- Continuous piecewise deterministic processes (PDP) with switching.
- PDP with jumps in the continuous variables.
- Averaged PDP.
- PDP with singular switching.

We justify rigorously the convergence for the four types of limits. In a second step of the study, we shall put a special emphasis on relationships between our model and real world data (experiment lead by O. Radulescu).

(4) Variable length Markov chains. A classical random walk \( (S_n, n \in \mathbb{N}) \) is defined by \( S_n := \sum_{k=0}^n X_k \), where \( (X_k)_{k \geq 1} \) are i.i.d random variables. When the increments \( (X_k)_{k \in \mathbb{N}} \) are a one-order Markov chain, a short memory is introduced in the dynamics of \( (S_n) \). This so-called “persistent” random walk is no longer Markovian and, under suitable conditions, the rescaled process converges towards the integrated telegraph noise (ITN) as the time-scale and
space-scale parameters tend to zero (see Hermann-Vallois [39]). The ITN process is effectively non-Markovian too. In [20] our aim has been to consider persistent random walks \((S_t)\) whose increments are Markov chains with variable order which can be infinite.

Associated with a process \((X_n)\) which takes its values in a finite set, we consider an integer valued process \((M_n)\) so that \((X_n, M_n)\) is Markov and \(M_n\) measures the size of the memory at time \(n\). This variable memory is justified by a one-to-one correspondence between \((X_n)\) and a suitable Variable Length Markov Chain (VLMC), since for a VLMC the dependency from the past can be unbounded. We prove in [20] that, under a suitable rescaling, \((S_n, X_n, M_n)\) converges in distribution towards a time continuous process \((S^0(t), X(t), M(t))\). The process \((S^0(t))\) is a semi-Markov and Piecewise Deterministic Markov Process whose paths are piecewise linear.

Observe that, though our study in [20] is made at a theoretical level, it leads to potentially interesting applications in growth models for tumors. This kind of link will be developed in the next future.

2.5.4 Collaborations

- Collaboration with S. Niclou (NorLux Neuro-Oncology Laboratory, Department of Oncology, Centre de Recherche Public de la Santé, Luxembourg) on tumor growth models.

- Other collaborations concerning photodynamic therapy: Céline Frochot (DR CNRS, LRGP) for the photo-physical characterization of photosensitizer-doped nano particles, Serge Mordon (DR INSERM, U703) for the imaging and instrumentation development associated with PDT and the clinicians involved in clinical studies, Olivier Tillement (PU, Lyon) and Bruno Therrien (PU, U. Neuchâtel) for the manufacturing of multifunctional nanoparticles in PDT, Eric Bullinger (PU, U. Liège) and Thomas Sauter (PU, U. Luxembourg) for system biology applied to PDT.

- Contacts with M. Llagostera (Departament de Microbiologia, Universitat Autònoma de Barcelona) concerning bacteriophage models.

- Joint project with A. Debussche (ENS Bretagne, Mathematics) and O. Radulescu (Montpellier 2, Biophysics) on stochastic gene networks.

- Collaboration with a probability team in Dijon on variable length Markov chains.

2.5.5 External support

PhD Thesis of R. Kejn funded by a special grant of Université de Lorraine. ANR and PEPS grants by T. Bastogne.

2.5.6 Self assessment

*Strong points.* Photodynamic therapy is a research speciality developed in Nancy since 1988 and we are positioned as one of the leader team in France and Europe. This is a very rich topic at a mathematical level, and yields motivating and challenging real world applications. We believe that it should be at the heart of our project during the next 4 years period.

*To be developed.* At a mathematical level, our projects on bacteriophages, stochastic gene networks and variable length Markov chains are certainly innovative enough and well motivated. However, we still have to work in order to connect those development to applied activities in the Biomedical world. This aspect is one of our priorities for the next future.
2.6 Inference for Gaussian systems objective: executive summary

2.6.1 Personnel

M. Besalú (Postdoc), A. Chronopoulou (Postdoc), C. Lacaux (MdC), S. Tindel (PR).

2.6.2 Project-team positioning

For this part of our project we should compare to other teams working in Gaussian analysis on the one hand and self-similar processes on the other hand.

- As mentioned above in the text, general dynamical systems driven by Gaussian processes can only been handled through rough paths type techniques. This theory was created by T. Lyons (Oxford) at the end of the 90s, and several prominent probabilist are still actively participating in its development: P. Friz (Berlin), M. Hairer (Warwick), M. Gubinelli (Paris Dauphine) might be the greatest exponents. Mostly theoretically oriented, the most common applications of the topic concern mathematical finance. Our team is thus the only one trying to find applications of rough paths in a Biomedical context.

- One should mention however the existence of the REGULARITY team at INRIA, which focuses on fractional calculus (another way to handle stochastic calculus outside of the semimartingale realm, but not as general as rough paths). Relationships with this team would certainly benefit to both of us, but their favorite applications seem to concern mostly geophysical quantities.

2.6.3 Scientific achievements

(1) LAN property for fractional Brownian motion. We have first focused on an important statistical property of fractional Brownian paths on their own. Indeed, the local asymptotic normality (LAN) property is a fundamental concept in asymptotic statistics, which gives the asymptotic normality of certain estimators such as the maximum likelihood estimator for instance. In [23], we focus on the LAN property for the model where we observe a sample of \( n \) observations \( X_n = (X_1, \ldots, X_n) \) of a Gaussian stationary sequence. The sequence \( (X_n)_{n \in \mathbb{N}} \), whose spectral density \( f_\theta \) is indexed by a parameter \( \theta \), can admit antipersitence, long memory or short memory and be noninvertible. To be more specific, our main assumption is:

\[
 f_\theta(x) \sim_{x \to 0} |x|^{-\alpha(\theta)} L_\theta(x)
\]

with \( L_\theta \) a slowly varying function and \( \alpha(\theta) \in (-\infty, 1) \). We prove the LAN property by studying an asymptotic expansion of the log likelihood and using some results on Toeplitz matrices. In particular, our assumptions are fulfilled by fractional Gaussian noises and autoregressive fractionally integrated moving average processes (ARFIMA(\( p, d, q \))). We also obtain the LAN property for fractional Brownian motion.

