IN PARTNERSHIP WITH:
Ecole Polytechnique

Activity Report 2019

Project-Team XPOP

Statistical modelling for life sciences

IN COLLABORATION WITH: Centre de Mathématiques Appliquées (CMAP)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Modeling and Control for Life Sciences

Table of contents

Project-Team XPOP

Creation of the Team: 2016 January 01, updated into Project-Team: 2017 July 01

Keywords:

Computer Science and Digital Science:

A3.1.1. - Modeling, representation
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
A3.4.5. - Bayesian methods
A3.4.6. - Neural networks
A3.4.7. - Kernel methods
A3.4.8. - Deep learning
A5.9.2. - Estimation, modeling
A6.1.1. - Continuous Modeling (PDE, ODE)
A6.2.2. - Numerical probability
A6.2.3. - Probabilistic methods
A6.2.4. - Statistical methods
A6.3.3. - Data processing
A6.3.5. - Uncertainty Quantification

Other Research Topics and Application Domains:

B1.1.4. - Genetics and genomics
B1.1.7. - Bioinformatics
B1.1.10. - Systems and synthetic biology
B2.2.3. - Cancer
B2.2.4. - Infectious diseases, Virology
B2.4.1. - Pharmaco kinetics and dynamics
B9.1.1. - E-learning, MOOC

1. Team, Visitors, External Collaborators

Research Scientists

Marc Lavielle [Team leader, Inria, Senior Researcher]
Erwan Le Pennec [École polytechnique, Researcher]
Laetitia Le [APHP - Ecole polytechnique, Researcher, until Oct 2019]

Faculty Members

Julie Josse [École polytechnique, Associate Professor]
Eric Moulines [École polytechnique, Professor]

Post-Doctoral Fellows

Angie Pineda [École polytechnique, Post-Doctoral Fellow]
Tom Rohmer [Inria, Post-Doctoral Fellow, until Aug 2019]

PhD Students

Nicolas Brosse [École polytechnique, PhD Student, until Aug 2019]
Wei Jiang [École polytechnique, PhD Student]
Mohammed Karimi [Inria, PhD Student, until Sep 2019]
Genevieve Robin [École polytechnique, PhD Student, until Aug 2019]
Marine Zulian [Dassault Systemes, PhD Student, granted by CIFRE]

Technical staff

Yao Xu [Inria, Engineer, until Jun 2019]

Administrative Assistants

Ines Dumontier [Inria, until May 2019]
Hanadi Dib [Inria, from May 2019]

Visiting Scientist

Ricardo Rios [Universidad Central de Venezuela, Sep 2019]

2. Overall Objectives

2.1. Developing sound, useful and usable methods

The main objective of XPOP is to develop new sound and rigorous methods for statistical modeling in the field of biology and life sciences. These methods for modeling include statistical methods of estimation, model diagnostics, model building and model selection as well as methods for numerical models (systems of ordinary and partial differential equations). Historically, the key area where these methods have been used is population pharmacokinetics. However, the framework is currently being extended to sophisticated numerical models in the contexts of viral dynamics, glucose-insulin processes, tumor growth, precision medicine, spectrometry, intracellular processes, etc.

Furthermore, an important aim of XPOP is to transfer the methods developed into software packages so that they can be used in everyday practice.

2.2. Combining numerical, statistical and stochastic components of a model

Mathematical models that characterize complex biological phenomena are defined by systems of ordinary differential equations when dealing with dynamical systems that evolve with respect to time, or by partial differential equations when there is a spatial component in the model. Also, it is sometimes useful to integrate a stochastic aspect into the dynamical system in order to model stochastic intra-individual variability.

In order to use such methods, we must deal with complex numerical difficulties, generally related to resolving the systems of differential equations. Furthermore, to be able to check the quality of a model (i.e. its descriptive and predictive performances), we require data. The statistical aspect of the model is thus critical in how it takes into account different sources of variability and uncertainty, especially when data come from several individuals and we are interested in characterizing the inter-subject variability. Here, the tools of reference are mixed-effects models.

