RESEARCH CENTRE

Nancy - Grand Est

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2020 ACTIVITY REPORT

Project-Team BIGS

Biology, genetics and statistics

IN COLLABORATION WITH: Institut Elie Cartan de Lorraine (IECL)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Contents

Project-Team BIGS

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

Keywords

Computer sciences and digital sciences

A3.1. – Data
A3.1.1. – Modeling, representation
A3.2. – Knowledge
A3.2.3. – Inference
A3.3. – Data and knowledge analysis
A3.3.1. – On-line analytical processing
A3.3.2. – Data mining
A3.3.3. – Big data analysis
A3.4.1. – Supervised learning
A3.4.2. – Unsupervised learning
A3.4.4. – Optimization and learning
A3.4.7. – Kernel methods
A6. – Modeling, simulation and control
A6.1. – Methods in mathematical modeling
A6.1.2. – Stochastic Modeling
A6.2. – Scientific computing, Numerical Analysis & Optimization
A6.2.3. – Probabilistic methods
A6.2.4. – Statistical methods
A6.4. – Automatic control
A6.4.2. – Stochastic control

Other research topics and application domains

B1. – Life sciences
B1.1. – Biology
B1.1.2. – Molecular and cellular biology
B1.1.10. – Systems and synthetic biology
B1.1.11. – Plant Biology
B2.2. – Physiology and diseases
B2.2.1. – Cardiovascular and respiratory diseases
B2.2.3. – Cancer
B2.3. – Epidemiology
B2.4. – Therapies
B5.5. – Materials

1 Team members, visitors, external collaborators

Research Scientists

- Nicolas Champagnat [Inria, Senior Researcher, from Dec 2020, HDR]
- Coralie Fritsch [Inria, Researcher, from Dec 2020]
- Ulysse Herbach [Inria, Researcher]
- Bruno Scherrer [Inria, Researcher, HDR]

Faculty Members

- Anne Gégout Petit [Team leader, Univ de Lorraine, Professor, HDR]
- Thierry Bastogne [Univ de Lorraine, Professor, HDR]
- Sandie Ferrigno [Univ de Lorraine, Associate Professor]
- Sophie Mezieres [Univ de Lorraine, Associate Professor]
- Jean-Marie Monnez [Univ de Lorraine, Emeritus, HDR]
- Aurélie Muller-Gueudin [Univ de Lorraine, Associate Professor]
- Samy Tindel [Univ de Lorraine, Professor, HDR]
- Pierre Vallois [Univ de Lorraine, Professor, HDR]
- Denis Villemonais [Univ de Lorraine, Associate Professor, from Sep 2020, HDR]

Post-Doctoral Fellows

- Emma Horton [Université de Bath - Angleterre, until Nov 2020]
- William Ocafrain [Inria, from Dec 2020]

PhD Students

- Vincent Hass [Inria, from Dec 2020]
- Clémence Karmann [Univ de Lorraine, until Aug 2020]
- Rodolphe Loubaton [Univ de Lorraine, from Dec 2020]
- Nassim Sahki [Inria]
- Nino Vieillard [Google, CIFRE]
- Nicolás Zalduendo Vidal [Inria, from Dec 2020]

Technical Staff

- Benoît Lalloué [Centre hospitalier universitaire de Nancy, Engineer]
- Nicolas Thorr [Inria, Engineer, from Dec 2020]

Interns and Apprentices

- Salma Aziz [Inria, from Aug 2020 until Sep 2020]
- Alfred Kamdem Tezanlekeu [Inria, from Sep 2020]

Administrative Assistant

- Celine Cordier [Inria]

External Collaborators

- Céline Lacaux [Univ d'Avignon et des pays du Vaucluse, HDR]
- Lionel Lenôtre [Univ de Haute Alsace]

2 Overall objectives

BIGS is a joint team of Inria, CNRS and Université Lorraine, via the Institut Élie Cartan, UMR 7502 CNRS-UL laboratory in mathematics, of which Inria is a strong partner. One member of BIGS, T. Bastogne, comes from the Research Center of Automatic Control of Nancy (CRAN), with which BIGS has strong relations in the domain "Health-Biology-Signal". Our research is mainly focused on stochastic modeling and statistics but also aiming at a better understanding of biological systems. BIGS involves applied mathematicians whose research interests mainly concern probability and statistics. More precisely, our attention is directed on (1) stochastic modeling, (2) estimation and control for stochastic processes, (3) algorithms and estimation for graph data and (4) regression and machine learning. The main objective of BIGS is to exploit these skills in applied mathematics to provide a better understanding of issues arising in life sciences, with a special focus on (1) tumor growth, (2) photodynamic therapy, (3) population studies of genomic data and of micro-organisms genomics, (4) epidemiology and e-health.

3 Research program

3.1 Introduction

We give here the main lines of our research that belongs to the domains of probability and statistics. For clarity, we made the choice to structure them in four items. Although this choice was not arbitrary, the outlines between these items are sometimes fuzzy because each of them deals with modeling and inference and they are all interconnected.

3.2 Stochastic modeling

Our aim is to propose relevant stochastic frameworks for the modeling and the understanding of biological systems. The stochastic processes are particularly suitable for this purpose. Among them, Markov chains give a first framework for the modeling of population of cells [72, 50]. Piecewise deterministic processes are non diffusion processes also frequently used in the biological context [40, 49, 42]. Among Markov model, we developed strong expertise about processes derived from Brownian motion and Stochastic Differential Equations [66, 48]. For instance, knowledge about Brownian or random walk excursions [73, 64] helps to analyse genetic sequences and to develop inference about it. However, 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 of fractional fields. The basic aspects of these differential equations are now well understood, mainly thanks to the so-called rough paths tools [56], but also invoking the Russo-Vallois integration techniques [65]. The specific issue of Volterra equations driven by fractional Brownian motion, which is central for the subdiffusion within proteins problem, is addressed in [41]. Many generalizations (Gaussian or not) of this model have been recently proposed for some Gaussian locally self-similar fields, or for some non-Gaussian models [53], or for anisotropic models [37].

3.3 Estimation and control for stochastic processes

We develop inference about stochastic processes that we use for modeling. Control of stochastic processes is also a way to optimise administration (dose, frequency) of therapy.

There are many estimation techniques for diffusion processes or coefficients of fractional or multi-fractional Brownian motion according to a set of observations [52, 33, 39]. But, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. Our team has a good expertise about inference of the jump rate and the kernel of Piecewise Deterministic Markov Processes (PDMP) [32, 29, 31, 30]. However, there are many directions to go further into. For instance, previous works made the assumption of a complete observation of jumps and mode, that is unrealistic in practice. We tackle the problem of inference of "Hidden PDMP". As an example, in pharmacokinetics modeling inference, we want to take into account for presence of timing noise and identification from longitudinal data. We have expertise on this subjects [34], and we also used mixed models to estimate tumor growth [35].

We consider the control of stochastic processes within the framework of Markov Decision Processes [63] and their generalization known as multi-player stochastic games, with a particular focus on infinite-horizon problems. In this context, we are interested in the complexity analysis of standard algorithms, as well as the proposition and analysis of numerical approximate schemes for large problems in the spirit of [36]. Regarding complexity, a central topic of research is the analysis of the Policy Iteration algorithm, which has made significant progress in the last years [75, 62, 47, 68], but is still not fully understood. For large problems, we have a long experience of sensitivity analysis of approximate dynamic programming algorithms for Markov Decision Processes [69, 71, 67, 55, 70], and we currently investigate whether/how similar ideas may be adapted to multi-player stochastic games.

3.4 Algorithms and estimation for graph data

A graph data structure consists of a set of nodes, together with a set of pairs of these nodes called edges. This type of data is frequently used in biology because they provide a mathematical representation of many concepts such as biological structures and networks of relationships in a population. Some attention has recently been focused in the group on modeling and inference for graph data.

Network inference is the process of making inference about the link between two variables taking into account the information about other variables. [74] gives a very good introduction and many references about network inference and mining. Many methods are available to infer and test edges in Gaussian graphical models [74, 57, 45, 46]. However, when dealing with abundance data, because inflated zero data, we are far from gaussian assumption and we want to develop inference in this case.

Among graphs, trees play a special role because they offer a good model for many biological concepts, from RNA to phylogenetic trees through plant structures. Our research deals with several aspects of tree data. In particular, we work on statistical inference for this type of data under a given stochastic model. We also work on lossy compression of trees via directed acyclic graphs. These methods enable us to compute distances between tree data faster than from the original structures and with a high accuracy.