(2) Inference for dynamical systems driven by Gaussian noises. As mentioned at Section 2.3.2 the problem of estimating the coefficients of a general differential equation driven by a Gaussian process is still largely unsolved. To be more specific, the most general (\( \mathbb{R} \)-valued) equation handled up to now as far as parameter estimation is concerned is of the form:

\[
 X^\theta_t = a + \theta \int_0^t b(X_u) \, du + B_t,
\]
where $\theta$ is the unknown parameter, $b$ is a smooth enough coefficient and $B$ is a one-dimensional fractional Brownian motion. In contrast with this simple situation, our applications of interest require the analysis of the following $\mathbb{R}^n$-valued equation:

$$X^\theta_t = a + \int_0^t b(\theta; X_u) \, du + \int_0^t \sigma(\theta; X_u) \, dB_u, \quad (5)$$

where $\theta$ enters non linearly in the coefficient, where $\sigma$ is a non-trivial diffusion term and $B$ is a $d$-dimensional fractional Brownian motion. We have thus decided to tackle this important scientific challenge first.

To this aim, here are the steps we have focused on:

- A better understanding of the underlying rough path structure for equation (5), carried out in [61], [65]. This step allows a proper definition of our equation of interest in a wide range of contexts.

- Gaussian type bounds for equations driven by a fractional Brownian motion, obtained in [10]. This is an important preliminary step for likelihood estimates for stochastic processes.

- An implementable numerical scheme for equations driven by irregular processes, which is one of the ingredients one needs in order to perform an accurate statistical estimation procedure (see [30]).

- A better understanding of the law of the solution $X^\theta_t$ to equation (5), carried out in [18]. This step allows to obtain smoothness of density for our equation of interest in a wide range of contexts, which is an essential prerequisite for a good estimation procedure.

- Another important preliminary step for likelihood estimates for stochastic equations is a good knowledge of their invariant measure in the ergodic case. This is the object of our article [24].

- Finally we have also progressed in our knowledge of noisy differential systems by extending the range of applications of rough paths methods [54, 63].

As far as statistical procedures are concerned, here are the main results we have obtained:

- Numerical aspects of a maximum likelihood type procedure for an equation of the form (5), expressed in terms of Malliavin calculus tools (see [22]).

- Convergence of a least square type estimator for an equation of the form (5) where the noise enters additively, handled in [59]. This is the first occurrence of a converging estimator for a general coefficient $b(\theta, \cdot)$.

(3) Local self-similarity properties and stable or Gaussian random fields. In 2009, C. Lacaux and H. Biermé carried on the study of some sample paths properties for an important class of anisotropic random fields called operator scaling random fields, which had been previously introduced by H. Biermé, M. Meerschaert and P. Scheffler (2007). To be more specific, the classical self-similarity property is replaced by the following operator scaling property:

$$\forall c > 0, \quad \left( X(c^E x) \right)_{x \in \mathbb{R}^d} \overset{(d)}{=} c \left( X(x) \right)_{x \in \mathbb{R}^d}, \quad (6)$$

where $c^E := \exp(E \ln(c))$. In particular, the Hölder regularity properties of operator scaling Gaussian or stable harmonizable random fields have been studied and can be expressed in
terms of the matrix $E$. C. Lacaux and H. Biarmé are now focusing on the key problem of the estimation of the parameter $E$, related to our osteoporosis application project.

Notice that processes defined by (6) enjoy a regularity which does not vary along the trajectories, and this can be too restrictive for some applications. In order to obtain more flexibility in the random fields we consider and construct some anisotropic random fields whose Hölder regularity properties are allowed to vary, we introduce in [15] a local version of the operator scaling property (similar to the local version of the classical self-similarity property defined in Benassi, Jaffard and Roux 1997). This local property is illustrated and we also define and study harmonizable multi-operator scaling stable random fields. For such a multi-operator random field, we obtain an accurate upper bound of both the modulus of continuity and global and directional Hölder regularities at any point $x$. As expected, the Hölder regularity properties vary along the trajectories.

The method we used in [15] can be applied to study the modulus of continuity of many stable or Gaussian random fields. In [14], it has been developed in the more general framework of conditionally sub-Gaussian random series. We have also proposed some conditions under which the series converges uniformly (on a random ball), with explicit rates of convergence.

In [50], we study the sample paths properties of an anisotropic random field, which is defined as limit of an invariance principle and is of the same type as a multifractional Brownian sheet. Our first aim was to generalize Cohen-Marty (2008), that is to obtain some multifractional random fields indexed by $\mathbb{R}^d$ with $d \geq 2$ and to allow Hurst indices to be lower than $1/2$. To overcome the problem of the values of the Hurst indices which characterize the limit field, we focus on stationary sequences $(X_n(H))_{n \in \mathbb{N}}$, where $H \in (0,1)^d$, defined by an harmonizable representation. Then, our limit field $S_h$ is defined as the limit of

$$S_h^N = \left\{ \sum_{n_1=1}^{[Nt_1]} \ldots \sum_{n_d=1}^{[Nt_d]} \frac{X_n(h_n^N)}{N^r_n} : t \in [0, +\infty)^d \right\}$$

for some suitable families $(h_n^N)_{n,N}$ and $(r_n^N)_{n,N}$. We then study the sample paths property of this limit field. In particular, we obtain some local self-similarity properties for its increments of order $k$ and its pointwise global and directional Hölder exponents. We also define (and obtain) some pointwise multi-Hölder exponents which characterize the Hölder property satisfied by the increments of order $d$ of $S_h$.