Confronted with such complex modeling problems, one of the goals of XPOP is to show the importance of combining numerical, statistical and stochastic approaches.

2.3. Developing future standards

Linear mixed-effects models have been well-used in statistics for a long time. They are a classical approach, essentially relying on matrix calculations in Gaussian models. Whereas a solid theoretical base has been developed for such models, *nonlinear* mixed-effects models (NLMEM) have received much less attention in the statistics community, even though they have been applied to many domains of interest. It has thus been the users of these models, such as pharmacometricians, who have taken them and developed methods, without really looking to develop a clean theoretical framework or understand the mathematical properties of the methods. This is why a standard estimation method in NLMEM is to linearize the model, and few people have been interested in understanding the properties of estimators obtained in this way.

Statisticians and pharmacometricians frequently realize the need to create bridges between these two communities. We are entirely convinced that this requires the development of new standards for population modeling that can be widely accepted by these various communities. These standards include the language used for encoding a model, the approach for representing a model and the methods for using it:

- **The approach.** Our approach consists in seeing a model as hierarchical, represented by a joint probability distribution. This joint distribution can be decomposed into a product of conditional distributions, each associated with a submodel (model for observations, individual parameters, etc.). Tasks required of the modeler are thus related to these probability distributions.
- **The methods.** Many tests have shown that algorithms implemented in MONOLIX are the most reliable, all the while being extremely fast. In fact, these algorithms are precisely described and published in well known statistical journals. In particular, the SAEM algorithm, used for calculating the maximum likelihood estimation of population parameters, has shown its worth in numerous situations. Its mathematical convergence has also been proven under quite general hypotheses.
- **The language.** Mlxtran is used by MONOLIX and other modeling tools and is today by far the most advanced language for representing models. Initially developed for representing pharmacometric models, its syntax also allows it to easily code dynamical systems defined by a system of ODEs, and statistical models involving continuous, discrete and survival variables. This flexibility is a true advantage both for numerical modelers and statisticians.

3. Research Program

3.1. Scientific positioning

"Interfaces" is the defining characteristic of XPOP:

The interface between statistics, probability and numerical methods. Mathematical modelling of complex biological phenomena require to combine numerical, stochastic and statistical approaches. The CMAP is therefore the right place to be for positioning the team at the interface between several mathematical disciplines.

The interface between mathematics and the life sciences. The goal of XPOP is to bring the right answers to the right questions. These answers are mathematical tools (statistics, numerical methods, etc.), whereas the questions come from the life sciences (pharmacology, medicine, biology, etc.). This is why the point of XPOP is not to take part in mathematical projects only, but also pluridisciplinary ones.

The interface between mathematics and software development. The development of new methods is the main activity of XPOP. However, new methods are only useful if they end up being implemented in a software tool. On one hand, a strong partnership with Lixoft (the spin-off company who continue developing MONOLIX) allows us to maintaining this positioning. On the other hand, several members of the team are very active in the R community and develop widely used packages.

3.2. The mixed-effects models

Mixed-effects models are statistical models with both fixed effects and random effects. They are well-adapted to situations where repeated measurements are made on the same individual/statistical unit.

Consider first a single subject i of the population. Let $y_i = (y_{ij}, 1 \leq j \leq n_i)$ be the vector of observations for this subject. The model that describes the observations y_i is assumed to be a parametric probabilistic model: let $p_Y(y_i; \psi_i)$ be the probability distribution of y_i, where ψ_i is a vector of parameters.

In a population framework, the vector of parameters ψ_i is assumed to be drawn from a population distribution $p_\Psi(\psi_i; \theta)$ where θ is a vector of population parameters.

Then, the probabilistic model is the joint probability distribution

$$p(y_i, \psi_i; \theta) = p_Y(y_i|\psi_i)p_\Psi(\psi_i; \theta) \tag{1}$$

To define a model thus consists in defining precisely these two terms.