3.5 Regression and machine learning

Regression models and machine learning aim at inferring statistical links between a variable of interest and covariates. In biological study, it is always important to develop adapted learning methods both in the context of *standard* data and also for data of high dimension (with sometimes few observations) and very massive or online data.

Many methods are available to estimate conditional quantiles and test dependencies [61, 51]. Among them we have developed nonparametric estimation by local analysis via kernel methods [43, 44] and we want to study properties of this estimator in order to derive a measure of risk like confidence band and test. We study also many other regression models like survival analysis, spatio temporal models with covariates. Among the multiple regression models, we want to develop omnibus tests that examine several assumptions together.

Concerning the analysis of high dimensional data, our view on the topic relies on the *French data analysis school*, specifically on Factorial Analysis tools. In this context, stochastic approximation is an essential tool [54], which allows one to approximate eigenvectors in a stepwise manner [59, 58, 60].

BIGS aims at performing accurate classification or clustering by taking advantage of the possibility of updating the information "online" using stochastic approximation algorithms [38]. We focus on several incremental procedures for regression and data analysis like linear and logistic regressions and PCA (Principal Component Analysis).

We also focus on the biological context of high-throughput bioassays in which several hundreds or thousands of biological signals are measured for a posterior analysis. We have to account for the inter-individual variability within the modeling procedure. We aim at developing a new solution based on an ARX (Auto Regressive model with eXternal inputs) model structure using the EM (Expectation-Maximisation) algorithm for the estimation of the model parameters.

4 Application domains

4.1 Tumor growth-oncology

On this topic, we want to propose branching processes to model appearance of mutations in tumor through new collaborations with clinicians. The observed process is the "circulating DNA" (ctDNA). The final purpose is to use ctDNA as a early biomarker of the resistance to an immunotherapy treatment. It is the aim of the ITMO project. Another topic is the identification of dynamic network of expression. In the ongoing work on low-grade gliomas, a local database of 400 patients will be soon available to construct models. We plan to extend it through national and international collaborations (Montpellier CHU, Montreal CRHUM). Our aim is to build a decision-aid tool for personalised medicine. In the same context, there is a topic of clustering analysis of a brain cartography obtained by sensorial simulations during awake surgery.

4.2 Genomic data and micro-organisms population

Despite of his 'G' in the name of BIGS, Genetics is not central in the applications of the team. However, we want to contribute to a better understanding of the correlations between genes trough their expression data and of the genetic bases of drug response and disease. We have contributed to methods detecting proteomics and transcriptomics variables linked with the outcome of a treatment.

4.3 Epidemiology and e-health

We have many works to do in our ongoing projects in the context of personalized medicine with CHU Nancy. They deal with biomarkers research, prognostic value of quantitative variables and events, scoring, and adverse events. We also want to develop our expertise in rupture detection in a project with APHP (Assistance Publique Hôpitaux de Paris) for the detection of adverse events, earlier than the clinical signs and symptoms. The clinical relevance of predictive analytics is obvious for high-risk patients such as those with solid organ transplantation or severe chronic respiratory disease for instance. The main challenge is the rupture detection in multivariate and heterogeneous signals (for instance daily measures of electrocardiogram, body temperature, spirometry parameters, sleep duration, etc.). Other collaborations with clinicians concern foetopathology and we want to use our work on conditional distribution function to explain fetal and child growth. We have data from the "Service de foetopathologie et de placentologie" of the "Maternité Régionale Universitaire" (CHU Nancy).

4.4 Dynamics of telomeres

Telomeres are disposable buffers at the ends of chromosomes which are truncated during cell division; so that, over time, due to each cell division, the telomere ends become shorter. By this way, they are markers of aging. Through a collaboration with Pr A. Benetos, geriatrician at CHU Nancy, we recently obtained data on the distribution of the length of telomeres from blood cells. With members of Inria team TOSCA, we want to work in three connected directions: (1) refine methodology for the analysis of the available data; (2) propose a dynamical model for the lengths of telomeres and study its mathematical properties (long term behavior, quasi-stationarity, etc.); and (3) use these properties to develop new

statistical methods. A slot of postdoc position is already planned in the Lorraine Université d'Excellence, LUE project GEENAGE (managed by CHU Nancy).

5 Social and environmental responsibility

We followed Inria's recommendations to get involved in the fight against COVID 19. We responded to the WHO's encouragement, relayed by our mathematical colleagues at the national level, to conduct seroprevalence studies in randomly drawn samples of the population. This is the purpose of the COVAL study described in the results section, initiated by Pierre Vallois.

6 Highlights of the year

The highlight of the year is the merger between BIGS and the members of the former TOSCA team, specialised in modelling for biological sciences and medicine: Nicolas Champagnat, Coralie Fritsch, Denis Villemonais and their post-doc and PhD students. The other highlights of the year are, unsurprisingly, those of the pandemic: most of our teachers devoted a lot of time to distance learning. Other researchers, especially PhD students, suffered from the lack of contacts and meetings. Part of the team was involved in supervising a seroprevalence study. Thanks to the quality of the collaboration with hospital doctors in this study, we are now involved in modelling the amount of coronavirus in wastewater in order to predict the number of hospital admissions.

7 New software and platforms

7.1 New software

7.1.1 Angio-Analytics

Keywords: Health, Cancer, Biomedical imaging

Scientific Description: This tool allows the pharmacodynamic characterization of anti-vascular effects in anti-cancer treatments. It uses time series of in vivo images provided by intra-vital microscopy. Such in vivo images are obtained owing to skinfold chambers placed on mice skin. The automatized analysis is split up into two steps that were completely performed separately and manually before. The first steps corresponds to image processing to identify characteristics of the vascular network. The last step is the system identification of the pharmacodynamic response and the statistical analysis of the model parameters.

Functional Description: Angio-Analytics allows the pharmacodynamic characterization of anti-vascular effects in anti-cancer treatments.

Contact: Thierry Bastogne

Participant: Thierry Bastogne

7.1.2 ARMADA

Name: A Statistical Methodology to Select Covariates in High-Dimensional Data under Dependence

Keywords: Biostatistics, Aggregated methods, High Dimensional Data, Personalized medicine, Variable selection

Functional Description: Two steps variable selection procedure in a context of high-dimensional dependent data but few observations. First step is dedicated to eliminate dependence between variables (clustering of variables, followed by factor analysis inside each cluster). Second step is a variable selection using by aggregation of adapted methods. <https://hal.archives-ouvertes.fr/hal-02173568>

News of the Year: This package is a new one.

URL: https://cran.r-project.org/web/packages/armada/

Publication: hal-02363338

Contacts: Aurélie Muller, Anne Gégout-Petit

Participants: Aurélie Muller, Anne Gégout-Petit

7.1.3 kosel

Name: Variable Selection by Revisited Knockoffs Procedures

Keywords: Variable selection, Regression

Functional Description: Performs variable selection for many types of L1-regularised regressions using the revisited knockoffs procedure. This procedure uses a matrix of knockoffs of the covariates independent from the response variable Y. The idea is to determine if a covariate belongs to the model depending on whether it enters the model before or after its knockoff. The procedure suits for a wide range of regressions with various types of response variables. Regression models available are exported from the R packages 'glmnet' and 'ordinalNet'. Based on the paper linked to via the URL below: Gegout A., Gueudin A., Karmann C. (2019) <arXiv:1907.03153>

News of the Year: This package is a new one.

URL: https://cran.r-project.org/web/packages/kosel/kosel.pdf

Publication: hal-01799914

Contacts: Clémence Karmann, Aurélie Muller

Participants: Clémence Karmann, Aurélie Muller, Anne Gégout-Petit

7.1.4 SesIndexCreatoR

Functional Description: This package allows computing and visualizing socioeconomic indices and categories distributions from datasets of socioeconomic variables (These tools were developed as part of the EquitArea Project, a public health program).

