IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2013

Project-Team BIGS

Biology, genetics and statistics

IN COLLABORATION WITH: Institut Elie Cartan de Lorraine

RESEARCH CENTER
Nancy - Grand Est

THEME
Computational Biology

Table of contents

Project-Team BIGS

Keywords: Computational Biology, Data Analysis, Identification, Statistical Methods, Stochastic Modeling

Creation of the Team: 2009 January 01*, updated into Project-Team:* 2011 January 01.

1. Members

Faculty Members

Samy Tindel [Team leader, Univ. Lorraine, Full Professor, HdR]
Thierry Bastogne [Univ. Lorraine, Full Professor, HdR]
Sandie Ferrigno [Univ. Lorraine, Associate Professor]
Céline Lacaux [Univ. Lorraine, Associate Professor, HdR]
Sophie Mezieres [Univ. Lorraine, Associate Professor]
Jean-Marie Monnez [Univ. Lorraine, Full Professor]
Aurélie Muller-Gueudin [Univ. Lorraine, Associate Professor]
Pierre Vallois [Univ. Lorraine, Full Professor]

PhD Students

Benoît Lalloué [Univ. Lorraine, until Dec 2013]
Romain Bar [IECL, until Dec 2013]
Kevin Duarte [IECL, from July 2013]
Yann Petot [IECL, from September 2013]

Post-Doctoral Fellows

Mireia Besalú [Inria, until Apr 2013]
Jean-Louis Marchand [Univ. Lorraine, until Sep 2013]
Laura Vinckenbosch [Inria]
Juan-José Viquez Rodriguez [Inria, from Jan 2013]

Administrative Assistant

Isabelle Blanchard [Inria]

2. Overall Objectives

2.1. Overall Objectives

BIGS is a team labeled by Inria, by CNRS and by Université de Lorraine, via the Institut Élie Cartan (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:

•*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.

•*Photodynamic therapy:* Since 1988, some control system scientists and biologists at the Centre de Recherche en Automatique de Nancy (CRAN in short) 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. 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.

•*Estimation for complex biological systems:* Numerous biological systems are accurately described by multi-dimensional noisy differential equations driven by Gaussian processes (beyond the realm of Brownian motion) or by fractional fields, for which asymptotic properties and parameter estimation are fruitful informations. We are thus interested in studying this kind of systems, having in mind 3 specific applications of interest for us: (i) Bacteriophage systems (ii) Random fluctuation of nanoparticles. (iii) Automatic detection of osteoporosis.

3. Research Program

3.1. Online data analysis

Participants: J-M. Monnez, R. Bar, P. Vallois. Generally speaking, there exists an overwhelming amount of articles dealing with the analysis of high dimensional data. Indeed, this is one of the major challenges in statistics today, motivated by internet or biostatistics applications. Within this global picture, the problem of classification or dimension reduction of online data can be traced back at least to a seminal paper by Mac Queen [53], in which the k-means algorithm is introduced. This popular algorithm, constructed for classification purposes, consists in a stepwise updating of the centers of some classes according to a stream of data entering into the system. The literature on the topic has been growing then rapidly since the beginning of the 90's.

Our point of view on the topic relies on the so-called *french data analysis school*, and more specifically on Factorial Analysis tools. In this context, it was then rapidly seen that stochastic approximation was an essential tool (see Lebart's paper [50]), which allows to approximate eigenvectors in a stepwise manner. A systematic study of Principal Component and Factorial Analysis has then been leaded by Monnez in the series of papers [56], [54], [55], in which many aspects of convergences of online processes are analyzed thanks to the stochastic approximation techniques.

3.2. Local regression techniques

Participants: S. Ferrigno, A. Muller-Gueudin. 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 which turns out to be useful 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 models such as the regression model with random regressors:

$$Y = g(X, \theta) + \sigma(X)\epsilon.$$

Many assumptions must be made to reach it as a possible model. Some require 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 ϵ. 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 an entertained structural part $g(\cdot, \theta)$. Some authors have also proposed tests about the variance term $\sigma(\cdot)$ (cf. [51]). Procedures testing the normality of the ϵ_i are given, but for other assumptions much less work has been done. Therefore the need of a global test which can evaluate the validity of a global structure emerges quite naturally.

With these preliminaries in mind, let us observe that one quantity which embodies all the information about the joint behavior of (X, Y) is the cumulative conditional distribution function, defined by

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

The (nonparametric) estimation of this function is thus of primary importance. To this aim, notice that modern estimators are usually based on the local polynomial approach, which has been recognized as superior to classical estimates based on the Nadaraya-Watson approach, and are as good as the recent versions based on spline and other methods. In some recent works [41], [42], we address the following questions:

- Construction of a global test by means of Cramer-von Mises statistic.
- Optimal bandwidth of the kernel used for approximation purposes.

We also obtain sharp estimates on the conditional distribution function in [43].

3.3. Stochastic modeling for complex and biological systems

In most biological contexts, mathematics turn out to be useful in producing accurate models with dual objectives: they should be simple enough and meaningful for the biologist on the one hand, and they should provide some insight on the biological phenomenon at stake on the other hand. We have focused on this kind of issue in various contexts that we shall summarize below.

Photodynamic Therapy: Photodynamic therapy induces a huge demand of interconnected mathematical systems, among which we have studied recently the following ones:

• The tumor growth model is of crucial importance in order to understand the behavior of the whole therapy. We have considered the tumor growth as a stochastic equation, for which we have handled the problem uncertainties on the measure times [27] as well as mixed effects for parameter estimation.

• Another important aspect to quantify for PDT calibration is the response to radiotherapy treatments. There are several valid mathematical ways to describe this process, among which we distinguish the so-called hit model. This model assumes that whenever a group of sensitive targets (chromosomes, membrane) in the cell are reached by a sufficient number of radiations, then the cell is inactivated and dies. We have elaborated on this scheme in order to take into account two additional facts: (i) The reduction of the cell situation to a two-state model might be an oversimplification. (ii) Several doses of radiations are inoculated as time passes. These observations have leaded us to introduce a new model based on multi-state Markov chains arguments (Keinj & al, 2012), in which cell proliferation can be incorporated.

Bacteriophage therapy: Let us mention a starting collaboration between BIGS and the Genetics and Microbiology department at the Universitat Autònoma de Barcelona, on the modeling of bacteriophage therapies. The main objective here is to describe how a certain family of benign viruses is able to weaken a bacterium induced disease, which naturally leads to the introduction of a noisy predator-prey system of equations. It should be mentioned that some similar problems have been treated (in a rather informal way, invoking a linearization procedure) by Carletti in [34]. These tools cannot be applied directly to our system, and our methods are based on concentration and large deviations techniques (on which we already had an expertise [57], [60]) in order to combine convergence to equilibrium for the deterministic system and deviations of the stochastic system. Notice that A. Muller-Gueudin is also working with A. Debussche and O. Radulescu on a related topic [37], namely the convergence of a model of cellular biochemical reactions.