In most applications, the observed data y_i are continuous longitudinal data. We then assume the following representation for y_i:

$$y_{ij} = f(t_{ij}, \psi_i) + g(t_{ij}, \psi_i)\varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i. \tag{2}$$

Here, y_{ij} is the observation obtained from subject i at time t_{ij}. The residual errors (ε_{ij}) are assumed to be standardized random variables (mean zero and variance 1). The residual error model is represented by function g in model (2).

Function f is usually the solution to a system of ordinary differential equations (pharmacokinetic/pharmacodynamic models, etc.) or a system of partial differential equations (tumor growth, respiratory system, etc.). This component is a fundamental component of the model since it defines the prediction of the observed kinetics for a given set of parameters.

The vector of individual parameters ψ_i is usually function of a vector of population parameters ψ_{pop}, a vector of random effects $\eta_i \sim \mathcal{N}(0, \Omega)$, a vector of individual covariates c_i (weight, age, gender, ...) and some fixed effects β.

The joint model of y and ψ depends then on a vector of parameters $\theta = (\psi_{\text{pop}}, \beta, \Omega)$.

3.3. Computational Statistical Methods

Central to modern statistics is the use of probabilistic models. To relate these models to data requires the ability to calculate the probability of the observed data: the likelihood function, which is central to most statistical methods and provides a principled framework to handle uncertainty.

The emergence of computational statistics as a collection of powerful and general methodologies for carrying out likelihood-based inference made complex models with non-standard data accessible to likelihood, including hierarchical models, models with intricate latent structure, and missing data.

In particular, algorithms previously developed by POPIX for mixed effects models, and today implemented in several software tools (especially MONOLIX) are part of these methods:

- the adaptive Metropolis-Hastings algorithm allows one to sample from the conditional distribution of the individual parameters $p(\psi_i|y_i; c_i, \theta)$,
- the SAEM algorithm is used to maximize the observed likelihood $\mathcal{L}(\theta; y) = p(y; \theta)$,
- Importance Sampling Monte Carlo simulations provide an accurate estimation of the observed log-likelihood $\log(\mathcal{L}(\theta; y))$.

Computational statistics is an area which remains extremely active today. Recently, one can notice that the incentive for further improvements and innovation comes mainly from three broad directions: the high dimensional challenge, the quest for adaptive procedures that can eliminate the cumbersome process of tuning "by hand" the settings of the algorithms and the need for flexible theoretical support, arguably required by all recent developments as well as many of the traditional MCMC algorithms that are widely used in practice.

Working in these three directions is a clear objective for XPOP.

3.4. Markov Chain Monte Carlo algorithms

While these Monte Carlo algorithms have turned into standard tools over the past decade, they still face difficulties in handling less regular problems such as those involved in deriving inference for high-dimensional models. One of the main problems encountered when using MCMC in this challenging settings is that it is difficult to design a Markov chain that efficiently samples the state space of interest.

The Metropolis-adjusted Langevin algorithm (MALA) is a Markov chain Monte Carlo (MCMC) method for obtaining random samples from a probability distribution for which direct sampling is difficult. As the name suggests, MALA uses a combination of two mechanisms to generate the states of a random walk that has the target probability distribution as an invariant measure:

1. new states are proposed using Langevin dynamics, which use evaluations of the gradient of the target probability density function;
2. these proposals are accepted or rejected using the Metropolis-Hastings algorithm, which uses evaluations of the target probability density (but not its gradient).

Informally, the Langevin dynamics drives the random walk towards regions of high probability in the manner of a gradient flow, while the Metropolis-Hastings accept/reject mechanism improves the mixing and convergence properties of this random walk.

Several extensions of MALA have been proposed recently by several authors, including fMALA (fast MALA), AMALA (anisotropic MALA), MMALA (manifold MALA), position-dependent MALA (PMALA), ...

MALA and these extensions have demonstrated to represent very efficient alternative for sampling from high dimensional distributions. We therefore need to adapt these methods to general mixed effects models.