URL: http://www.equitarea.org/documents/packages_1.0-0/

Contact: Benoît Lalloué

Participants: Benoît Lalloué, Jean-Marie Monnez, Nolwenn Le Meur, Severine Deguen

7.1.5 In silico

Name: In silico design of nanoparticles for the treatment of cancers by enhanced radiotherapy

Keywords: Bioinformatics, Cancer, Drug development

Functional Description: To speed up the preclinical development of medical engineered nanomaterials, we have designed an integrated computing platform dedicated to the virtual screening of nanostructured materials activated by X-ray making it possible to select nano-objects presenting interesting medical properties faster. The main advantage of this in silico design approach is to virtually screen a lot of possible formulations and to rapidly select the most promising ones. The platform can currently handle the accelerated design of radiation therapy enhancing nanoparticles and medical imaging nano-sized contrast agents as well as the comparison between nano-objects and the optimization of existing materials.

Contact: Thierry Bastogne

Participant: Thierry Bastogne

7.1.6 HSPOR

Name: Hidden Smooth Polynomial Regression for Rupture Detection

Keywords: Polynomial regression, Rupture detection

Functional Description: Several functions that allow by different methods to infer a piecewise polynomial regression model under regularity constraints, namely continuity or differentiability of the link function. The implemented functions are either specific to data with two regimes, or generic for any number of regimes, which can be given by the user or learned by the algorithm.

News of the Year: This package is a new one

URL: https://cran.r-project.org/web/packages/HSPOR/

Contact: Florine Greciet

Participants: Florine Greciet, Romain Azais, Anne Gégout-Petit

7.1.7 cvmgof

Keywords: Regression, Test, Estimators

Scientific Description: Many goodness-of-fit tests have been developed to assess the different assumptions of a (possibly heteroscedastic) regression model. Most of them are "directional" in that they detect departures from a given assumption of the model. Other tests are "global" (or "omnibus") in that they assess whether a model fits a dataset on all its assumptions. cvmgof focuses on the task of choosing the structural part of the regression function because it contains easily interpretable information about the studied relationship. It implements 2 nonparametric "directional" tests and one nonparametric "global" test, all based on generalizations of the Cramer-von Mises statistic.

Functional Description: cvmgof is an R library devoted to Cramer-von Mises goodness-of-fit tests. It implements three nonparametric statistical methods based on Cramer-von Mises statistics to estimate and test a regression model.

News of the Year: New version available on CRAN website since Jan 11 2021 Preprint available on HAL since Jan 7 2021

URL: https://cran.r-project.org/web/packages/cvmgof/index.html

Publication: hal-03101612v1

Contacts: Sandie Ferrigno, Romain Azais

Participants: Sandie Ferrigno, Marie-José Martinez, Romain Azais

7.1.8 starm R

Name: Spatio-Temporal Autologistic Regression Model, package R

Keywords: Spatio-temporal, Autologistic model

Functional Description: Estimation and model selection of the two-time centered autologistic regression model based on Gegout-Petit A., Guerin-Dubrana L., Li S. "A new centered spatio-temporal autologistic regression model. Application to local spread of plant diseases." 2019 <arXiv:1811.06782>. Application for the spatio-temporal modelling of the spread of a disease on a grid over time.

Contact: Anne Gégout-Petit

8 New results

8.1 Stochastic modelling

Participants Anne Gégout-Petit, Ulysse Herbach, Sophie Wantz-Mézières, Pierre Vallois.

8.1.1 Modelling of diffuse low-grade gliomas growth

We are continuing our research on the modelling of the growth of low grade diffuse gliomas. We propose an original MRI-based method to quantify gliomas brain infiltration, easy to implement and to interpret for Neuro-oncologists. The aim is to guide the treatment strategy in giving functional information using only anatomical knowlege and conventional MRI sequences. This work has been the subject of a conference paper [15].

A retrospective survival study over 35 years follow-up has been done [9].

8.1.2 Reconstruction of epigenetic landscapes from single-cell data

The aim is to better understand how living cells make decisions (e.g., differentiation of a stem cell into a particular specialized type), seeing decision-making as an emergent property of an underlying complex molecular network. Indeed, it is now proven that cells react probabilistically to their environment: cell types do not correspond to fixed states, but rather to "potential wells" of a certain energy landscape (representing the energy of the possible states of the cell) that we are trying to reconstruct. A first paper proposing a reconstruction method has been submitted [24] in the framework of an international collaboration (USA, Switzerland, France). Another paper is about to be submitted, dealing more specifically with the inference of the underlying networks.
Joint work with Nan Papili Gao (ETH Zurich), Olivier Gandrillon (ENS Lyon), András Páldi (EPHE, Paris), and Rudiyanto Gunawan (University at Buffalo, New York)

8.2 Optimal control of Markov processes

Participants Bruno Scherrer, Nino Vieillard.

In [13], we adapt the optimization's concept of momentum to reinforcement learning. Seeing the state-action value functions as an analog to the gradients in optimization, we interpret momentum as an average of consecutive q-functions. We derive Momentum Value Iteration (MoVI), a variation of Value iteration that incorporates this momentum idea. Our analysis shows that this allows MoVI to average errors over successive iterations. We show that the proposed approach can be readily extended to deep learning. Specifically, we propose a simple improvement on DQN based on MoVI, and experiment it on Atari games. This work has been published in the AISTATS conference.

Recent Reinforcement Learning (RL) algorithms making use of Kullback-Leibler (KL) regularization as a core component have shown outstanding performance. Yet, only little is understood theoretically about why KL regularization helps, so far. In [12], we study KL regularization within an approximate value iteration scheme and show that it implicitly averages q-values. Leveraging this insight, we provide a very strong performance bound, the very first to combine two desirable aspects: a linear dependency to the horizon (instead of quadratic) and an error propagation term involving an averaging effect of the estimation errors (instead of an accumulation effect). We also study the more general case of an additional entropy regularizer. The resulting abstract scheme encompasses many existing RL algorithms. Some of our assumptions do not hold with neural networks, so we complement this theoretical analysis

with an extensive empirical study. This work has been accepted to the Neurips conference and selected for oral presentation (selection rate: 1.1% of all submissions)
Joint work with Matthieu Geist, Olivier Pietquin, Rémi Munos and Tadashi Kozuno (Google Brain Paris).

8.3 Regression and machine learning

Participants Thierry Bastogne, Sandie Ferrigno, Anne Gégout-Petit, Clémence Karmann, Benoît Lalloué, Jean-Marie Monnez, Pauline Guyot, Aurélie Gueudin, Clémence Karmann, Sophie Wantz-Mézières.

8.3.1 Cramér–von Mises goodness-of-fit tests in regression models

Many goodness-of-fit tests have been developed to assess the different assumptions of a (possibly heteroscedastic) regression model. Most of them are 'directional' in that they detect departures from a given assumption of the model. Other tests are 'global' (or 'omnibus') in that they assess whether a model fits a dataset on all its assumptions. We focus on the task of choosing the structural part of the regression function because it contains easily interpretable information about the studied relationship. We consider 2 nonparametric 'directional' tests and one nonparametric 'global' test, all based on generalizations of the Cramér–von Mises statistic.

To perform these goodness-of-fit tests, we develop the R package cvmgof (https://hal.archives-ouvertes.fr/hal-02014516), an easy-to-use tool for practitioners, available from the Comprehensive R Archive Network (https://CRAN.R-project.org/package=cvmgof). The use of the library is illustrated through a tutorial on real data and simulation studies are carried out in order to show how the package can be exploited to compare the 3 implemented tests. The practitioner can also easily compare the test procedures with different kernel functions, bootstrap distributions, numbers of bootstrap replicates, or bandwidths. A first article [22] has been submitted on this work.

To complete this work, it would be interesting to assess the other assumptions of a regression model such as the functional form of the variance or the additivity of the random error term. It should be noted that this can already be done using Ducharme and Ferrigno test implemented in cvmgof since it is a global test. However, it would be relevant to compare the results obtained from Ducharme and Ferrigno test with the ones obtained from other directional tests, especially developed to assess one of these specific assumptions. The implementation of these directional tests would enrich cvmgof package and offer a complete easy-to-use tool for validating regression models. Moreover, the assessment of the overall validity of the model when using several directional tests could be compared with that done when using only a global test. In particular, the well-known problem of multiple testing could be discussed by comparing the results obtained from multiple test procedures with those obtained when using a global test strategy. Another perspective of this work would be to develop a similar tool for other statistical models widely used in practice such as generalized linear models.
Join work with Romain Azaïs (INRIA, ENS Lyon) and Marie-José Martinez (LJK, Université Grenoble Alpes).