Gaussian signals: Nature provides us with many examples of systems such that the observed signal has a given Hölder regularity, which does not correspond to the one we might expect from a system driven by ordinary Brownian motion. This situation is commonly handled by noisy equations driven by Gaussian processes such as fractional Brownian motion or (in higher dimensions of the parameter) fractional fields.

The basic aspects of differential equations driven by a fractional Brownian motion (fBm) and other Gaussian processes are now well understood, mainly thanks to the so-called *rough paths* tools [52], but also invoking the Russo-Vallois integration techniques [59]. The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in [38].

Fractional fields are very often used to model irregular phenomena which exhibit a scale invariance property, fractional Brownian motion being the historical fractional model. Nevertheless, its isotropy property is a serious drawback for instance in hydrology or in medecine (see [33]). Moreover, the fractional Brownian motion cannot be used to model some phenomena for which the regularity varies with time. Hence, many generalization (gaussian or not) of this model has been recently proposed, see for instance [28] for some Gaussian locally self-similar fields, [46] for some non-Gaussian models, [31] for anisotropic models.

Our team has thus contributed [36], [47], [46], [48], [58] and still contributes [30], [32], [31], [49], [44] to this theoretical study: Hölder continuity, fractal dimensions, existence and uniqueness results for differential equations, study of the laws to quote a few examples. As we shall see below, this line of investigation also has some impact in terms of applications: we shall discuss how we plan to apply our results to osteoporosis on the one hand and to fluctuations within protein molecules on the other hand.

3.4. Parameter identifiability and estimation

When one desires to confront theoretical probabilistic models with real data, statistical tools are obviously crucial. We have focused on two of them: parameter identifiability and parameter estimation.

Parameter identifiability [62] deals with the possibility to give a unique value to each parameter of a mathematical model structure in inverse problems. There are many methods for testing models for identifiability: Laplace transform, similarity transform, Taylor series, local state isomorphism or elimination theory. Most of the current approaches are devoted to *a priori* identifiability and are based on algebraic techniques. We are particularly concerned with *a posteriori* identifiability, *i.e.* after experiments or in a constrained experimental framework and the link with experimental design techniques. Our approach is based on statistical techniques through the use of variance-based methods. These techniques are strongly connected with global sensitivity approaches and Monte Carlo methods.

The parameter estimation for a family of probability laws has a very long story in statistics, and we refer to [29] for an elegant overview of the topic. Moving to the references more closely related to our specific projects, let us recall first that the mathematical description of photodynamic therapy can be split up into three parametric models : the uptake model (pharmacokinetics of the photosensitizing drug into cancer cells), the photoreaction model and the tumor growth model. (i) Several papers have been reported for the application of system identification techniques to pharmacokinetics modeling problems. But two issues were ignored in these previous works: presence of timing noise and identification from longitudinal data. In [27], 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. Statistical inference from longitudinal data based on mixed effects models can be performed by the *Monolix* software (http://www.monolix.org) developed the Monolix group chaired by Marc Lavielle and France Mentré, and supported by Inria. In the recent past, we have used this tool for tumor growth modeling. (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 [39]. The main open problem is to access to data. We currently build on an experimental platform which aims at overcoming this technical issue. Moreover, an identifiability study coupled to a global sensitivity analysis of the photoreaction model are currently in progress. (iii) Tumor growth is generally described by population dynamics models or by cell cycle models. Faced with this wide variety of descriptions, one of the main open problems is to identify the suitable model structure. As mentioned above, we currently investigate alternative representations based on branching processes and Markov chains, with a model selection procedure in mind.

A few words should be said about the existing literature on statistical inference for diffusion or related processes, a topic which will be at the heart of three of our projects (namely photodynamic and bacteriophage therapies, as well as fluctuations within molecules). The monograph [45] is a good reference on the basic estimation techniques for diffusion processes. The problem of estimating diffusions observed at discrete times, of crucial importance for applications, has been addressed mainly since the mid 90s. The maximum likelihood techniques, which are also classical for parameter estimation, are well represented by the contributions [40].

Some attention has been paid recently to the estimation of the coefficients of fractional or multifractional Brownian motion according to a set of observations. Let us quote for instance the nice surveys [26], [35]. On the other hand, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. A good reference on the question is [61], dealing with some very particular families of equations, which do not cover the cases of interest for us.

4. Application Domains

4.1. Data analysis and local regression

Our expertise in data analysis and advanced statistics methods has given raise no a wide number of interdisciplinary collaborations. Among those, here are the most challenging at a scientific level:

(i) Health inequalities: We have recently developed a statistical procedure in order to create a neighborhood socioeconomic index and investigate its influence on health inequalities. The study setting is composed with 3 major French metropolitan areas (Lille, Lyon and Marseille), and we collaborate for this project with a medical team at EHESP (Ecole des Hautes Etudes en Santé Publique) lead by D. Zmirou (see (Lalloue & al, 2012) for further details).

(ii) Fetal pathology: 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 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.

(iii) Cohorts analysis: 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. We have thus initiated a common project together with the Inria team Orpailleur (particularly with Marie-Dominique Desvignes and Malika Smail) in this direction. The goal of this collaboration 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. 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.

4.2. Estimation for complex and biological systems

Our main application for this line of investigation is the photodynamic therapy developed by T. Bastogne. We shall also focus on bacteriophage therapies and subdiffusion within molecules.

(i) Photodynamic therapy. One of the main application we have in mind for our identification problems is to model photodynamic therapy. This promising cancer treatment involves selective uptake and retention of a photosensitive drug in a tumor, followed by irradiation with light at an appropriate wavelength. Photosensitizers are photoactive compounds such as for instance porphyrins and chlorins. The activated photosensitizer is thought to produce singlet oxygen at high doses and thereby to initiate apoptotic and necrotic death of tumor. Due to the lack of response reproducibility, the complexity of interactions between physical, chemical and biological aspects and the high cost of experiments, there is a real demand in good mathematical and physical models which might help to better control and understand PDT responses. We are particularly concerned with modeling the drug uptake into cancer cells, the photoreactions induced by light exposition and tumor growth kinetics.

(ii) Bacteriophage systems. A collaboration between our team, the Mathematics and the Genetics and Microbiology Departments at the *Universitat Autònoma de Barcelona* (UAB) is being set up, focusing on probabilistic aspects of bacteriophage therapies for animal diseases like hemorrhagic septicemia in cattle or atrophic rhinitis in swine. This kind of therapy consists in inoculating a (benign) virus to animals in order to kill the bacteria known to be responsible of the disease. It was in use in the Soviet Union until the 80s, and is now re-emerging, still at an experimental level, due to the progressive slowdown in antibiotic efficiency.