We are also interested in self-similar processes indexed by manifolds in [42]. This study is motivated by the fact various spatial data are indexed by a manifold and not by the Euclidean space $\mathbb{R}^d$ in practical situations such as image analysis.

### 2.6.4 Collaborations

- Statistics team (Cohen, Gamboa, Loubes, Panloup) in Toulouse for the LAN property for fBm and ergodic properties of Gaussian equations.
- H. Biarmé (Tours), J. Istas (Grenoble) and P. Scheffler (Siegen) are collaborators of C. Lacaux on self-similar random fields.
- T. Cass (Imperial College London), F. Baudoin (Purdue), M. Hairer (Warwick), J. León (Cinvestav Mexico), D. Nualart (Kansas), C. Ouyang (Michigan), A. Neweunkirch (Manheim), L. Quer (Barcelona) are regular international collaborators of S. Tindel on stochastic differential equations driven by fractional Brownian motion.
2.6.5 External support

C. Lacaux has been member of the MATAIM ANR grant. She is currently member of GDR Multifractal, GDR GeoSto and GDR Maths & Entreprises. S. Tindel has been member of the ECRU grant.

2.6.6 Self assessment

Strong points. Self similar fields and rough paths theory are obviously two research areas where our results have a strong international impact. We have recently been invited to several important international conferences on the topic, and we have also organized a high level workshop in Nancy last year. These directions should thus be privileged in the next few years.

To be developed. Recall that the main applications we have in mind for this kind of systems concern anomalous fluctuations of proteins on the one hand and osteoporosis on the other hand. However, the process of getting data for those applications being quite long, we have focused for the moment on theoretical aspects of our study. Now we understand much better the objects we are dealing with, and are presumably ready to handle data sets.

3 Knowledge dissemination

3.1 Publications

<table>
<thead>
<tr>
<th></th>
<th>10-11</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhD Thesis</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H.D.R (*)</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Journal</td>
<td>14</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Conference proceedings (**)</td>
<td>10</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Patent</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Technical report</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

(*) HDR: Habilitation à diriger des Recherches  
(**) Conference with a program committee

Indicate the major journals in the field and, for each, indicate the number of papers coauthored by members of the project-team that have been accepted during the evaluation period.


2. Two major journals in mathematical biosciences in which we have published articles are: Journal of Theoretical Biology and Mathematical Biosciences.

Indicate the major conferences in the field and, for each, indicate the number of papers coauthored by members of the project-team that have been accepted during the evaluation period.

1. There are many prestigious conferences in Statistics. Our team is involved in the yearly meeting of Société Francophone de Statistiques, which is the major event of its kind in the French-speaking world.
2. We have also been speakers (or organizers) in invited sessions of prestigious conferences in Probability Theory: International Congress of Mathematicians satellite conference (2010), Stochastic Processes ans their Applications (2011), European Congress of Mathematics (2012), European Meeting of Statisticians (2013).


3.2 Software

*Socio-economic index:* A R package called SesIndexCreatoR has been written by M. Lalloué and J-M. Monnez in order to implement the socio-economic index for health inequalities described at Section 2.4.3.

3.3 Valorization and technology transfert (Socio-economic impact and transfer)

*Start-up project by T. Bastogne:*

- Industrial partner: Cybernano (Contract Research Organization in NanoMedicine).
- Status: SAS created in July 2013.
- Comments: Cybernano has received the "emergence" award in 2012 from the French Research ministry for the creation of start-up based on innovative technology. Cybernano proposes innovating services to reduce the cost and control the risk during the preclinical development of nanoparticles in oncology applications. The engineering approach used by this spin-off is strongly based on the use of suited mathematical models. Concerning the BIGS program for the next four years, Cybernano is particularly interested by two items: (i) Development of a Matlab toolbox for cost-effectiveness analysis in clinical studies. (ii) Development of algorithms for treatment planning systems associated with nano-therapies.

3.4 Teaching

BIGS is a team whose composition includes University staff only. All members teach numerous courses, ranging from L1 to M2 levels.

- Samy Tindel (192h, University)
- Thierry Bastogne (192h, University).
- Sandie Ferrigno (192h, Engineering schools)
- Céline Lacaux (192h, Engineering schools)
- Jean-Marie Monnez (192h, IUT and University)
- Aurélie Muller-Gueudin (192h, Engineering schools)
• Pierre Vallois (192h, University)
• Sophie Wantz-Mézières (192h, IUT)

**Note:** An innovative teaching program specialized in Biocybernetics (L3,M1,M2) was proposed by Thierry Bastogne and will start in Sep. 2013 at the Faculté des Sciences et Technologies.

### 3.5 Visibility

In 2011-2012, our team has been involved in the organization of the following scientific events:

- **STochastic ANalysis days, 9-11 May, 2012:** A 3 days international meeting gathering some of the best specialists in stochastic analysis and applications (including statistics and fractional fields). Organizers: C. Lacaux, I. Nourdin, S. Tindel.


- **Journée Identification de systèmes biologiques, 11 April 2013,** within the framework of the MACS network. Organizer: Thierry Bastogne.

- **Seminar on systems biology for biologists and engineers,** within the framework of Université de la Grande Région, Lultzhausen, Luxembourg, 19-20 June 2011. Organizer: Thierry Bastogne.