8.3.2 The revisited knockoffs method for variable selection in L1-penalized regressions

We consider the problem of variable selection in regression models. In particular, we are interested in selecting explanatory covariates linked with the response variable and we want to determine which covariates are relevant, that is which covariates are involved in the model. In this framework, we deal with L1-penalized regression models. To handle the choice of the penalty parameter to perform variable selection, we develop a new method based on the knockoffs idea. This revisited knockoffs method is general, suitable for a wide range of regressions with various types of response variables. Besides, it also works when the number of observations is smaller than the number of covariates and gives an order of importance of the covariates. Finally, we provide many experimental results to corroborate our method

and compare it with other variable selection methods. This work is published in [5] and is implemented in package 'kosel'.

The next subsections are dedicated to online data analysis

8.3.3 Widening the scope of an eigenvector stochastic approximation process and application to streaming PCA and related methods

Accepted in Journal of Multivariate Analysis in October 2020 [8].

We prove the almost sure convergence of processes of Oja type to eigenvectors of the expectation of a random matrix while relaxing the i.i.d. assumptions on the observed random matrices. As an application of this generalization, we can perform the online PCA of a random vector Z when there is a data stream of i.i.d. observations of Z, even when both the metric used M and the expectation of Z are unknown and estimated online. Moreover, in order to update the stochastic approximation process at each step we are no more bound to using only a data mini-batch of observations of Z, but we can use all the previous observations up to the current step without storing them. This is useful not only when dealing with streaming data but also with Big Data as on can process it sequentially as a data stream. In addition, the general framework of this process, unlike other algorithms in the literature, covers also the case of factorial methods related to PCA.

In collaboration with A. Skiredj.

8.3.4 Streaming constrained binary logistic regression with online standardized data

Accepted in "Journal of Applied Statistics" in December 2020 [7].

Online learning is a method for analyzing very large datasets ("big data") as well as data streams. In this article, we consider the case of constrained binary logistic regression and show the interest of using processes with an online standardization of the data, in particular to avoid numerical explosions or to allow the use of shrinkage methods. We prove the almost sure convergence of such a process and propose using a piecewise constant step-size such that the latter does not decrease too quickly and does not reduce the speed of convergence. We compare twenty-four stochastic approximation processes with raw or online standardized data on five real or simulated datasets. Results show that, unlike processes with raw data, processes with online standardized data can prevent numerical explosions and yield the best results.

In collaboration with E. Albuisson.

8.3.5 Construction and update of an online ensemble score involving linear discriminant analysis and logistic regression

Submitted in Februray 2021 [26], [20].

The present aim is to update, upon arrival of new learning data, the parameters of a score constructed with an ensemble method involving linear discriminant analysis and logistic regression in an online setting, without the need to store all of the previously obtained data. Poisson bootstrap and stochastic approximation processes were used with online standardized data to avoid numerical explosions, the convergence of which has been established theoretically. This empirical convergence of online ensemble scores to a reference "batch" score was studied on five different datasets from which data streams were simulated, comparing six different processes to construct the online scores. For each score, 50 replications using a total of $10N$ observations (N being the size of the dataset) were performed to assess the convergence and the stability of the method, computing the mean and standard deviation of a convergence criterion. A complementary study using $100N$ observations was also performed. The best processes were averaged processes using online standardized data and a piecewise constant step-size.

8.3.6 Change-point detection theresholds in the sequential context

Our work around change-point theresholds for the score-based CUSUM statistic in a sequential context has been published [11]. In this paper, we consider the score-based cumulative sum statistic and propose to evaluate the detection performance of somethresholds on simulated data. Three thresholds come from the literature: the Wald constant, the empirical constant, and the conditional empirical instantaneous threshold. Two new thresholds are built by a simulation-based procedure: the first one is instantaneous, the second is a dynamical version of the previous one. The thresholds' performance measured by an estimation of the mean time between false alarm (MTBFA) and the average detection delay (ADD) are evaluated on independent and autocorrelated data for several scenarios, according to the detection objective and the real change in the data. The simulations allow us to compare the difference between the thresholds' results and to see that their performances prove to be robust when a parameter of the prechange regime is poorly estimated or when the data independence assumption is violated. We found also that the conditional empirical threshold is the best at minimizing the detection delay while maintaining the given false alarm rate. However, on real data, we suggest to use the dynamic instantaneous threshold because it is the easiest to build for practical implementation.

Our collaboration with APHP could not succeed because of the great delay in data collection. To apply our algorithms to real data, we turned to some EMG signal data provided by INRS. The study concerns the development of trapezius muscle myalgia in the workplace. We apply change-point detection to characterise different computer activities carried out during an experimental day.

8.4 Statistical learning and application in health

Participants Ulysse Herbach, Sandie Ferrigno, Anne Gégout-Petit, Aurélie Gueudin, Pierre Vallois, Benoît Lalloué, Jean-Marie Monnez, Nicolas Thorr, Pierre Vallois.

8.4.1 Estimation of reference curves for fetal weight

In Epidemiology, we are working with INSERM to study fetal development in the last two trimesters of pregnancy. Reference or standard curves are required in this kind of biomedical problems. Values which lie outside the limits of these reference curves may indicate the presence of disorder. Data are from the French EDEN mother-child cohort (INSERM). It's a mother-child cohort study investigating the prenatal and early postnatal determinants of child health and development. 2002 pregnant women were recruited before 24 weeks of amenorrhoea in two maternity clinics from middle-sized French cities (Nancy and Poitiers). From May 2003 to September 2006, 1899 newborns were then included. The main outcomes of interest are fetal (via ultra-sound) and postnatal growth, adiposity development, respiratory health, atopy, behaviour and bone, cognitive and motor development. We are studying fetal weight that depends on the gestional age in the second and the third trimesters of mother's pregnancy. Some classical empirical and parametric methods as polynomial are first used to construct these curves. Polynomial regression is one of the most common parametric approach for modelling growth data espacially during the prenatal period. However, some of them requires strong assumptions. So, we propose to work with semi-parametric LMS method, by modifying the response variable (fetal weight) with a Box-cox transformation. A first article detailing these methodologies applied to the data is being written.

Alternative nonparametric methods as Nadaraya-Watson kernel estimation, local polynomial estimation, B-splines or cubic splines are also developed in this context to construct these curves. The practical implementation of these methods required working on smoothing parameters or choice of knots for the different types of nonparametric estimation. In particular, optimal choice of these parameters has been proposed. Then, a first version of an R package has been developed to propose a tool to construct nonparametric reference curves. This should be submitted to CRAN very soon. In addition, a graphical interface (GUI) intended for practitioners has been developed to allow intuitive visualization of the results given by the package.

Join work with Myriam Maumy-Bertrand (IRMA, Université de Strasbourg) and INSERM.

8.4.2 Construction of parsimonious event risk scores by an ensemble method. An illustration for short-term predictions in chronic heart failure patients from the GISSI-HF trial

Submitted in December 2020 [27].

Heart failure (HF) is a worldwide major cause of mortality and morbidity for which many predictive scores have been defined. Selecting which explanatory variables to include in a given score is a common difficulty, as a balance must be found between statistical fit and practical application. This article presents a methodology for constructing parsimonious event scores combining a stepwise selection of variables with ensemble scores obtained by aggregation of several scores, using several classifiers, bootstrap samples and various modalities of random selection of variables. The stepwise selection allows constructing a succession of scores with the practitioner able to choose which score best fits his or her needs. The methods proposed herein can be reproduced on any set of variables as long as the training dataset comprises a sufficient number of cases. Three methods were compared in an application to construct parsimonious short-term scores in chronic HF patients. The working sample consisted of 11,411 couples patient-visit dyads from the GISSI-HF database, with 5,595 events and 5,816 non-events. Sixty-two candidate explanatory variables were studied. Focusing on the fastest method, four scores were constructed, yielding out-of-bag AUCs ranging from 0.81 (26 variables) to 0.76 (2 variables). These results are slightly better than those obtained by other scores reported in the literature using a similar number of variables.

In collaboration with E. Albuisson and D. Lucci.