Within this context, our analysis of a noisy predator-prey competition modeling the treatment helps to calibrate and to understand better the behavior of the system in terms of fluctuations around an equilibrium. Note that our preliminary contacts with the Genetics and Microbiology Departments at UAB also open the way to a particle model in order to represent the couple bacteria/virus living on a surface.

5. Software and Platforms

5.1. Light diffusion into tissues

We are currently considering the possibility to implement our Matlab algorithms concerning light diffusion into tissues into the Matlab toolbox *Contsid*, developed by the System Identification team of the CRAN (http://www.iris.cran.uhp-nancy.fr/contsid/).

5.2. Online data analysis

An R package performing most of the methods of factorial analysis in an online way has been developed by R. Bar and J-M. Monnez. Starting from a simulated data flow, the main goal of the program is to perform online factorial analyses (Principal Component Analyses, Canonical Correlation Analysis, Canonical Discriminant Analysis, Correspondence Analysis). Data are supposed to be independent and identically distributed observations of a random vector (whose distribution is a priori unknown). Defining stochastic approximation processes, the procedure is adaptative in the sense that the results of the analyses are updated recursively each time that a new data is taken into account.

From a theoretical point of view, the i.i.d case has been recently extended to the case of an expectation and/or covariance matrix of the random vector varying with time. We plan to include these improvements into our software.

5.3. Socio-economic index

A R package called SesIndexCreatoR has been written by B. Lalloué and J-M. Monnez in order to implement our socio-economic index for health inequalities. The version 1.0 of this package is currently freely available on the website of the Equit'Area project: http://www.equitarea.org/documents/packages_1.0-0/. It contains the functions needed to run the procedure (either integrally or partially) and obtain the corresponding SES index. The user may also create categories of this index with different methods (hierarchical clustering with or without k-nearest neighbors, quantiles, or intervals) and generate automatic reports of the results. Visualization and plotting functions are provided in the package.

6. New Results

6.1. Modern methods of data analysis

Participants: R. Bar, S. Ferrigno, B. Lalloué, J-M. Monnez, A. Muller-Gueudin, S. Tindel

6.1.1. Help to medical decision and telemedecine in the monitoring of heart failure

We describe here a project started in 2013, for which we expect some concrete output in 2014. This project fits in the general framework of telemedecine 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 his rehospitalization.

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 rehospitalization 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.

We will use to carry out this study 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 treated applications.

6.1.2. Online factorial data analysis methods

Nowadays data analysts are often faced with the problem of dealing with a rapid and infinite flow of data. Examples include web, telecommunications, process control or financial data. We made first the assumption that the data are generated at random according to a stationary distribution, but in many cases this assumption does not hold true. We developed in [13] the online adaptation of principal component analysis and other dimension reduction statistical algorithms by using stochastic approximation. An R package was developed by Romain Bar.

6.1.3. Data analysis techniques and Bayesian models applied to the context of social inequalities and environmental exposures

The aim of [10] is to improve the knowledge about and apply data mining techniques and some Bayesian model in the field of social and environmental health inequalities. The health event considered is infant mortality. We try to explain its risk with socio-economic data retrieved from the national census and environmental exposures such as air pollution, noise, proximity to traffic, green spaces and industries. The data mining part details the development of a procedure of creation of multi-dimensional socio-economic indices and of an R package that implements it.

6.1.4. A simultaneous stepwise variable selection and clustering algorithm to discriminate a class variable with numerous levels

In supervised learning the number of levels of a categorical variable to explain can be high. When some of its levels are of low frequency, clustering them in order to reduce the number of classes can be useful to perform relevant discriminant analyses. On the other hand selecting relevant predictors is a crucial step to build robust and efficient classification rules, especially when too many variables are available in regard to the overall sample size. We are currently carrying out an extension of an algorithm we had devised to solve both these problems using an alternate minimization of Wilks' Lambda. We show through simulations the interest of adding Akaike Information Criterion as another optimality criterion. We also moved forward to stepwise selection and applied this new version of our algorithm to real allergology datasets.

6.1.5. Local polynomial regression. Application to the estimation of the fetal growth.

This topic is an ongoing collaboration with M. Maumy-Bertrand, for which we expect a publication in 2014. We have established exact rate of strong uniform consistency for the local linear estimator of the conditional distribution function. We want to extend our results to obtain exact rates of strong uniform consistency for the local linear estimator of other conditional quantities: the conditional mean $\mathbb{E}(Y|X)$, and the conditional quantiles $q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\}$, for $\alpha \in (0,1)$.

Another crucial problem with the non parametric regression methods is the choice of the bandwidth parameter h. It is common in practice to choose $h > 0$ so 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 $h_n = C(X_1, ..., X_n)n^{-1/5}$, where n is the sample size, and $X_1, ..., X_n$ are the n independent copies of the random variable X. This bandwidth is called a *data-driven bandwidth* to enhance its dependence to the data. Our current project in this direction consists in establishing the 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 $h_n = C(X_1, ..., X_n)n^{-1/5}$.

Turning to applications, note that we have a contact with Professor Bernard Foliguet at the Maternité Régionale de Nancy. We will continue to collaborate with him, to estimate growing curves of the fetal weight, and other fetal quantities thanks to the techniques mentioned above.

6.1.6. Cohort analysis

In an ongoing work with the INSERM team of P. Guéant, we aim at describing the complex interactions between genetic, phenotypic and biologic variables that are available in medical cohorts, in different contexts (cognitive decline; inflammatory intestinal diseases; liver cancer).

A firt step in our analysis, which should be finished in 2014, consists in giving an overview of 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:

1. with exploratory methods, or factorial models,
2. with regression models to predict qualitative variable by the use of qualitative or quantitative factors.

In the sequel, we will test non association or independence between the variables. The objective is to develop new methods, adapted to the studied cohorts (matching cases/controls, high number of individuals, high number of explicative variables, missing data problem). The particularity of our work is to combine statistical and symbolical methods.

After having identified and choice the relevant variables, we will have to give a model for classifying the data. The proposed models will allow us to identify subgroups of invidious, with common genetic, biologic and phenotypic characteristics.

6.1.7. Local polynomial estimation and goodness-of-fit tests

We describe here an ongoing work with Marie-José Martinez, assistant professor at the IUT of Grenoble and member of the Inria MISTIS team. A related publication should be finished at the end of 2014. 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). It is an extension of the linear model, in the sense that its satisfies a relation of the form $Y = g(X) + \epsilon$ where:

- The stochastic component ϵ follows other distributions than the Gaussian.
- The function g can be non linear.

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.