- **Weekly Biostats Seminar at IECN,** organized by Sandie Ferrigno (until 2012) and Aurélie Muller-Gueudin. See [http://www.iecn.u-nancy.fr/ muller/gt.html](http://www.iecn.u-nancy.fr/ muller/gt.html).

### 4 Funding

#### 4.1 Funding external to Inria

<table>
<thead>
<tr>
<th>£ initiatives</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>National initiatives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANR</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>PEPS</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>INCa</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>European projects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Scholarships</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhD</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Post Doc</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>14</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>
4.2 Inria competitive funding

<table>
<thead>
<tr>
<th># grants</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scholarships</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internship</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Post Doc</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

\(^4\) INRIA postdoctoral research contract (‘post-doc Inria sur subvention’).

National initiatives


- **PDTX** (2010-2013), Active Nanoplatforms for Photodynamic Therapy, Funding organism: **French National Agency for Research (ANR)**, Leader: M. Verelst (U. Paul Sabatier, Toulouse).


- **ECRU** (2010-2012), Exploration des Chemins Rugueux, Funding organism: **French National Agency for Research (ANR)**, Leader: M. Gubinelli (Paris Dauphine).

- **MASTERIE** (2010-2013), Malliavin Stein Random Irregular Equation, Funding organism: **French National Agency for Research (ANR)**, Leader: F. Russo (ENSTA, Paris).

- **Computation of profit probabilities in sports gambling** (2011-2012), Funding organism: CNRS - **PEPS**, Leader: A. Muller-Gueudin. Two Internships involved.

European projects

- **I-DERBI** (2010-2012), Infrastructure Distribuée d’Enseignement et de Recherche en Biologie Intégrative, Funding organism: **Université de la Grande Région** (Nancy-Université, Université de Liège, Université du Luxembourg, Universität des Saarlandes), Leader: E. Bullinger (GIGA, Université de Liège), T. Bastogne (CRAN, U. Lorraine).

Other funding

PhD funding


Internships


• Yosra Chemli: Applicability of an Exponential-Linear (E-L) model to describe the in vitro cell responses in photodynamic therapy. Advisor: T. Bastogne.

• Kevin Ziegelmeyer: Data Analysis for liver cirrhosis prediction. Advisor: A. Muller-Gueudin.

5 Objectives for the next four years

Since our team is a relatively new one, our main objective is to strengthen the dynamics we have launched in data analysis and stochastic models in a biomedical context. We shall also make an additional effort in order to focus, as a group, on our photodynamic therapy application. Let us now detail our projects for the next future according to the scheme we have already used at Section 2.3.
5.1 Analysis of high dimensional data

This research line gives raise to interesting theoretical developments, but represents also one of the main interaction flow with other discipline, with a strong application oriented flavor. We shall make progresses in both directions in the next 4 years.

5.1.1 Online data analysis

We still have several interesting issues to tackle in this direction. Let us mention for instance the 2 following problems:

(1) Truncated stochastic approximation of a dynamic generalized linear model under convex constraints. Recall that we have already investigated several aspect of online data analysis, like principal component analysis or $k$-medians. Romain Bar’s PhD dissertation (still under progress) focuses on dynamical versions of linear models, and gives raise to the following natural generalization: consider a sequence of random variables $(Y_n, Z_n)$ in $\mathbb{R}^p \times \mathbb{R}$ and a sequence of real $(m, p)$ matrices $(\Gamma_n)$. Suppose that there exist a function $h : \mathbb{R} \rightarrow \mathbb{R}$ and a parameter $\theta$ in $\mathbb{R}^m$ such that

$$E[Z_n|Y_n] = h(Y_n^\prime \Gamma_n^\prime \theta) .$$

Suppose moreover that $\theta$ belongs to a non empty convex subset $K$. The aim is to estimate $\theta$, by means of a truncated stochastic approximation process. Notice that our results will then be applied in both generalized linear regression analysis and in partial principal component analysis.

(2) Large data set clustering with an averaged $k$-kernel stochastic algorithm. Consider the problem of partitioning a large and high dimensional data set in a fixed number $k$ of classes. Furthermore, we wish to be able to consider a sequential flow of data, and aim at a fast algorithm.

In this context, we suppose that vectors of data are i.i.d. observations of a random vector. After defining a representative (kernel) of each cluster, a dissimilarity measure and a classification criterion, we will define a sequential $k$-kernel algorithm as a stochastic gradient method of averaging to accelerate its convergence. Particular cases of this type of approach include the sequential $k$-means algorithm of MacQueen and the sequential $k$-medians algorithm of Cardot, Cénac, Monnez [17].

5.1.2 Local polynomial regression

The two main objectives we have in this direction can be summarized as follows:

(1) Rates of convergence in local polynomial regression. As mentioned at Section 2.3.1 this work is a collaboration with Myriam Maumy-Bertrand (assistant professor at the University of Strasbourg). Recall that we have established exact rate of strong uniform consistency for the local linear estimator of the conditional distribution function $F(y|X = x)$. We now wish to extend our results to obtain exact rates of strong uniform consistency for the local linear estimator of other conditional quantities: obvious examples are the conditional mean $E(Y|X)$, and the conditional quantiles $q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\}$, for $\alpha \in (0, 1)$.

Another crucial problem in the non parametric regression methods is the choice of the bandwidth parameter $h$. It is common in practice to choose $h > 0$ so as to minimize asymptotically the mean square error (MSE) or the mean integrated square error (MISE). This minimization leads to an optimal choice of $h$ of the form $h_n = C(X_1, \ldots, X_n)n^{-1/5}$, where $n$ is the sample size, and $X_1, \ldots, X_n$ are the $n$ independent copies of the random variable $X$. 28
This bandwidth is called a *data-driven bandwidth* to highlight its dependence to the data. In this context, we plan to establish consistency of the local linear estimator when the bandwidth \( h \) is allowed to range in a small interval which may decrease in length with the sample size. Such a result would be immediately applicable to prove uniform consistency of the local linear estimator when the bandwidth is a data-driven bandwidth \( h_n = C(X_1, \ldots, X_n)n^{-1/3} \).