8.4.3 Modeling and estimation of circulating tumor DNA (ctDNA) dynamics for detecting resistance to targeted therapies

Continuation of the ITMO Cancer project, supervised by Nicolas Champagnat, concerning the modeling of circulating tumor DNA (ctDNA) to detect the appearance of resistance to targeted therapies (personalized medicine). After a phase of investigation of possible scenarios in collaboration with Alexandre Harlé of the Institute of Cancerology of Lorraine (ICL), a final model was selected. Based on a mathematical analysis, the members of the project then designed a statistical inference algorithm (learning the parameters of the model, including the genealogical tree of mutations for each patient) which is intended to be validated on real data currently being acquired at the Nancy CHRU. The general idea is to exploit a "variational principle" that allows to explore the discrete space of family trees, of very large size, through a "pivot" space of continuous parameters, easy to optimize (and in reasonable numbers). An article detailing the model and its inference is currently being written.

In collaboration with N. Champagnat and C. Fritsch.

8.4.4 A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles in oncology

We propose a new methodology for selecting and ranking covariates associated with a variable of interest in a context of high-dimensional data under dependence but few observations. The methodology successively intertwines the clustering of covariates, decorrelation of covariates using Factor Latent Analysis, selection using aggregation of adapted methods and finally ranking. A simulation study shows the interest of the decorrelation inside the different clusters of covariates. We first apply our method to transcriptomic data of 37 patients with advanced non-small-cell lung cancer who have received chemotherapy, to select the transcriptomic covariates that explain the survival outcome of the treatment. Secondly, we apply our method to 79 breast tumor samples to define patient profiles for a new metastatic biomarker and associated gene network in order to personalize the treatments. This work is published in [2] and is implemented in R package 'ARMADA'.

In collaboration with T. Boukhobza and H. Dumond from CRAN and B. Bastien from biopharmaceutical industry Transgene.

8.4.5 Project linked with the COVID 19 pandemic

Pierre Vallois is the scientific coordinator of the seroprevalence study COVAL Nancy held in Nancy in July 2020 in collaboration with CHRU de Nancy (CIC épidémiologie clinique and Laboratoire de Virologie).

Background. The World Health Organisation recommends monitoring the circulation of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). We aimed to estimate anti–SARS-CoV-2 total immunoglobulin (IgT) antibody seroprevalence and describe symptom profiles and in vitro seroneutralization in Nancy, France, in spring 2020.

Methods. Individuals were randomly sampled from electoral lists and invited with household members over 5 years old to be tested for anti–SARS-CoV-2 (IgT, i.e. IgA/IgG/IgM) antibodies by ELISA (Bio-rad). Serum samples were classified according to seroneutralization activity 50 % (NT50) on Vero CCL-81 cells. Age- and sex-adjusted seroprevalence was estimated. Subgroups were compared by chi-square or Fisher exact test and logistic regression.

Results. Among 2006 individuals, 43 were SARS-CoV-2–positive; the raw seroprevalence was 2.1 % (95 % confidence interval 1.5 to 2.9), with adjusted metropolitan and national standardized seroprevalence 2.5 % (1.8 to 3.3) and 2.3 % (1.7 to 3.1). Seroprevalence was highest for 20- to 34-year-old participants (4.7 % [2.3 to 8.4]), within than out of socially deprived area (2.5 % vs 1 %, P=0.02) and with than without intra-family infection (p<10-6). Moreover, 25 % (23 to 27) of participants presented at least one COVID-19 symptom associated with SARS-CoV-2 positivity (p<10-13), with anosmia or ageusia highly discriminant (odds ratio 27.8 [13.9 to 54.5]), associated with dyspnea and fever. Among the SARS-CoV-2-positives, 16.3 % (6.8 to 30.7) were asymptomatic. For 31 of these individuals, positive seroneutralization was demonstrated in vitro.

Conclusions. In this population of very low anti-SARS-CoV-2 antibody seroprevalence, a beneficial effect of the lockdown can be assumed, with frequent SARS-CoV-2 seroneutralization among IgT-positive patients.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

- R. Azaïs, A. Gégout-Petit, F. Greciet collaborated with SAFRAN Aircraft Engines (through a 2016-2019 contract). SAFRAN Aircraft Engines designs and products aircraft engines. For the design of pieces, they have to understand the mechanism of crack propagation under different conditions. BIGS models crack propagation with Piecewise Deterministic Markov Processes (PDMP).

- B. Scherrer collaborate with Google brain on reinforcement learning in the framework of the PhD thesis of Nino Vieillard

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other international programs

In Fall 2020, Bruno Scherrer was invited for 4 months in Berkeley to participate to Simons Institute Programme on the Theory of Reinforcement Learning. Due to the Covid constraints, the semester was eventually hold online.

10.2 International research visitors

10.2.1 Visits of international scientists

Juhyun Park (Lancaster University) visited Nancy for one week in the framework of her collaboration with A. Gégout-Petit on statistical test for paired distribution.

10.3 National initiatives

- FHU CARTAGE (Fédération Hospitalo Universitaire Cardial and ARTerial AGEing ; leader : Pr Athanase Benetos), Jean-Marie Monnez, Benoît Lalloué, Anne Gégout-Petit.
- RHU Fight HF (Fighting Heart Failure; leader: Pr Patrick Rossignol), located at the University Hospital of Nancy, Jean-Marie Monnez, Benoît Lalloué.
- Project "Handle your heart", team responsible for the creation of a drug prescription support software for the treatment of heart failure, head: Jean-Marie Monnez.
- A. Gégout-Petit, N. Sahki, S. Mézières are involved in the learning aspect of the clinical protocol "EOLEVAL" with Assistance Publique des Hopitaux de Paris (APHP).
- "ITMO Physics, mathematics applied to Cancer" (2017-2019): "Modeling ctDNA dynamics for detecting targeted therapy", Funding organisms: ITMO Cancer, ITMO Technologies pour la santé de l'alliance nationale pour les sciences de la vie et de la santé (AVIESAN), INCa, Leader: N. Champagnat (Inria TOSCA), Participants: A. Gégout-Petit, A. Muller-Gueudin, P. Vallois, U. Herbach.
- PEPS AMIES (2019-2020), Etude Biométrique en foetopathologie et développement de l'enfant, Collaboration between Institut Elie Cartan and the CRESS INSERM, S. Ferrigno.
- Modular, multivalent and multiplexed tools for dual molecular imaging (2017-2020), Funding organism: ANR, Leader: B Kuhnast (CEA). Participant: T. Bastogne.
- Sophie Mézières belongs to GDR 720 ISIS, Funding organism: CNRS, leader: Laure Blanc-Féraud.

10.4 Regional initiatives

- CHRU de Nancy. We have good collaborations with several researchers from CHRU de Nancy. We are involved in LUE Impact Geenage in research axis telomeres.
- CHRU de Nancy. Joint initiave of the Sars-Cov2 seroprevalence study COVAL Nancy with CIC épidémiologie. https://clinicaltrials.gov

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Journal

- Ulysse Herbach was a guest editor for the journal "Mathematical Biosciences and Engineering" (special edition "Cells as dynamical systems").

11.1.2 Invited talks

- Anne Gégout-Petit was invited to a plenary communication in "Journées de Statistique", Nice, France.
- Ulysse Herbach was invited to a plenary communication in conference "Interplay between Oncology, Mathematics and Numerics", Paris, France.

11.1.3 Research administration

- Anne Gégout-Petit is the head of “Institut Élie Cartan de Lorraine” (mathematics laboratory of Université de Lorraine) since September 1st.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

Bruno Scherrer and Ulysse Herbach excepted, BIGS members have teaching obligations at "Université Lorraine" and are teaching at least 192 hours each year. They teach probability and statistics at different levels (Licence, Master, Engineering school). Many of them have pedagogical responsibilities.