3.5. Parameter estimation

The Stochastic Approximation Expectation Maximization (SAEM) algorithm has shown to be extremely efficient for maximum likelihood estimation in incomplete data models, and particularly in mixed effects models for estimating the population parameters. However, there are several practical situations for which extensions of SAEM are still needed:

High dimensional model: a complex physiological model may have a large number of parameters (in the order of 100). Then several problems arise:

- when most of these parameters are associated with random effects, the MCMC algorithm should be able to sample, for each of the N individuals, parameters from a high dimensional distribution. Efficient MCMC methods for high dimensions are then required.
- Practical identifiability of the model is not ensured with a limited amount of data. In other words, we cannot expect to be able to properly estimate all the parameters of the model, including the fixed effects and the variance-covariance matrix of the random effects. Then, some random effects should be removed, assuming that some parameters do not vary in the population. It may also be necessary to fix the value of some parameters (using values from the literature for instance). The strategy to decide which parameters should be fixed and which random effects should be removed remains totally empirical. XPOP aims to develop a procedure that will help the modeller to take such decisions.

Large number of covariates: the covariate model aims to explain part of the inter-patient variability of some parameters. Classical methods for covariate model building are based on comparisons with respect to some criteria, usually derived from the likelihood (AIC, BIC), or some statistical test (Wald test, LRT, etc.). In other words, the modelling procedure requires two steps: first, all possible models are fitted using some estimation procedure (e.g. the SAEM algorithm) and the likelihood of each model is computed using a numerical integration procedure (e.g. Monte Carlo Importance Sampling); then, a model selection procedure chooses the "best" covariate model. Such a strategy is only possible with a reduced number of covariates, i.e., with a "small" number of models to fit and compare.

As an alternative, we are thinking about a Bayesian approach which consists of estimating simultaneously the covariate model and the parameters of the model in a single run. An (informative or uninformative) prior is defined for each model by defining a prior probability for each covariate to be included in the model. In other words, we extend the probabilistic model by introducing binary variables that indicate the presence or absence of each covariate in the model. Then, the model selection procedure consists of estimating and maximizing the conditional distribution of this sequence of binary variables. Furthermore, a probability can be associated to any of the possible covariate models.

This conditional distribution can be estimated using an MCMC procedure combined with the SAEM algorithm for estimating the population parameters of the model. In practice, such an approach can only deal with a limited number of covariates since the dimension of the probability space to explore increases exponentially with the number of covariates. Consequently, we would like to have methods able to find a small number of variables (from a large starting set) that influence certain parameters in populations of individuals. That means that, instead of estimating the conditional distribution of all the covariate models as described above, the algorithm should focus on the most likely ones.

Fixed parameters: it is quite frequent that some individual parameters of the model have no random component and are purely fixed effects. Then, the model may not belong to the exponential family anymore and the original version of SAEM cannot be used as it is. Several extensions exist:

- introduce random effects with decreasing variances for these parameters,
- introduce a prior distribution for these fixed effects,
- apply the stochastic approximation directly on the sequence of estimated parameters, instead of the sufficient statistics of the model.

None of these methods always work correctly. Furthermore, what are the pros and cons of these methods is not clear at all. Then, developing a robust methodology for such model is necessary.

Convergence toward the global maximum of the likelihood: convergence of SAEM can strongly depend on thie initial guess when the observed likelihood has several local maxima. A kind of simulated annealing version of SAEM was previously developed and implemented in MONOLIX. The method works quite well in most situations but there is no theoretical justification and choosing the settings of this algorithm (i.e. how the temperature decreases during the iterations) remains empirical. A precise analysis of the algorithm could be very useful to better understand why it "works" in practice and how to optimize it.

Convergence diagnostic: Convergence of SAEM was theoretically demonstrated under very general hypothesis. Such result is important but of little interest in practice at the time to use SAEM in a finite amount of time, i.e. in a finite number of iterations. Some qualitative and quantitative criteria should be defined in order to both optimize the settings of the algorithm, detect a poor convergence of SAEM and evaluate the quality of the results in order to avoid using them unwisely.