6.1.8. Model selection for SVM

Support vector machines provide a very powerful method of data classification, for which model selection is one of the key issues. For a support vector machine, model selection consists in selecting the kernel function, the values of its parameters, and the amount of regularization. To set the value of the regularization parameter, one can minimize an appropriate objective function over the regularization path. A priori, this requires the availability of two elements: the objective function and an algorithm computing the regularization path at a reduced cost. The literature provides us with several upper bounds and estimates for the leave-one-out cross-validation error of the ℓ_2-SVM. However, no algorithm was available so far for fitting the entire regularization path of this machine. In our contribution [3], we introduce the first algorithm of this kind. It is involved in the specification of new methods to tune the corresponding penalization coefficient, whose objective function is a leave-one-out error bound or estimate. From a computational point of view, these methods appear especially

appropriate when the Gram matrix is of low rank. A comparative study involving state-of-the-art alternatives provides us with an empirical confirmation of this advantage.

6.2. Estimation for complex and biological systems

Participants: T. Bastogne, C. Lacaux, S. Mézières, S. Tindel, P. Vallois

6.2.1. Tumor growth modeling

This project is an extension of our article [15], which will be written in 2014. 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 are currently working on a model in which we suppose that the size V_t at time t of the tumor is a diffusion process of the type :

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

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 (1). This diffusion process lives between two frontiers : 0 and κ. In this stochastic setting, the role of κ 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).

6.2.2. Local score associated with long biological sequences

Statistical properties of the distribution of the local score is largely used by molecular biologists to extract important features in biological sequences and in particular to determine the most significant one among a collection of biological sequences. The probabilistic model which is commonly used is the following. Associated with a sequence $(\epsilon_i)_{i\geq 1}$ of independent, centered and reduced random variables, consider $S_n = \epsilon_1 + \cdots + \epsilon_n$ and

$$\underline{S}_n = \min_{0\leq i\leq n} S_i, \quad U_n = S_n - \underline{S}_n = S_n - \min_{i\leq n} S_i \quad n \geq 0.$$

In biological sequence analysis, (ϵ_i) can for example correspond to the physical or chemical properties of the i-th amino acid or nucleotid of the sequence ; it can also reflect the similarity between components of two sequences. The local score $\overline{U}_n$ is the supremum of (U_n) up to time n. Molecular biologists are interested by this supremum as it highlights the best part of the studied sequence, the eventual segment of DNA transmitted by a common ancestor for sequence comparison or the best hydrophobic segment of a protein that would thus naturally move in a transmembrane place. It is clear that the trajectory of (U_n) can be decomposed in a succession of 0 and excursions above 0. These excursions have an important biological interpretation and in particular the highest one corresponds to the best segment due to the physico chemical property or similarity scores that have been chosen. Note that the local score $\overline{U}_n$ can be viewed as the maximum of the heights of all the excursions up to time n. In the article [22], we are interested in complete excursions up to a fixed time.

This leads us to introduce the maximum U_n^* of the heights of all the excursions up to time n. The second variable which will play an important role is θ_n^* the time necessary to reach its maximal height U_n^*.
We believe that the knowledge of the joint distribution of the pair (U_n^*, θ_n^*) would permit to get more efficient statistical tests than the ones only based on the local score. This point should be developed in a forthcoming paper.

However, it seems difficult to determine explicitly the law of $(U_n^*,\ \theta_n^*)$ for a fixed n. This difficulty can be overcome considering biological sequences which have a large number of bases and approximating the initial random walk (S_n) by a Brownian motion (B_t) started at 0. Using the functional theorem of convergence of Donsker, the process (U_k) can be compared to

$$\widehat{U}(t) := B(t) - \inf_{0\leq s\leq t} B(s), \quad t \geq 0. \tag{2}$$

This leads us to consider:

1. the local score $\overline{U}(t)$ which is the maximum of the heights of all the excursions of $U(s)$ up to time t,
2. the maximum $U^*(t)$ of the heights of all the complete excursions up to time t,
3. the time $\theta^*(t)$ taken by $U(s)$ starting from the beginning of the largest excursion to hit the maximal level $U^*(t)$.

The approximation of (U_n) by $(\widehat{U}_t)$ permits to prove that the asymptotic distribution of $\left(\frac{U_n^*}{\sqrt{n}},\ \frac{\theta_n^*}{n}\right)$ as $n \to \infty$ is the one of $(U^*(1), \theta^*(1))$. Consequently, our initial problem in the discrete setting reduces to determine the joint law of $(U^*(t), \theta^*(t))$, where $t > 0$ is given. We determine in [22] the distribution and the density functions of $(U^*(t), \theta^*(t))$.

6.2.3. Bacteriophage therapies

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 applying the treatment to 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.

At a mathematical level, 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) is as follows:

$$\begin{cases} \mathrm{d}S_t &= [\alpha - k\,Q_t]\,S_t \mathrm{d}t + \varepsilon\,S_t\,\mathrm{d}W_t^1 \\ \mathrm{d}Q_t &= \left[d - mQ_t - k\,Q_t\,S_t + k\,b\,e^{-\mu\zeta}\,Q_{t-\zeta}\,S_{t-\zeta}\right]\mathrm{d}t + \varepsilon\,Q_t\,\mathrm{d}W_t^2, \end{cases} \tag{3}$$

where α is the reproducing rate of the bacteria and k is the adsorption rate. In equation (3), 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, ζ 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 μ is the bacteria's death rate). A given initial condition (S_0, Q_0) is also specified, and the term $\varepsilon\,\mathrm{d}W_t$ 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 ζ (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 [1]) have been the following:

- Quantification of the exponential convergence to a bacteria-free equilibrium of equation (3) 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.

6.2.4. Light transport in tissues with probabilistic methods

Photodynamic therapy (PDT) is a type of phototherapy used for treating several diseases such as acne, bacterial infection, viruses and some cancers. The aim of this treatment is to kill pathological cells with a photosensitive drug that is absorbed by the target cells and that is then activated by light. For appropriate wavelength and power, the light beam makes the photosensitizer produce singlet oxygen at high doses and induces the apoptosis and necrosis of the malignant cells. Our project focuses on an innovative application: the interstitial PDT for the treatment of high-grade brain tumors. This strategy requires the installation of optical fibers to deliver the light directly into the tumor tissue to be treated, while nanoparticles are used to carry the photosensitizer into the cancer cells. 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. In [25], in the framework of homogeneous biological tissues, we propose an alternative MC method to Wang's algorithm. Then we also propose a variance reduction method. Interestingly enough, our formulation also allows us to design quite easily a Markov chain Monte Carlo (MCMC) method based on Metropolis-Hastings algorithm and to handle the inverse problem (of crucial importance for practitioners), consisting in estimating the optical coefficients of the tissue according to a series of measurements. We have compared the proposed MC and MCMC method and Wang's algorithm: we see that our MC method is much more consistent. However, MCMC methods induce quick mutations, which paves the way to very promising algorithms in the inhomogenous case. To handle the inverse problem, we derive a probabilistic representation of the variation of the fluence with respect to the absorption and scattering coefficients. This leads us to the implementation of a Levenberg-Marquardt type algorithm that gives an approximate solution to the inverse problem.