- A. Gégout-Petit: Head of the Master 2 "Ingénierie Mathématique pour la science des données (Mathematical Engineering for data science)", Université de Lorraine
- T. Bastogne is in charge research master program "Santé Numérique et Imagerie Médicale" with the Faculty of Medicine, Université de Lorraine, France
- Master: S. Ferrigno, Experimental designs, 4.5h, M1, fourth year of EEIGM, Université de Lorraine, France
- Master: S. Ferrigno, Data analyzing and mining, 63h, M2, third year of Ecole des Mines, Université de Lorraine, France
- Master: S.Ferrigno, Modeling and forecasting, 43h, M1, second year of Ecole des Mines, Université de Lorraine, France
- Master: S.Ferrigno, Training projects, 18h, M1/M2, second and third year of Ecole des Mines, Université de Lorraine, France
- Master: A. Muller-Gueudin, Probability and Statistics, 160h, second year of ENSEM and ENSAIA, University of Lorraine, France.
- Master: A. Muller-Gueudin, Scientific calculation with Matlab, 20h, second year of ENSAIA, University of Lorraine, France.
- Master: A.Gégout-Petit, Statistics, modeling, 15h, future teacher, Université de Lorraine, France
- Master: A.Gégout-Petit, Statistics, modeling, data analysis, 80h, master in applied mathematics, Université de Lorraine, France
- Master: S. Wantz-Mézières, Learning and analysis of medical data, 36h, with J.M. Moureaux, Master SNIM, Université de Lorraine, France
- Licence: S. Wantz-Mézières, Applied mathematics for management, financial mathematics, Probability and Statistics, 160h, I.U.T. (L1/L2/L3)
- Licence: S. Wantz-Mézières, Probability, 100h, first year in Telecom Nancy engineering school (initial and apprenticeship cursus)
- Licence: A. Muller-Gueudin, Statistics, 60h, first year of ENSAIA, University of Lorraine, France.
- Licence: S. Ferrigno, Descriptive and inferential statistics, 60h, L2, second year of EEIGM, Université de Lorraine, France
- Licence: S. Ferrigno, Statistical modeling, 60h, L2, second year of EEIGM, Université de Lorraine, France
- Licence: S. Ferrigno, Mathematical and computational tools, 20h, L3, third year of EEIGM, Université de Lorraine, France
- Licence: S. Ferrigno, Training projects, 20h, L1/L3, first, second and third year of EEIGM, Université de Lorraine, France

11.2.2 Supervision

Defended PhD thesis

- PhD: Florine Greciet, "Modèles markoviens déterministes par morceaux cachés pour la propagation de fissures", grant CIFRE SAFRAN AIRCRAFT ENGINES, Advisors : R. Azaïs, A. Gégout-Petit, Université de Lorraine, defense on January, 2020.

PhD thesis

- PhD: Pauline Guyot, "Modélisation et Simulation de l'Electrocardiogramme d'un Patient Numérique", Grant : CIFRE-Cybernano. Advisors: T. Bastogne, E. H. Djermoune.
- PhD: Nassim Shaki, "Détection de rupture dans des signaux multivariés pour la prédiction d'événement redouté à partir de paramètres physiologiques recueillis par capteurs connectés après greffe pulmonaire", grant Inria-Cordis. Advisors: A. Gégout-Petit, S. Wantz-Mézières, M. d'Ortho.
- PhD: Nino Vieillard, "Deep Reinforcement Learning", CIFRE grand with Google Brain Paris. Advisors: B. Scherrer, M. Geist.

Post-doctoral positions

- Benoît Lalloué, contract research engineer for two years, RHU Fight RF, supervised by Jean-Marie Monnez.
- Postdoc: Emma Horton, Telomer Modelling, grant LUE GEENAGE. Advisors: A. Gégout-Petit, D. Villemonais. Emma was hired CR Inria at Bordeaux Sud-Ouest (ASTRAL team)

Other

- Master: all BIGS members regularly supervise project and internship of master IMOI students.
- Engineering school: all BIGS members regularly supervise projects of "École des Mines", ENSEM, EEIGM or Télécom-Nancy students.

11.2.3 Juries

- Anne Gégout-Petit wrote the report and participated to the jury of the Phd defense of Titin Agustin NENGSIH, Strasbourg University, March 16th.
- Anne Gégout-Petit wrote the report and participated to the jury of the HDR defense of Maud Delattre, Paris-Saclay University, November 6th.
- Anne Gégout-Petit is member of the "Jury du prix de thèse AMIES".
- Bruno Scherrer participated to the jury of the Phd defense of Matthieu Guillot, G-SCOP lab, Grenoble INP, July 3rd.
- Bruno Scherrer participated to the jury of the Phd defense of Rituraj Kaushik, July 23rd.

11.3 Popularization

11.3.1 Education

- Sandie Ferrigno: Advisor of a group of students (EEIGM), "La main à la Pâte" project, elementary schools, Nancy, January-June 2020.
- Sandie Ferrigno: Advisor of a group of students (EEIGM), "Energies renouvelables", "La main à la Pâte" project, Institut médico-éducatif (IME), Commercy, January 2020.
- Sandie Ferrigno: Advisor of a group of students (EEIGM), "L'Astronomie", Cgénial project, Collège Paul Verlaine, Malzéville, January 2020.
- Sandie Ferrigno: Advisor of a group of students (EEIGM), "Le Chocolat", Cgénial project, Collège de la Craffe, Nancy, January 2020.

11.3.2 Interventions

- Sophie Wantz-Mézières was part of the organization of a thematic and multidisciplinary week "Neurosciences, Neuro-oncologie et Numérique" for students from Télécom-Nancy and Faculté de Médecine de Nancy, janvier 2020.
- Bruno Scherrer made detailed simulations of the reform for the retirement system that has been considered by Philippe's government in France [28].

12 Scientific production

12.1 Publications of the year

International journals

[1] J.-B. Barbry, A.-S. Poinsard, T. Bastogne and O. Balland. 'Short-term effects of ocular 2% dorzolamide, 0.5% timolol or 0.005% latanoprost on the anterior segment architecture in healthy cats: a prospective study.' In: *Open Veterinary Journal* (2020). URL: https://hal.archives-ouvertes.fr/hal-02396549.

[2] B. Bastien, T. Boukhobza, H. Dumond, A. Gégout-Petit, A. Muller-Gueudin and C. Thiébaut. 'A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles in oncology'. In: *Journal of Applied Statistics* (2021), p. 23. DOI: 10.1080/02664763.2020.1837083. URL: https://hal.archives-ouvertes.fr/hal-02173568.

[3] A. Buessler, T. Chouihed, K. Duarte, A. Bassand, M. Huot-Marchand, Y. Gottwalles, A. Pénine, E. André, L. Nace, D. Jaeger, M. Kobayashi, S. Coiro, P. Rossignol and N. Girerd. 'Accuracy of Several Lung Ultrasound Methods for the Diagnosis of Acute Heart Failure in the ED: A Multicenter Prospective Study'. In: *Chest* 157.1 (Jan. 2020), pp. 99–110. DOI: 10.1016/j.chest.2019.07.017. URL: https://hal.univ-lorraine.fr/hal-02512447.

[4] M. Ferrua, E. Minvielle, A. Fourcade, B. Lalloué, C. Sicotte, M. Di Palma and O. Mir. 'How to Design a Remote Patient Monitoring System? A French Case Study'. In: *BMC Health Services Research* 20.1 (Dec. 2020). DOI: 10.1186/s12913-020-05293-4. URL: https://hal.archives-ouvertes.fr/hal-02950440.

[5] A. Gégout-Petit, A. Gueudin-Muller and C. Karmann. 'The revisited knockoffs method for variable selection in L 1 -penalized regressions'. In: *Communications in Statistics - Simulation and Computation* (July 2020). DOI: 10.1080/03610918.2020.1775850. URL: https://hal.archives-ouvertes.fr/hal-02903837.

[6] P. Guyot, E.-H. Djermoune, B. Chenuel and T. Bastogne. 'A signal demodulation-based method for the early detection of Cheyne-Stokes respiration'. In: *PLoS ONE* 15.3 (12th Mar. 2020), e0221191. DOI: 10.1371/journal.pone.0221191. URL: https://hal.archives-ouvertes.fr/hal-02513384.

[7] B. Lalloué, J.-M. Monnez and E. Albuisson. 'Streaming constrained binary logistic regression with online standardized data'. In: *Journal of Applied Statistics* (2021). DOI: 10.1080/02664763.2020.1870672. URL: https://hal.archives-ouvertes.fr/hal-02156324.

[8] J.-M. Monnez and A. Skiredj. 'Widening the scope of an eigenvector stochastic approximation process and application to streaming PCA and related methods'. In: *Journal of Multivariate Analysis* 182 (Mar. 2021), p. 19. DOI: 10.1016/j.jmva.2020.104694. URL: https://hal.inria.fr/hal-03038206.

[9] T. Obara, M. Blonski, C. Brzenczek, S. Mézières, Y. Gaudeau, C. Pouget, G. Gauchotte, A. Verger, G. Vogin, J.-M. Moureaux, H. Duffau, F. Rech and L. Taillandier. 'Adult diffuse low-grade gliomas: 35-year experience at the Nancy France neurooncology unit'. In: *Frontiers in Oncology* 10 (Oct. 2020), p. 574679. DOI: 10.3389/fonc.2020.574679. URL: https://hal.archives-ouvertes.fr/hal-03024376.