At the same time, we will try to obtain a central limit theorem for the process defined by \((\hat{F}_n(y|x))_{y \in \mathbb{R}}\). Indeed, a central limit theorem would allow us to propose confidence intervals or tests for the cumulative conditional distribution function \( y \mapsto F(y|x) \). It should be noted that even for i.i.d. data, at our knowledge, no asymptotic normality has so far been established (except in the U-statistic paradigm, with the PhD of Mint El Mouvid, 2000). Note also that the pointwise asymptotic normality of \( \hat{F}_n(y|x) \) (for a fixed \( y \)) is useless, since such a result cannot be extended to functionals of the process \((\hat{F}_n(y|x))_{y \in \mathbb{R}}\) (for instance, the conditional mean and the conditional quantiles). We will then consider the process \((\hat{F}_n(y|x))_{y \in \mathbb{R}}\) as a process defined in the Skorohod space \( D(\mathbb{R}) \).

Also notice that we plan to reactivate our contact with Professor Foliguet (Maternité Régionale de Nancy). We will continue to collaborate with him, in order to estimate growing curves of the fetal weight and other fetal quantities thanks to our local regression techniques.

(2) **Local polynomial direction.** Many clinical trials and other medical studies involve responses that might be considered to have a normal distribution. However, this is not invariably the case and models based on this distribution are often indiscriminately applied to data which might be better handled otherwise. This is especially true for discrete data. An approach which may yields models that are more biologically reasonable in many situations is to use generalized linear models (GLM).

In statistical theory, generalized linear models were formulated by John Nelder and Robert Wedderburn (1972) as a way of unifying various other statistical models including for examples linear regression, logistic regression and poisson regression. Such a technique was developed by McCullagh and Nelder (1989).

In order to understand better the generalized linear model, let us go back to the linear model itself. It is well known that its general form can be written as \( Y = X\beta + \epsilon \), where \( Y \) is a dependent variable, \( X \) is a matrix of \( p \) independent variables or predictors, \( \beta \) is a \((p+1)\) vector of unknown parameters and \( \epsilon \) is a zero-mean stochastic disturbance. Typically, \( \epsilon \) is assumed to be independent across observations with constant variance \( \sigma^2 \) and normally distributed.

That is, the normal linear regression model is characterized by the following features:

- **Stochastic component**: \( Y \) is usually assumed to have independent normal distribution of the form \( Y \sim \mathcal{N}(\mu, \sigma^2) \).

- **Systematic component**: \( X \) combines linearly with the coefficients to form the linear predictor \( \eta = X\beta \).

- **Link between the random and systematic component**: the linear predictor \( X\beta = \eta \) is a function of the mean parameter \( \mu \) via a link function \( g(\mu) \). Note that for the usual linear model, \( g \) is simply the identity.

Generalized linear models follows from two extensions of this setup in the following sense:

- The stochastic component follows other distributions than the Gaussian.

- The link functions are different from the identity.
Notice that those models are well-suited to analyze dependences between variables following distributions in the so-called exponential family, like Poisson, Binomial and Gamma distributions. In practice, link functions are chosen such that the inverse link, $\mu = g^{-1}(\eta)$ is easily computed. For instance, for binomial data, logit and probit link functions are commonly used.

Our aim in this project is to use generalized linear models in order to extend our global test of goodness-of-fit to a wide range of models used in biological and medical applications. We wish to use the cumulative conditional distribution $F(y|X = x)$ again, which embodies all the information about the joint behavior of two random variables. The expected outcome is a global goodness of fit test for the relationship between two random variables in the exponential family. The test will compare a nonparametric estimator of the cumulative distribution function with the value of the cumulative distribution function under the null hypothesis.

5.1.3 Data analysis in biomedical context

We shall keep on working on the following interdisciplinary projects in the next future:

(1) Medical decision and telemedicine in the monitoring of heart failure. This project fits in the general framework of telemedicine and more precisely in the monitoring of heart failure patients. From measurements performed automatically and daily on a patient at home through a new process under development at the Pluri-Thematic Clinical Investigation Center of the University Hospital of Nancy, the aim is to propose therapeutic adjustments to improve the prognosis of patient in order to increase his chances of survival or to avoid useless hospitalizations.

The patient’s condition and its evolution are determined by the initial values of his biological or clinical parameters, as well as those collected throughout his follow-up. The treatments are intended to stabilize or change the values of parameters in order to avoid the occurrence of adverse events, in particular the death of the patient. This is why the first part of the study will consist in building survival scores or hospitalization scores according to the values of biological or clinical parameters.

In a second part, we will seek to build models of the evolution of the values of biological or clinical parameters depending on treatments (average or cumulative drug doses, drug combinations) and patients’ characteristics. This will allow to predict the potential effect of an adjustment proposal or modification of treatment and then predict a new survival score to conclude the relevance or not of the proposed medication. The physician will have this help to confirm or change his decision which belongs finally to him.

In order to carry out this study, we shall resort to a wide range of classic and recent methods of data analysis (in particular discriminant analysis) without a priori: several methods will be used, compared and selected according to their performance in the aforementioned applications. Interesting links with our recent online data analysis activity are obviously an extra motivation for us, and a PhD student should be involved in the project.