6.2.5. System identification of gap junctional intercellular communication channels of two cancer cell lines.

The main challenge addressed in this work [12], [14] was to assess the relevance of a proposed model-based approach to detect differences between gap junctional intercellular communication channels of two cancer cell lines. To that aim, KB and FaDu, two human head and neck carcinoma cell lines, were used. The former expresses connexin proteins (positive line) while the latter does not (negative line). Moreover, each cell line was tested on spheroid (3-D) and monolayer (2-D) slices and *in vitro* assays were repeated six times. Continuous-time system identification algorithms of the Matlab System Identification and CONTSID toolboxes are tested and applied to a set of *in vitro* data. Results firstly show an acceptable fit of the biological responses but they mainly emphasize the possibility to use several model parameters as statistics to discriminate different cancer cell lines. So, this study exemplifies the potential contribution of dynamic system identification methods and tools to the discovery of new diagnostic biomarkers in systems biology.

6.2.6. Photodynamic therapy modeling and control.

We have also carried on the development of methodological and technological innovations for the realtime control of the therapeutic efficiency in photodynamic therapy (Tylcz:2013). One part of the innovation has been protected by a patent submitted in 2012 (No.1261339, INPI) and extended in 2013. A demonstration platform is currently in development.

6.2.7. Bio-inspired system reliability method.

Based on previously developed works (Keinj, 2011, 2012), we have also proposed in [15] an extension of the target theory in biology applied to system reliability. In this development, we consider rough products produced by a factory. Each product coming from the plant has m vital elements and some elements can be damaged. To obtain a perfect product (i.e. all the constitutive m elements are safe) all the damaged elements are repaired and a test phase follows. The result of this two-steps procedure is random. We suppose that the number Z_k of non-damaged elements is a Markov chain valued in the set $\{0, 1, \cdots, m\}$, where k is the number of applied repairing-test phases. We have a qualitative result which says that if the repair phase is efficient then $P(Z_k = m)$ is close to 1. As for production of a large number n of products, the former result allows us to give conditions under which either the n elements or a fraction of these n elements are (is) safe after the application of k previous maintenance phases.

6.2.8. Dynamical Global Sensitivity Analysis as an Early Warning for System's Critical Transition.

In biology, systems with bifurcations may experience abrupt irreversible and often unwanted shifts in their performance, called critical transitions. For many systems like climate, economy, ecosystems it is highly desirable to identify indicators serving as early warnings of such regime shifts. Several statistical measures were recently proposed as early warnings of critical transitions including increased variance, autocorrelation and skewness of experimental or model-generated data. The lack of automatized tool for model-based prediction of critical transitions led to designing DyGloSA, a Matlab toolbox for dynamical global parameter sensitivity analysis (GPSA) of ordinary differential equations models. One part of our research activity in 2013 was focused on the implementation of a global sensitivity analysis method developed in (Dobre, 2011, 2012) into DyGloSA for dynamical global parameter sensitivity analysis (GPSA) of ordinary differential equations models. This work has been carried out in the context of a collaboration with the University of Luxembourg and more precisely the Thomas Sauter's team. We have shown in [2] that tools developed in this toolbox are efficient to analyze several models with bifurcations and predict the time periods when systems can still avoid going to a critical transition by manipulating certain parameter values, which is not detectable with the existing SA techniques.

6.3. Inference for gaussian systems

Participants: C. Lacaux, S. Tindel

6.3.1. Inference for dynamical systems driven by Gaussian noises.

As mentioned at the *Scientific Foundations* Section, 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:

$$X_t^\theta = a + \theta \int_0^t b(X_u)\, du + B_t,$$

where θ 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 (see the *Application Domains* Section) require the analysis of the following $\mathbb{R}^n$-valued equation:

$$X_t^\theta = a + \int_0^t b(\theta; X_u)\, du + \int_0^t \sigma(\theta; X_u)\, dB_t, \tag{4}$$

where θ enters non linearly in the coefficient, where σ 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 in 2013:

- A better understanding of the underlying rough path structure for equation (4), carried out in [6]. This step allows a proper definition of stochastic integrals with respect to fractional Brownian motion in a wide range of contexts.
- Extension of pathwise stochastic integration to processes indexed by the plane in [19], which helps to the definition of noisy systems such as partial differential equations.
- Gaussian type bounds for equations driven by a fractional Brownian motion, obtained in [18], [7]. This is an important preliminary step for likelihood estimates for stochastic processes. Also notice the interesting central limit theorems exhibited in [24], in a context which is similar to our equation of interest.
- Numerical aspects of a maximum likelihood type procedure for an equation of the form (4), expressed in terms of Malliavin calculus tools (see [4]).

6.3.2. 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 [5], we focus on the LAN property for the model where we observe a sample of n observations $\mathbf{X_n} = (X_1, ..., X_n)$ of a Gaussian stationary sequence. The sequence $(X_n)_{n\in\mathbb{N}}$, whose spectral density f_θ is indexed by a parameter θ, can admit antipersistence, 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_θ 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.

6.3.3. 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:

$$\forall c > 0,\ \left(X(c^E x)\right)_{x\in\mathbb{R}^d} \overset{(d)}{=} c\,(X(x))_{x\in\mathbb{R}^d}\,, \tag{5}$$

where $c^E := \exp(E \ln(c))$. In particular, the Hölder regularity properties of operator scaling Gaussian or stable harmonizable random fields have been expressed in terms of the matrix E. The method they used can be applied to study the modulus of continuity of many stable or Gaussian random fields. As example in 2011, with P. Scheffler, they have followed it to study multi-operator harmonizable stable random fields, which satisfy a local version of the operator scaling property and enjoy a regularity which may vary along the trajectories. In [20], it has been developed in the more general framework of conditionally sub-Gaussian random series. This allows to also study for example some multistable random fields, which have been introduced in (Falconer & al, 2009); for such a field X, the marginal $X(x)$ is a stable random variable whose index of stability can depend on x. In this paper, some conditions have been proposed to establish the uniform convergence of the series (on an eventually random ball), an upper bound for the modulus of continuity of its limit, an uniform control of the partial series ones and an explicit rate of convergence. Focusing on LePage random series, upper bounds of the modulus of continuity of some harmonizable stable or multistable random fields are provided. In the conference paper [11], [20] has then been applied to study the class of linear multifractional multistable motions. In particular, the upper bound obtained for the modulus of linear multifractional stable motion is the sharpest available.

We are also interested in self-similar processes indexed by manifolds in [8]. 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.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

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.