[10] P. Rossignol, K. Duarte, N. Girerd, M. Karoui, J. J. McMurray, K. Swedberg, D. Veldhuisen, S. Pocock, K. Dickstein, F. Zannad and B. Pitt. 'Cardiovascular risk associated with serum potassium in the context of mineralocorticoid receptor antagonist use in patients with heart failure and left ventricular dysfunction'. In: *European Journal of Heart Failure* (9th Jan. 2020). DOI: 10.1002/ejhf.1724. URL: https://hal.univ-lorraine.fr/hal-02516141.

[11] N. Sahki, A. Gégout-Petit and S. Wantz-Mézières. 'Performance study of change-point detection thresholds for cumulative sum statistic in a sequential context'. In: *Quality and Reliability Engineering International* 1-21 (14th July 2020), p. 21. DOI: 10.1002/qre.2723. URL: https://hal.inria.fr/hal-02389331.

International peer-reviewed conferences

[12] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos and M. Geist. 'Leverage the Average: an Analysis of KL Regularization in Reinforcement Learning'. In: NeurIPS - 34th Conference on Neural Information Processing Systems. Vancouver / Online, Canada, 6th Dec. 2020. URL: https://hal.inria.fr/hal-03137351.

[13] N. Vieillard, B. Scherrer, O. Pietquin and M. Geist. 'Momentum in Reinforcement Learning'. In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR : Volume 108. Copyright 2020 by the author(s)*. AISTATS 2020 - 23rd International Conference on Artificial Intelligence and Statistics. Vol. 108. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020. Palermo / Virtual, Italy, 2020. URL: https://hal.inria.fr/hal-03137343.

Conferences without proceedings

[14] L. Batista, M. Milhem, T. Bastogne, F. Clanché, G. Personeni, J.-P. Jehl and G. Gauchard. 'A data-driven classification solution for the timed-up and go test in risk falling assessment'. In: EMBC 2020 - 42nd Engineering in Medicine and Biology Conference. Montréal, Canada, 20th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02568440.

[15] C. Brzenczek, S. Wantz-Mézières, Y. Gaudeau, M. Blonski, F. Rech, T. Obara, J.-M. Moureaux and L. Taillandier. 'An original MRI-based method to quantify the diffuse low-grade glioma brain infiltration'. In: 10th International Conference on Image Processing Theory, Tools and Applications, IPTA'20. Paris, France, 9th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03024433.

[16] J. Deleforterie, L. Hassler and T. Bastogne. 'A dendrogram clustering of lipid nanoparticles'. In: 15th annual event of the ETPN – European Technology Platform on Nanomedicine, ETPN2020. Heraklion, Greece, 14th Oct. 2020. URL: https://hal.univ-lorraine.fr/hal-03109525.

[17] L. Hassler and T. Bastogne. 'Approche bayésienne du Quality-by-Design appliquée à un bioprocédé d'extraction de principe actif'. In: 5th Bioproduction Congress. Lyon, France, 29th Sept. 2020. URL: https://hal.univ-lorraine.fr/hal-03109560.

[18] Y. Kolasa, E. Gandiole and T. Bastogne. 'Quality-by-design development of a patient mobility e-monitoring system'. In: 2nd EAI International Conference on Wearables in Healthcare, EAI HealthWear 2020. Virtual, France, 2020. URL: https://hal.univ-lorraine.fr/hal-03109552.

[19] Y. Kolasa, T. Bastogne, J.-P. Georges and S. Kubler. 'Quality-by-design-engineered pBFT consensus configuration for medical device development'. In: EMBC 2020 - 42nd Engineering in Medicine and Biology Conference. Montreal, Canada, 20th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02568428.

[20] B. Lalloué, J.-M. Monnez and E. Albuisson. 'Convergence d'un score d'ensemble en ligne : étude empirique'. In: 52e Journées de Statistique. Nice, France: https://jds2020.sciencesconf.org/, July 2020. URL: https://hal.archives-ouvertes.fr/hal-02894908.

Doctoral dissertations and habilitation theses

[21] F. Greciet. 'Piecewise polynomial regression for crack propagation'. Université de Lorraine, 22nd Jan. 2020. URL: https://hal.univ-lorraine.fr/tel-02510850.

Reports & preprints

[22] R. Azaïs, S. Ferrigno and M.-J. Martinez. *cvmgof: an R package for Cramér-von Mises goodness-of-fit tests in regression models*. 7th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03101612.

[23] T. Bastogne. *Supplementary material iQbD: a TRL-indexed Quality-by-Design Paradigm for Medical Device Development*. 13th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-02937273.

[24] N. P. Gao, O. Gandrillon, A. Páldi, U. Herbach and R. Gunawan. *Universality of cell differentiation trajectories revealed by a reconstruction of transcriptional uncertainty landscapes from single-cell transcriptomic data*. 5th Feb. 2021. DOI: 10.1101/2020.04.23.056069. URL: https://hal.inria.fr/hal-03132652.

[25] B. Lalloué and J.-M. Monnez. *Ensemble methods and online learning for creation and update of prognostic scores in HF patients*. Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03066040.

[26] B. Lalloué, J.-M. Monnez and E. Albuisson. *Construction and update of an online ensemble score involving linear discriminant analysis and logistic regression*. 8th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03134248.

[27] B. Lalloué, J.-M. Monnez, D. Lucci and E. Albuisson. *Construction of parsimonious event risk scores by an ensemble method. An illustration for short-term predictions in chronic heart failure patients from the GISSI-HF trial*. 23rd Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03040390.

[28] B. Scherrer. *Simulations de carrières et retraites à points dans 3 cadres macro-économiques: modèle du gouvernement Philippe (âge-pivot bloqué), modèle du gouvernement Philippe corrigé (âge-pivot glissant), modèle Destinie2 (avec revalorisation de la fonction publique)*. INRIA, 4th Mar. 2020. URL: https://hal.inria.fr/hal-03137362.

12.2 Cited publications

[29] R. Azaïs, F. Dufour and A. Gégout-Petit. 'Non-Parametric Estimation of the Conditional Distribution of the Interjumping Times for Piecewise-Deterministic Markov Processes'. In: *Scandinavian Journal of Statistics* 41.4 (Dec. 2014), pp. 950–969. DOI: 10.1111/sjos.12076. URL: https://hal.archives-ouvertes.fr/hal-01103700.

[30] R. Azaïs and A. Muller-Gueudin. 'Optimal choice among a class of nonparametric estimators of the jump rate for piecewise-deterministic Markov processes'. In: *Electronic journal of statistics* (2016). URL: https://hal.archives-ouvertes.fr/hal-01168651.

[31] R. Azaïs. 'A recursive nonparametric estimator for the transition kernel of a piecewise-deterministic Markov process'. In: *ESAIM: Probability and Statistics* 18 (2014), pp. 726–749.

[32] R. Azaïs, F. Dufour and A. Gégout-Petit. 'Nonparametric estimation of the jump rate for non-homogeneous marked renewal processes'. In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 49. 4. Institut Henri Poincaré. 2013, pp. 1204–1231.

[33] J. M. Bardet, G. Lang, G. Oppenheim, A. Philippe, S. Stoev and 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.

[34] T. Bastogne, S. Mézières-Wantz, N. Ramdani, P. Vallois and M. Barberi-Heyob. 'Identification of pharmacokinetics models in the presence of timing noise'. In: *Eur. J. Control* 14.2 (2008), pp. 149–157. DOI: 10.3166/ejc.14.149-157. URL: http://dx.doi.org/10.3166/ejc.14.149-157.

[35] T. Bastogne, A. Samson, P. Vallois, S. Wantz-Mézières, S. Pinel, D. Bechet and M. Barberi-Heyob. 'Phenomenological modeling of tumor diameter growth based on a mixed effects model'. In: *Journal of theoretical biology* 262.3 (2010), pp. 544–552.

[36] D. Bertsekas and J. Tsitsiklis. *Neurodynamic Programming*. Athena Scientific, 1996.

[37] H. Biermé, C. Lacaux and H.-P. Scheffler. 'Multi-operator Scaling Random Fields'. Anglais. In: *Stochastic Processes and their Applications* 121.11 (2011). MAP5 2011-01, pp. 2642–2677. DOI: 10.1016/j.spa.2011.07.002. URL: http://hal.archives-ouvertes.fr/hal-00551707/en/.