(2) Cohort analysis. As mentioned at Section 2.3.1, this is a joint project with the INSERM team U954 and the INRIA team Orpailleur. Our main task is to describe the complex interactions between genetic, phenotypic and biologic variables available in medical cohorts, in different contexts (cognitive decline; inflammatory intestinal diseases; liver cancer).

We shall first resort to the existing methods given in the literature, for the analysis of qualitative and quantitative data. Indeed, we have to describe links between qualitative and quantitative data, and we plan to perform a thorough study of available tools in our high dimensional context:

1. Exploratory methods, or factorial models.
2. Regression models to predict qualitative variable by the use of qualitative or quantitative factors.

We will also describe the association variables (via dispersion measures, prediction measures, Goodman and Kruskal’s $\lambda$ coefficients, Cohen’s $\kappa$ coefficient).

In a second step of the study, we will test non association or independence between variables. The objective is to develop new methods, adapted to the cohorts under consideration (matching cases/controls, high number of individuals, high number of explicative variables, missing data problem). This is where the joint work with Orpailleur will prove to be particularly useful.

After having identified and extracted the relevant variables, we shall give a model in order to classify the data. The proposed models will allow us to identify subgroups of individuals, with common genetic, biologic and phenotypic characteristics.

5.2 Modeling for complex and biological systems

For this part of our project, photodynamic therapy will still be our central application. However, we shall maintain an activity in interesting directions such as bacteriophage and gene networks.

5.2.1 Photodynamic therapy

We still have a lot of modeling work in this direction, in collaboration with biomedical teams. Here are the main projects we have in mind for the next period:

(1) Real time PDT controller. Based on some recent innovative results that have been submitted for a patent (No.1261339, Nov. 2012), our main objective for this project is to develop the model-based control of the photodynamic therapy. This includes the following demanding tasks:

(i) Material innovation. In collaboration with the INSERM U703 (Thérapies interventionnelles assistées par l’image et la simulation: www.u703.fr), we would like to associate two of our innovations (PDT realtime controller: CRAN and an optical fiber-based tissue: U703) in a prototype devoted to the treatment of non-melanoma skin cancers by photodynamic therapy. The PDT controller developed in Nancy is a world innovation which improves the reproducibility of the therapeutic response. This innovative system is based on a state-space observer. Numerous modeling improvements are planned to be developed and implemented in the existing technical platform.

(ii) Model-based controller. The new device we have just patented would allow patients to be treated at home with a reduced cost. Moreover, since the new system uses low irradiance impulses we expect to reduce significantly side effects such as the pain sensation during the illumination phases. Since each part has already been tested separately, the risk associated with this technological research is minimized. For the development part realized in Nancy, we need a system control engineer (IR) for the realtime implementation of the controller, the development of the technical platform and the \textit{in vitro} & \textit{in vivo} validation tests of new model-based control strategies.

(iii) Optical fiber positioning optimization. Another ambition is to use the PDT realtime controller to treat brain tumors by interstitial PDT. To this aim, we need to carry on our research works on the optimization of the optical fibers positioning within the brain. This item will require at least one 2-years post-doc student.
5.2.2 Tumor growth modeling

Up to now, we have focused on discrete time and space models for tumor growth in relation to PDT applications. However, continuous models are sometimes easier to handle and more realistic from a biological point of view. Indeed, a cancer tumor can be represented for simplicity as an aggregate of cancer cells, each cell behaving according to the same discrete model and independently of the others. Therefore to measure its size evolution, it seems natural to use tools coming from dynamics of population, for instance the logistic model. This deterministic framework is well-known but the stochastic one is worthy of interest. We suppose that the size $V_t$ at time $t$ of the tumor is a diffusion process of the type:

$$
\begin{cases}
\frac{dV_t}{dt} = r V_t \left(1 - \frac{V_t}{\kappa}\right) - c V_t + \beta V_t dB_t \\
V_0 = v > 0
\end{cases}
$$

(7)

where $(B_t)_{t \geq 0}$ is a standard brownian motion starting from zero. Then (i) We define a family of time continuous Markov chains which models the evolution of the rate of malignant cells and approximate (under some conditions) the diffusion process $(V_t)$. (ii) We study in depth the solution to equation (7). This diffusion process leaves between two frontiers: 0 and $\kappa$. In this stochastic setting, the role of $\kappa$ is not so clear and we contribute to understand it. We describe the asymptotic behavior of the diffusion according to the values of the parameters. The tools we resort to are boundary classification criteria and Laplace transform of the hitting time to biological worthwhile level. We believe we are able in particular to express the mean of the hitting time.

The next step in this project can be summarized as follows: at this point in our investigations on tumor growth modelization, we have identified a pertinent and consistent model. Nevertheless our study remains theoretical. A statistical estimation of the parameters $r, \kappa, c, \beta$ is thus in order. This would permit to apply our model to real data. A further objective could be to consider a more complex form for the logistic term, see e.g. Schurtz (2007).

5.2.3 Light transport in tissues with probabilistic methods

This project is related with the optimization problem alluded to at Section 5.2.1. Indeed, in order to optimize the intra-cerebral position of our optical fiber, two fundamental questions have to be answered:

1. What is the optimal shape and position of the light source in order to optimize the damage on malignant cells?

2. Is there a way to identify the physical parameters of the tissue which drive the light propagation?

Notice that we are obviously not the first ones to address these issues, and there is nowadays a consensus in favor of the algorithm proposed by L. Wang and S. L. Jacques for the simulation of light transport in biological tissues. However, our starting point is the observation that the usual methods slightly lack of formalism and miss formal representations that answer the questions of identifiability. By proposing a completely rigorous framework to the simulation we thus wish to answer some convergence questions in a sharp way, and also open the way to new improvements of the method.