8. Partnerships and Cooperations

8.1. National Initiatives

- *Optique-PDT* (2012-2014), mOdélisation et oPTimisation de l'Irradiance dans les tissus biologiQUEs hétérogènes traités par Thérapie PhotoDynamique interstitielle, Funding organism: *PEPS CNRS-INSERM-Inria*, Leader: M. Thomassin (CRAN, U. Lorraine).
- *Nano-Xrays* (2011-2014), Nanoparticles-based X ray-induced photodynamic therapy in glioblastoma multiforme, Funding organism: *Institut National du Cancer (INCa)*, Leader: M. Barberi-Heyob (CRAN, U. Lorraine).
- *PDTX* (2010-2013), Active Nanoplatforms for Photodynamic Therapy, Funding organism: *French National Agency for Research (ANR)*, Leader: M. Verelst (U. Paul Sabatier, Toulouse).
- *MASTERIE* (2010-2013), Malliavin Stein Random Irregular Equation, Funding organism: *French National Agency for Research (ANR)*, Leader: F. Russo (ENSTA, Paris).

8.2. European Initiatives

8.2.1. FP7 Projects

- *Target-PDT* (2009-2013), Photodynamic Therapy using photosensitizer-doped targeted organic nanoparticles, Funding organism: FP7 ERA-NET EuroNanoMed: **European Innovative RTD Projects Proposals in Nanomedicine**, Leader: P. Boisseau (CEA LETI, Grenoble).

9. Dissemination

9.1. Scientific Animation

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

9.2. Teaching - Supervision - Juries

9.2.1. 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) has been proposed by Thierry Bastogne and started in Sep. 2013 at the Faculté des Sciences et Technologies.

9.2.2. Supervision

PhD: Jean Baptiste Tylcz, Identification et contrôle de systèmes biologiques. Application à la thérapie photodynamique, Université de Lorraine, 4 décembre 2013. Advisors: T. Bastogne and E. Bullinger (U. de Liège).

PhD : Romain Bar, Développement de méthodes d'analyse de données en ligne, Université de Lorraine, 29 novembre 2013. Advisor : J-M. Monnez.

PhD : Benoît Lalloué, Méthodes d'analyse de données et modèles bayesiens appliqués au contexte des inégalités socio-territoriales de santé et des expositions environnementales, Université de Lorraine, 6 décembre 2013. Advisors : J-M. Monnez and S. Deguen.

PhD: Rémi Bonidal, Sélection de modèle par chemin de régularisation pour les machines à vecteurs support à coût quadratique, Université de Lorraine, 19 juin 2013. Advisors: Y. Guermeur and S. Tindel.

9.2.3. Juries

- PhD : Ilham Ben Abbes, Développement d'un nouveau modèle dédié à la commande du métabolisme glucidique appliqué aux patients diabétiques de type 1, Supélec Rennes, 28 juin 2013. Referee: T. Bastogne.

- HDR: Landy Rabehasaina, Contribution à quelques problèmes de premier passage et de ruine multidimensionnels; Lien avec les réseaux de files fluides, Université de Franche Comté, 3 décembre 2013. Referee: P. Vallois.

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] X. BARDINA, D. BASCOMPTE, C. ROVIRA, S. TINDEL. *A stochastic model for bacteriophage therapies*, in "Math. Biosci.", 2013, vol. 241, n^{o} 1, pp. 99-108, http://hal.inria.fr/hal-00619307

[2] T. BAUMURATOVA, S. DOBRE, T. BASTOGNE, T. SAUTER. *Switch of sensitivity dynamics revealed with DyGloSA toolbox for dynamical global sensitivity analysis as an early warning for system's critical transition*, in "PLoS ONE", December 2013, vol. 8, n^{o} 12 [*DOI :* 10.1371/JOURNAL.PONE.0082973], http://hal.inria.fr/hal-00911807

[3] R. BONIDAL, S. TINDEL, Y. GUERMEUR. *Model Selection for the l2-SVM by Following the Regularization Path*, in "Transactions on Computational Collective Intelligence", 2013, vol. 13, pp. 83-112, http://hal.inria.fr/hal-00849720

[4] A. CHRONOPOULOU, S. TINDEL. *On inference for fractional differential equations*, in "Statistical Inference for Stochastic Processes", 2013, vol. 16, n^{o} 1, pp. 29-61, http://hal.inria.fr/hal-00587087

[5] S. COHEN, F. GAMBOA, C. LACAUX, J.-M. LOUBES. *LAN property for some fractional type Brownian motion*, in "ALEA : Latin American Journal of Probability and Mathematical Statistics", 2013, vol. 10, n^{o} 1, pp. 91-106, http://hal.inria.fr/hal-00638121

[6] Y. HU, M. JOLIS, S. TINDEL. *On Stratonovich and Skorohod stochastic calculus for Gaussian processes*, in "Annals of Probability", 2013, vol. 41, n^{o} 3, pp. 1656-1693, http://hal.inria.fr/hal-00556955

[7] Y. HU, S. TINDEL. *Smooth density for some nilpotent rough differential equations*, in "Journal of Theoretical Probability", 2013, vol. 26, pp. 722-749, Contains a Norris type lemma in the rough paths context. A similar result is to be posted soon by M. Hairer and N. Pillai, http://hal.inria.fr/hal-00584938

[8] J. ISTAS, C. LACAUX. *On locally self-similar fractional random fields indexed by a manifold*, in "Stochastics: An International Journal of Probability and Stochastic Processes", 2013, vol. 85, n^{o} 3, pp. 489-499 [*DOI :* 10.1080/17442508.2012.654609], http://hal.inria.fr/hal-00193203

[9] C. LACAUX, A. MULLER, R. RANTA, S. TINDEL. *Convergence and performance of the peeling wavelet denoising algorithm*, in "Metrika", 2014, Online first article [*DOI :* 10.1007/S00184-013-0451-Y], http://hal.inria.fr/hal-00903593

[10] B. LALLOUÉ, J.-M. MONNEZ, C. PADILLA, W. KIHAL, N. LE MEUR, D. ZMIROU-NAVIER, S. DEGUEN. *A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis*, in "International Journal for Equity in Health", March 2013, vol. 12, n^{o} 1, 21 p. [*DOI :* 10.1186/1475-9276-12-21], http://hal.inria.fr/hal-00854028

International Conferences with Proceedings

[11] H. BIERMÉ, C. LACAUX. *Linear multifractional multistable motion: LePage series representation and modulus of continuity*, in "XIème Colloque Franco-Roumain de Mathématiques Appliquées", Bucarest, Romania, 2013, vol. 1/2013, pp. 345-360, http://hal.inria.fr/hal-00855092

[12] J.-B. TYLCZ, M. ABBACI, T. BASTOGNE, W. BLONDEL, D. DUMAS, M. BARBERI-HEYOB. *System identification of the fluorescence recovery after photobleaching in gap junctional intracellular communications*, in "52nd IEEE Conference on Decision and Control, CDC 2013", Florence, Italy, December 2013, http://hal.inria.fr/hal-00877361