[38] H. Cardot, P. Cénac and J.-M. Monnez. 'A fast and recursive algorithm for clustering large datasets with k-medians'. In: *Computational Statistics & Data Analysis* 56.6 (2012), pp. 1434–1449.

[39] J. F. Coeurjolly. 'Simulation and identification of the fractional brownian motion: a bibliographical and comparative study'. In: *Journal of Statistical Software* 5 (2000), pp. 1–53.

[40] M. H. Davis. 'Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1984), pp. 353–388.

[41] A. Deya and S. Tindel. 'Rough Volterra equations. I. The algebraic integration setting'. In: *Stoch. Dyn.* 9.3 (2009), pp. 437–477. DOI: 10.1142/S0219493709002737. URL: http://dx.doi.org/10.1142/S0219493709002737.

[42] M. Doumic, M. Hoffmann, N. Krell and L. Robert. 'Statistical estimation of a growth-fragmentation model observed on a genealogical tree'. In: *Bernoulli* 21.3 (2015), pp. 1760–1799.

[43] S. Ferrigno and G. Ducharme. 'Un test d'adéquation global pour la fonction de répartition conditionnelle'. In: *C. R. Math. Acad. Sci. Paris* 341.5 (2005), pp. 313–316. DOI: 10.1016/j.crma.2005.07.003. URL: http://dx.doi.org/10.1016/j.crma.2005.07.003.

[44] S. Ferrigno, M. Maumy-Bertrand and A. Muller-Gueudin. 'Uniform law of the logarithm for the local linear estimator of the conditional distribution function'. In: *C. R. Math. Acad. Sci. Paris* 348.17-18 (2010), pp. 1015–1019. DOI: 10.1016/j.crma.2010.08.003. URL: http://dx.doi.org/10.1016/j.crma.2010.08.003.

[45] J. Friedman, T. Hastie and R. Tibshirani. 'Sparse inverse covariance estimation with the graphical lasso'. In: *Biostatistics* 9.3 (2008), pp. 432–441.

[46] C. Giraud, S. Huet and N. Verzelen. 'Graph selection with GGMselect'. In: *Statistical applications in genetics and molecular biology* 11.3 (2012).

[47] T. Hansen and U. Zwick. 'Lower Bounds for Howard's Algorithm for Finding Minimum Mean-Cost Cycles'. In: *ISAAC (1)*. 2010, pp. 415–426.

[48] S. Herrmann and P. Vallois. 'From persistent random walk to the telegraph noise'. In: *Stoch. Dyn.* 10.2 (2010), pp. 161–196. DOI: 10.1142/S0219493710002905. URL: http://dx.doi.org/10.1142/S0219493710002905.

[49] J. Hu, W.-C. Wu and S. Sastry. 'Modeling subtilin production in bacillus subtilis using stochastic hybrid systems'. In: *Hybrid Systems: Computation and Control*. Springer, 2004, pp. 417–431.

[50] R. Keinj, T. Bastogne and P. Vallois. 'Multinomial model-based formulations of TCP and NTCP for radiotherapy treatment planning'. In: *Journal of Theoretical Biology* 279.1 (June 2011), pp. 55–62. DOI: 10.1016/j.jtbi.2011.03.025. URL: http://hal.inria.fr/hal-00588935/en.

[51] R. Koenker. *Quantile regression*. 38. Cambridge university press, 2005.

[52] Y. A. Kutoyants. *Statistical inference for ergodic diffusion processes*. Springer Series in Statistics. London: Springer-Verlag London Ltd., 2004, pp. xiv+481.

[53] C. Lacaux. 'Real Harmonizable Multifractional Lévy Motions'. In: *Ann. Inst. Poincaré*. 40.3 (2004), pp. 259–277.

[54] 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.

[55] B. Lesner and B. Scherrer. 'Non-Stationary Approximate Modified Policy Iteration'. In: *ICML 2015*. Lille, France, July 2015. URL: https://hal.inria.fr/hal-01186664.

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

[57] N. Meinshausen and P. Bühlmann. 'High-dimensional graphs and variable selection with the lasso'. In: *The Annals of Statistics* (2006), pp. 1436–1462.

[58] J.-M. Monnez. 'Approximation stochastique en analyse factorielle multiple'. In: *Ann. I.S.U.P.* 50.3 (2006), pp. 27–45.

[59] J.-M. Monnez. 'Convergence d'un processus d'approximation stochastique en analyse factorielle'. In: *Publ. Inst. Statist. Univ. Paris* 38.1 (1994), pp. 37–55.

[60] J.-M. Monnez. 'Stochastic approximation of the factors of a generalized canonical correlation analysis'. In: *Statist. Probab. Lett.* 78.14 (2008), pp. 2210–2216. DOI: 10.1016/j.spl.2008.01.088. URL: http://dx.doi.org/10.1016/j.spl.2008.01.088.

[61] E. Nadaraya. 'On non-parametric estimates of density functions and regression curves'. In: *Theory of Probability & Its Applications* 10.1 (1965), pp. 186–190.

[62] I. Post and Y. Ye. *The simplex method is strongly polynomial for deterministic Markov decision processes*. Tech. rep. arXiv:1208.5083v2, 2012.

[63] M. Puterman. *Markov Decision Processes*. Wiley, New York, 1994.

[64] B. Roynette, P. Vallois and M. Yor. 'Brownian penalisations related to excursion lengths, VII'. In: *Annales de l'IHP Probabilités et statistiques*. Vol. 45. 2. 2009, pp. 421–452.

[65] F. Russo and P. Vallois. 'Elements of stochastic calculus via regularization'. In: *Séminaire de Probabilités XL*. Vol. 1899. Lecture Notes in Math. Berlin: Springer, 2007, pp. 147–185. DOI: 10.1007/978-3-540-71189-6_7. URL: http://dx.doi.org/10.1007/978-3-540-71189-6%5C_7.

[66] F. Russo and P. Vallois. 'Stochastic calculus with respect to continuous finite quadratic variation processes'. In: *Stochastics: An International Journal of Probability and Stochastic Processes* 70.1-2 (2000), pp. 1–40.

[67] B. Scherrer. 'Approximate Policy Iteration Schemes: A Comparison'. In: *ICML - 31st International Conference on Machine Learning - 2014*. Pékin, China, June 2014. URL: https://hal.inria.fr/hal-00989982.

[68] B. Scherrer. 'Improved and Generalized Upper Bounds on the Complexity of Policy Iteration'. In: *Mathematics of Operations Research* (Feb. 2016). Markov decision processes ; Dynamic Programming ; Analysis of Algorithms. DOI: 10.1287/moor.2015.0753. URL: https://hal.inria.fr/hal-00829532.

[69] B. Scherrer. 'Performance Bounds for Lambda Policy Iteration and Application to the Game of Tetris'. In: *Journal of Machine Learning Research* 14 (Jan. 2013), pp. 1175–1221. URL: https://hal.inria.fr/hal-00759102.

[70] B. Scherrer, M. Ghavamzadeh, V. Gabillon, B. Lesner and M. Geist. 'Approximate Modified Policy Iteration and its Application to the Game of Tetris'. In: *Journal of Machine Learning Research* 16 (2015). A paraître, pp. 1629–1676. URL: https://hal.inria.fr/hal-01091341.

[71] B. Scherrer and B. Lesner. 'On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes'. In: *NIPS 2012 - Neural Information Processing Systems*. South Lake Tahoe, United States, Dec. 2012. URL: https://hal.inria.fr/hal-00758809.

[72] P. Vallois. 'The range of a simple random walk on Z'. In: *Advances in applied probability* (1996), pp. 1014–1033.

[73] P. Vallois and C. S. Tapiero. 'Memory-based persistence in a counting random walk process'. In: *Phys. A.* 386.1 (2007), pp. 303–307. DOI: 10.1016/j.physa.2007.08.027. URL: http://dx.doi.org/10.1016/j.physa.2007.08.027.

[74] N. Villa-Vialaneix. *An introduction to network inference and mining*. http://wikistat.fr/. (consulté le 22/07/2015). 2015. URL: http://www.nathalievilla.org/doc/pdf//wikistat-network_compiled.pdf.

[75] Y. Ye. 'The Simplex and Policy-Iteration Methods Are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate'. In: *Math. Oper. Res.* 36.4 (2011), pp. 593–603.