Let us specify a little our model for the light propagation, and formalize its probabilistic representation. The physical phenomenon involves three processes: absorption, emission and scattering, which are described by the so-called equation of radiative transfer. In the particular
case of an homogeneous tissues, this equation takes the following form: let $D = \mathbb{R}^3$ be the set of positions in the tissue and $S^2$ be the unit sphere in $\mathbb{R}^3$. The quantity of light at $x \in \mathbb{R}^3$ in the direction $\omega \in S^2$ is denoted by $L(x, \omega)$ and satisfies:

$$L(x, \omega) = L_e(x, \omega) + TL(x, \omega),$$

where $L_e(x, \omega)$ is the emitted light from $x$ in direction $\omega$ and $T$ is an integral operator over $S^2 \times \mathbb{R}^3$ whose exact expression is omitted for sake of conciseness.

It can be shown that the operator $T$ is a strict contraction in a proper function space, and thus we can expand $(I - T)^{-1}$ in Neumann series. This allows us to formally solve (8) and to write it as $L(x, \omega) = \sum_{n \geq 0} T^n L_e(x, \omega)$. Some additional elementary considerations allow then to write $L(x, \omega)$ as the expected value of a random variable of the form $G_{x,\omega}(Y)$:

$$L(x, \omega) = E\left[G_{x,\omega}(Y)\right],$$

where $\{G_{x,\omega}; x \in \mathbb{R}^3, \omega \in S^2\}$ is a family of bounded functions. The random variable $Y$ lives in an infinite dimensional space of piecewise linear paths of arbitrary length (like light ray themselves), but it admits a density with an explicit expression and it can be simulated exactly. Therefore, $L(x, \omega)$ can be approximated by means of a Monte Carlo method.

With these preliminaries in mind, let us now describe some directions in which we wish to improve the original algorithm naturally induced by (9): (i) Improvements of the simulations and MCMC-based algorithm. (ii) Specific treatment of inhomogeneous tissues, especially in presence of tumors. (iii) Inverse problem in order to retrieve optical parameters of the tissue. (iv) Use of particle systems in the MCMC framework. This project is obviously a long range one, and shall require the help of 1-2 postdoc students.

5.2.4 Controlled therapies

Related to the controlled PDT project of Section 5.2.1, we are interested in a simple enough mathematical model able to predict biological damages of ionizing treatments in brain tumors. In order to describe this project, let us refer again to [46, 45], in which a multinomial model in discrete time is proposed in order to describe the effect of radiotherapy on tumor cells. Those articles are based on a discrete-time Markov chain, which takes into account cell repair and cell damage heterogeneity. The model relies on the target theory in radiobiology and assumes that a cancer cell contains $m$ targets which must be deactivated to produce cell death. Therefore each cell presents at any time a particular state expressing the number of deactivated targets. The model takes into account the number of cells of each state type at a given instant $t$, and can be simulated exactly.

Our aim is to introduce an optimized planning for the radiotherapy treatment according to the evolution of the cells states. Indeed, our theoretical target is to kill the highest number of cancerous cells while sparing healthy cells. For that purpose, we consider two cell populations: cancerous cells and healthy ones. The size of the global population remains constant in a fixed volume. Our goal is to control two parameters: the daily dose $A_t$ and the last treatment time $\tau$.

The methodology we have in mind in order to handle this problem is based on dynamical programming, for which we refer e.g. to Szepesvari (2010). Such an application of the reinforcement learning method is quite new; to the best of our knowledge the only reference in this direction is Kim, Ghate and Phillips (2012), in a much simpler setting than ours. Our model is based on the following three objects:

- The State Process $(X_t)_{t \in \mathbb{N}^*}$, which represents the populations evolution (different representations are possible, for instance the current state of each cells).
• The Action Process $\{(A_t)_{t \in \mathbb{N}}^\ast\}$, which models the dose-monitoring (living in a finite set).
• The Reward Process $\{(R_t)_{t \in \mathbb{N}}^\ast\}$, giving the quantity to maximize at each step. This process summarizes in some way the constraints of the problem.

We are thus within the landmark of application of Markov Decision Processes, and our aim is to determine an optimal strategy that maximizes the mean of the reward process. Notice that this project will be a collaboration between our team (J-L. Marchand, S. Mézières, S. Tindel) and B. Scherrer (MAIA team at INRIA).

5.2.5 Cost-Effectiveness Analysis of Photodynamic Therapy

The clinical development of PDT requires to prove its relevance in comparison with competing treatments. We will particularly focus on precancerous and cancerous skin lesions for which the main alternative therapies are surgery and cryotherapy. Comparison must be performed not only in terms of therapeutic efficiency but also in terms of economical cost and quality of life for the patient. A cost-effectiveness analysis is particularly suited to determine which treatment is the most cost-effective.

In the context of pharmacoeconomics, the cost-effectiveness of a therapeutic is the ratio of the cost of the intervention to a relevant measure of its effect. Cost refers to the resource expended for the intervention and the measure of effects depends on the intervention being considered. Examples include the number of people cured of a disease and the number of symptom-free days experienced by a patient.

Markov modeling has become an accepted tool for medical decision and cost-effectiveness analyses (CEA), and cohort analysis has become the standard procedure for solving such models to obtain expected costs, life years, and quality-adjusted life years (see e.g the book by Briggs and Claxton, 2006). However, recent works have identified several problems with economic evaluations undertaken alongside controlled trials that can have potentially serious impacts on the ability of decision makers to draw valid conclusions. At the same time, the use of cost-effectiveness models has been drawn into question, due to the alleged arbitrary nature of their construction. This has led researchers to try and identify ways of improving the quality of cost-effectiveness models through identifying best practice, producing guidelines for peer review and identifying tests of validity.