National Conferences with Proceedings

[13] R. BAR, J.-M. MONNEZ. *Analyse en composantes principales partielle de données séquentielles d'espérance et de matrice de covariance variables dans le temps*, in "45èmes Journées de Statistiques - 2013", Toulouse, France, May 2013, http://hal.inria.fr/hal-00841181

Conferences without Proceedings

[14] J.-B. TYLCZ, M. ABBACI, T. BASTOGNE, W. BLONDEL, D. DUMAS, M. BARBERI-HEYOB. *A system-engineering model to analyze gap-FRAP in multicellular models.*, in "Interdisciplinary Symposium on Signals and Systems for Medical Applications, ISSSMA 2013", Paris, France, June 2013, http://hal.inria.fr/hal-00920143

Scientific Books (or Scientific Book chapters)

[15] T. BASTOGNE, P. VALLOIS. *An extension of the target theory in biology applied to system reliability*, in "Models and Methods in Economics and Management Science: Essays in Honor of Charles S. Tapiero", F. E. OUARDIGHI, K. KOGAN (editors), International Series in Operations Research & Management Science, Springer, 2014, pp. 155-181, http://hal.inria.fr/hal-00911872

[16] A. DEYA, S. TINDEL. *Malliavin calculus for fractional heat equation*, in "Malliavin Calculus and stochastic analysis : a festschrift in honor of David Nualart", Springer Proceedings in Mathematics & Statistics, Springer, 2013, 583 p. , Dedicated to David Nualart on occasion of his 60th birthday, http://hal.inria.fr/hal-00618597

Research Reports

[17] F. BAUDOIN, E. NUALART, C. OUYANG, S. TINDEL. , *On probability laws of solutions to differential systems driven by a fractional Brownian motion.*, January 2014, http://hal.inria.fr/hal-00931118

[18] M. BESALÚ, A. KOHATSU-HIGA, S. TINDEL. , *Gaussian type lower bounds for the density of solutions of SDEs driven by fractional Brownian motions*, 2013, 1 figure, http://hal.inria.fr/hal-00875378

[19] K. CHOUK, S. TINDEL. , *Skorohod and Stratonovich integration in the plane*, September 2013, http://hal.inria.fr/hal-00866067

Other Publications

[20] H. BIERMÉ, C. LACAUX. , *Modulus of continuity of some conditionally sub-Gaussian fields, application to stable random fields*, 2013, http://hal.inria.fr/hal-00780684

[21] P. CAZEAUX, P. CHARTON, N. PHAM, L. VINCKENBOSCH, R. ZEINEDDINE. , *Semaine d'Etude Mathématiques et Entreprises 5 : Propriétés asymptotiques de processus à volatilité stochastique*, 2013, http://hal.inria.fr/hal-00833375

[22] C. CHABRIAC, A. LAGNOUX, S. MERCIER, P. VALLOIS. , *Element related to the largest complete excursion of a reflected BM stopped at a fixed time. Application to local score*, September 2013, http://hal.inria.fr/hal-00857402

[23] C. DESMONTS, R. BAR, M. HAMZA, I. CHOURABI, P.-A. FAYE. , *Semaine d'Etude Mathématiques et Entreprises 5 : Sélection de variables statistiquement représentatives pour la production électrique photovoltaïque*, 2013, http://hal.inria.fr/hal-00833410

[24] A. DEYA, D. NUALART, S. TINDEL. , *On L2 modulus of continuity of Brownian local times and Riesz potentials*, April 2013, http://hal.inria.fr/hal-00814171

[25] L. VINCKENBOSCH, S. TINDEL, M. THOMASSIN, C. LACAUX, O. TIPHAINE. , *Monte Carlo methods for light propagation in biological tissues*, 2013, arXiv:1311.5343, http://hal.inria.fr/hal-00907391

References in notes

[26] J. M. BARDET, G. LANG, G. OPPENHEIM, A. PHILIPPE, S. STOEV, M. TAQQU. *Semi-parametric estimation of the long-range dependence parameter: a survey*, in "Theory and applications of long-range dependence", Birkhauser Boston, 2003, pp. 557-577

[27] T. BASTOGNE, S. MÉZIÈRES-WANTZ, N. RAMDANI, P. VALLOIS, M. BARBERI-HEYOB. *Identification of pharmacokinetics models in the presence of timing noise*, in "Eur. J. Control", 2008, vol. 14, n^{o} 2, pp. 149–157, http://dx.doi.org/10.3166/ejc.14.149-157

[28] A. BENASSI, S. JAFFARD, D. ROUX. *Elliptic Gaussian random processes*, in "Rev. Mat. Iberoamericana", 1997, vol. 13, n^{o} 1, pp. 19–90

[29] J. O. BERGER. , *Statistical decision theory and Bayesian analysis*, Springer Series in Statistics, Springer-VerlagNew York, 1993, xvi+617 p. , Corrected reprint of the second (1985) edition

[30] H. BIERMÉ, C. LACAUX. *Hölder regularity for operator scaling stable random fields*, in "Stochastic Process. Appl.", 2009, vol. 119, n^{o} 7, pp. 2222–2248, http://dx.doi.org/10.1016/j.spa.2008.10.008

[31] H. BIERMÉ, C. LACAUX, H.-P. SCHEFFLER. *Multi-operator Scaling Random Fields*, in "Stochastic Processes and their Applications", 2011, vol. 121, n^{o} 11, pp. 2642-2677, MAP5 2011-01 [*DOI :* 10.1016/J.SPA.2011.07.002], http://hal.archives-ouvertes.fr/hal-00551707/en/

[32] H. BIERMÉ, C. LACAUX, Y. XIAO. *Hitting probabilities and the Hausdorff dimension of the inverse images of anisotropic Gaussian random fields*, in "Bull. Lond. Math. Soc.", 2009, vol. 41, n^{o} 2, pp. 253–273, http://dx.doi.org/10.1112/blms/bdn122

[33] B. BRUNET-IMBAULT, G. LEMINEUR, C. CHAPPARD, R. HARBA, CLAUDE-LAURENT. BENHAMOU. *A new anisotropy index on trabecular bone radiographic images using the fast Fourier transform*, in "BMC Med. Imaging", 2005, vol. 5, n^{o} 1, 4 p.