Our objective for this project is to contribute to improve the quality of cost-effectiveness models. A PhD thesis is starting in September 2013 on this subject in collaboration with clinicians of the CRAN (Health-Biology-Signal department) to assess the relevance of the proposed innovative modeling method in practice.

5.2.6 Modelisation of gene networks with piecewise deterministic processes

As mentioned in Section 2.5.3, our collaboration with A. Debussche (ENS Bretagne) and O. Radulescu (University of Montpellier 2) calls for and application oriented step, that we wish to take in the next future.

Indeed, we have established in [58, 27] convergence results that can be used to simplify the stochastic dynamics of gene network models arising in molecular biology. In the sequel of these works, we want to apply our theoretical models with gene expression data, in a gene network concerning the bacterium *Bacillus subtilis*. The data are currently being obtained by the INSERM team of Catherine Royer, in Montpellier, with the collaboration of the European consortium Bacillus Systems Biology. We want to study the parameters identifiability and to develop bayesian methods for the estimation of the parameters in this context, the interest of bayesian methods lying in the ability to include a priori informations, given by Biologists.
Let us be more specific about the methodology we shall use for this project: The Markovian models we have proposed in [58, 27] contain a lot of parameters which describe the kinetics of the chemical reactions. However, the kinetics parameters are not exhaustively known. The bayesian methods we wish to introduce make use of Markov Chains Monte-Carlo (MCMC) techniques, and allow to estimate the parameters when the set of parameters lives in a domain with low dimension. But the MCMC techniques become useless in practice, when the set of parameters lies in a high dimensional domain, since the MCMC algorithm cannot visit all the space of the parameters. This problem has been partially solved by the particular Monte-Carlo algorithms. Despite this improvement, the computation time is too long. The method we have in mind in order to reach identifiability is thus to replace some parameters by a linear combination of them. This technique will obviously yield reduced models.

Moreover, we have proposed in [58, 27] several approximated models. We shall introduce a method allowing to select models which can explain the data, in order to compare the molecular mechanisms. More precisely, we will resort to the Bayesian model selection, which can be briefly summarized as follows: let $M_1$ and $M_2$ be two models of molecular mechanisms, among which we wish to select a model explaining the data $x_0$ in an appropriate way. The Bayes factor is then defined as

$$B_{12} = \frac{\pi(M_1|x_0)/\pi(M_2|x_0)}{\pi(M_1)/\pi(M_2)},$$

where $\pi(M_i)$ is the a priori and $\pi(M_i|x_0)$ is the a posteriori distribution of model $M_i$. The Bayes factor $B_{12}$ can be interpreted as a summary of the evidence provided by the data in favor of one statistical model over another. Note that the models being compared do not need to be nested, which is important in our setting.

5.2.7 Bacteriophage systems

After discussion with our partner M. Llagostera (Departament de Microbiologia, Universitat Autònoma de Barcelona), she wishes us to address the problem of modeling bacteriophages therapies in a different experimental context than the one dealt with in [3].

Indeed, in certain situations our microorganisms are somehow trapped by their host media, a phenomenon which reduces dramatically their mobility (a typical example of this situation is given by a system salmonella/phage living in intestine walls). In this context, one can no longer rely on ODEs or SDEs methods in order to accurately model the system. A natural alternative is to introduce some particle models living on a discrete graph (say $\mathbb{Z}^2$), where mobility only happens by means of two mechanisms:

1. Artificial introduction of phages (consider the example of cattle for which bacteriophages are mixed with food).

2. Reproduction of both bacteria and bacteriophages, which sends descendent on nearest neighbor sites.

We plan to model, simulate and study theoretically this kind of behavior thanks to particle systems techniques.

5.3 Inference for Gaussian systems

This section is devoted to describe our projects in Gaussian analysis related to anomalous protein fluctuations and osteoporosis. Its is divided into two subprojects.
5.3.1 Gaussian operator scaling random fields

This section presents a project on osteoporosis, which is linked to the ANR MATAIM. The proceedings Biermé-Richard-Benhamou (2009) is a first step to model trabecular bone x-ray images by Gaussian operator scaling random fields, that is Gaussian random fields which satisfy the scaling property \( \gamma \) (with \( d = 2 \)). In view of this promising work and with the aim to help to the diagnostic of the osteoporosis, we are then interested in the estimation of the parameter \( E \in M_2(R) \), which characterizes the operator scaling property, the anisotropy and the Hölder regularity properties of the Gaussian field proposed as model. Specifically, we shall estimate the eigenvalues and eigenvectors of \( E \).

Let us notice that this project also leads to real world biological application, in collaboration with some practitioners of CHU Orléans. Those Biologists have been members of the ANR MATAIM as H. Biermé and C. Lacaux, and within this framework they have built an important database concerning radiographs of calcaneus. Our estimation method is designed to help them to detect automatically osteoporosis.

5.3.2 Inference for Gaussian differential systems

Recall that we are still interested in parameter estimation for equation (5). In spite of some extensive work during the last years, we are still far from a full generality in this demanding task. The ideas we wish to implement, possibly with the help of a PhD student, are the following:

- Normalization of numerical schemes for equation (5), since numerical schemes are always a part of the implementation for statistical procedures.
- Gaussian upper and lower bounds for the law of the solution \( X_t \) to equation (5).
- Theoretical study of a full maximum likelihood related to our system, by going back to the underlying Wiener process.
- Interactions with practitioners of the field. In particular, some preliminary contacts have been established with S. Kou (Harvard).

6 Bibliography of the project-team

References


[50] Céline Lacaux and Renaud Marty. From invariance principles to a class of multifractional fields related to fractional sheets. 2011.