[34] M. CARLETTI. *Mean-square stability of a stochastic model for bacteriophage infection with time delays*, in "Math. Biosci.", 2007, vol. 210, n° 2, pp. 395–414, http://dx.doi.org/10.1016/j.mbs.2007.05.009

[35] J. F. COEURJOLLY. *Simulation and identification of the fractional brownian motion: a bibliographical and comparative study.*, in "Journal of Statistical Software", 2000, vol. 5, pp. 1–53

[36] S. COHEN, C. LACAUX, M. LEDOUX. *A general framework for simulation of fractional fields*, in "Stochastic Process. Appl.", 2008, vol. 118, n° 9, pp. 1489–1517, http://dx.doi.org/10.1016/j.spa.2007.09.008

[37] A. CRUDU, A. DEBUSSCHE, O. RADULESCU. *Hybrid stochastic simplifications for multiscale gene networks*, in "BMC Systems Biology", 2009, vol. 3, 89 p. , http://hal.inria.fr/inria-00431227/en/

[38] A. DEYA, S. TINDEL. *Rough Volterra equations. I. The algebraic integration setting*, in "Stoch. Dyn.", 2009, vol. 9, n° 3, pp. 437–477, http://dx.doi.org/10.1142/S0219493709002737

[39] S. DOBRE, T. BASTOGNE, M. BARBERI-HEYOB, A. RICHARD. *Practical identifiability of photophysical parameters in photodynamic therapy*, in "29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'2007", Lyon, France, August 2007, http://hal.archives-ouvertes.fr/hal-00167475/en/

[40] G. B. DURHAM, A. R. GALLANT. *Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes*, in "J. Bus. Econom. Statist.", 2002, vol. 20, n° 3, pp. 297–338, With comments and a reply by the authors, http://dx.doi.org/10.1198/073500102288618397

[41] S. FERRIGNO, G. DUCHARME. *Un test d'adéquation global pour la fonction de répartition conditionnelle*, in "C. R. Math. Acad. Sci. Paris", 2005, vol. 341, n° 5, pp. 313–316, http://dx.doi.org/10.1016/j.crma.2005.07.003

[42] S. FERRIGNO, G. DUCHARME. *Un choix de fenêtre optimal en estimation polynomiale locale de la fonction de répartition conditionnelle*, in "C. R. Math. Acad. Sci. Paris", 2008, vol. 346, n° 1-2, pp. 83–86, http://dx.doi.org/10.1016/j.crma.2007.11.026

[43] S. FERRIGNO, M. MAUMY-BERTRAND, A. MULLER-GUEUDIN. , *Uniform law of the logarithm for the conditional distribution function and application to certainty bands*, September 2011, http://hal.inria.fr/hal-00641027

[44] M. GUBINELLI, S. TINDEL. *Rough evolution equations*, in "Ann. Probab.", 2010, vol. 38, n° 1, pp. 1–75, http://dx.doi.org/10.1214/08-AOP437

[45] Y. A. KUTOYANTS. , *Statistical inference for ergodic diffusion processes*, Springer Series in Statistics, Springer-Verlag London Ltd.London, 2004, xiv+481 p.

[46] C. LACAUX. *Real Harmonizable Multifractional Lévy Motions*, in "Ann. Inst. Poincaré.", 2004, vol. 40, n° 3, pp. 259–277

[47] C. LACAUX. *Series representation and simulation of multifractional Lévy motions*, in "Adv. in Appl. Probab.", 2004, vol. 36, n° 1, pp. 171–197, http://dx.doi.org/10.1239/aap/1077134469

[48] C. LACAUX, J.-M. LOUBES. *Hurst exponent estimation of fractional Lévy motion*, in "ALEA Lat. Am. J. Probab. Math. Stat.", 2007, vol. 3, pp. 143–164

[49] C. LACAUX, R. MARTY. , *From invariance principles to a class of multifractional fields related to fractional sheets*, 2011, MAP5 2011-08, http://hal.archives-ouvertes.fr/hal-00592188/en/

[50] L. LEBART. *On the Benzecri's method for computing eigenvectors by stochastic approximation (the case of binary data)*, in "Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)", Vienna, Physica Verlag, 1974, pp. 202–211

[51] H. LIERO. *Testing homoscedasticity in nonparametric regression*, in "Journal of Nonparametric Statistics", 2003, vol. 15, n^{o} 1, pp. 31-51, http://www.tandfonline.com/doi/abs/10.1080/10485250306038

[52] T. LYONS, Z. QIAN. , *System control and rough paths*, Oxford mathematical monographs, Clarendon Press, 2002, http://books.google.com/books?id=H9fRQNIngZYC

[53] J. MACQUEEN. , *Some methods for classification and analysis of multivariate observations.*, 1967, Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967).

[54] J.-M. MONNEZ. *Approximation stochastique en analyse factorielle multiple*, in "Ann. I.S.U.P.", 2006, vol. 50, n^{o} 3, pp. 27–45

[55] J.-M. MONNEZ. *Stochastic approximation of the factors of a generalized canonical correlation analysis*, in "Statist. Probab. Lett.", 2008, vol. 78, n^{o} 14, pp. 2210–2216, http://dx.doi.org/10.1016/j.spl.2008.01.088

[56] J.-M. MONNEZ. *Convergence d'un processus d'approximation stochastique en analyse factorielle*, in "Publ. Inst. Statist. Univ. Paris", 1994, vol. 38, n^{o} 1, pp. 37–55

[57] D. MÁRQUEZ-CARRERAS, C. ROVIRA, S. TINDEL. *Asymptotic behavior of the magnetization for the perceptron model*, in "Ann. Inst. H. Poincaré Probab. Statist.", 2006, vol. 42, n^{o} 3, pp. 327–342, http://dx.doi.org/10.1016/j.anihpb.2005.04.005

[58] A. NEUENKIRCH, I. NOURDIN, A. RÖSSLER, S. TINDEL. *Trees and asymptotic expansions for fractional stochastic differential equations*, in "Ann. Inst. Henri Poincaré Probab. Stat.", 2009, vol. 45, n^{o} 1, pp. 157–174, http://dx.doi.org/10.1214/07-AIHP159

[59] F. RUSSO, P. VALLOIS. *Elements of stochastic calculus via regularization*, in "Séminaire de Probabilités XL", Berlin, Lecture Notes in Math., Springer, 2007, vol. 1899, pp. 147–185, http://dx.doi.org/10.1007/978-3-540-71189-6_7

[60] S. TINDEL. *Quenched large deviation principle for the overlap of a p-spins system*, in "J. Statist. Phys.", 2003, vol. 110, n^{o} 1-2, pp. 51–72, http://dx.doi.org/10.1023/A:1021062510565

[61] C. A. TUDOR, F. G. VIENS. *Statistical aspects of the fractional stochastic calculus*, in "Ann. Statist.", 2007, vol. 35, n^{o} 3, pp. 1183–1212, http://dx.doi.org/10.1214/009053606000001541

[62] É. WALTER, L. PRONZATO. , *Identification of parametric models*, Communications and Control Engineering Series, Springer-VerlagBerlin, 1997, xviii+413 p. , From experimental data, Translated from the 1994 French original and revised by the authors, with the help of John Norton