3.6. Model building

Defining an optimal strategy for model building is far from easy because a model is the assembled product of numerous components that need to been evaluated and perhaps improved: the structural model, residual error model, covariate model, covariance model, etc.

How to proceed so as to obtain the best possible combination of these components? There is no magic recipe but an effort will be made to provide some qualitative and quantitative criteria in order to help the modeller for building his model.

The strategy to take will mainly depend on the time we can dedicate to building the model and the time required for running it. For relatively simple models for which parameter estimation is fast, it is possible to fit many models and compare them. This can also be done if we have powerful computing facilities available (e.g., a cluster) allowing large numbers of simultaneous runs.

However, if we are working on a standard laptop or desktop computer, model building is a sequential process in which a new model is tested at each step. If the model is complex and requires significant computation time (e.g., when involving systems of ODEs), we are constrained to limit the number of models we can test in a reasonable time period. In this context, it also becomes important to carefully choose the tasks to run at each step.

3.7. Model evaluation

Diagnostic tools are recognized as an essential method for model assessment in the process of model building. Indeed, the modeler needs to confront "his" model with the experimental data before concluding that this model is able to reproduce the data and before using it for any purpose, such as prediction or simulation for instance.

The objective of a diagnostic tool is twofold: first we want to check if the assumptions made on the model are valid or not ; then, if some assumptions are rejected, we want to get some guidance on how to improve the model.

As is the usual case in statistics, it is not because this "final" model has not been rejected that it is necessarily the "true" one. All that we can say is that the experimental data does not allow us to reject it. It is merely one of perhaps many models that cannot be rejected.

Model diagnostic tools are for the most part graphical, i.e., visual; we "see" when something is not right between a chosen model and the data it is hypothesized to describe. These diagnostic plots are usually based on the empirical Bayes estimates (EBEs) of the individual parameters and EBEs of the random effects: scatterplots of individual parameters versus covariates to detect some possible relationship, scatterplots of pairs of random effects to detect some possible correlation between random effects, plot of the empirical distribution of the random effects (boxplot, histogram,...) to check if they are normally distributed, ...

The use of EBEs for diagnostic plots and statistical tests is efficient with rich data, i.e. when a significant amount of information is available in the data for recovering accurately all the individual parameters. On the contrary, tests and plots can be misleading when the estimates of the individual parameters are greatly shrunk.

We propose to develop new approaches for diagnosing mixed effects models in a general context and derive formal and unbiased statistical tests for testing separately each feature of the model.

3.8. Missing data

The ability to easily collect and gather a large amount of data from different sources can be seen as an opportunity to better understand many processes. It has already led to breakthroughs in several application areas. However, due to the wide heterogeneity of measurements and objectives, these large databases often exhibit an extraordinary high number of missing values. Hence, in addition to scientific questions, such data also present some important methodological and technical challenges for data analyst.

Missing values occur for a variety of reasons: machines that fail, survey participants who do not answer certain questions, destroyed or lost data, dead animals, damaged plants, etc. Missing values are problematic since most statistical methods can not be applied directly on a incomplete data. Many progress have been made to properly handle missing values. However, there are still many challenges that need to be addressed in the future, that are crucial for the users.

- State of arts methods often consider the case of continuous or categorical data whereas real data are very often mixed. The idea is to develop a multiple imputation method based on a specific principal component analysis (PCA) for mixed data. Indeed, PCA has been used with success to predict (impute) the missing values. A very appealing property is the ability of the method to handle very large matrices with large amount of missing entries.
- The asymptotic regime underlying modern data is not any more to consider that the sample size increases but that both number of observations and number of variables are very large. In practice first experiments showed that the coverage properties of confidence areas based on the classical methods to estimate variance with missing values varied widely. The asymptotic method and the bootstrap do well in low-noise setting, but can fail when the noise level gets high or when the number of variables is much greater than the number of rows. On the other hand, the jackknife has good coverage properties for large noisy examples but requires a minimum number of variables to be stable enough.
- Inference with missing values is usually performed under the assumption of "Missing at Random" (MAR) values which means that the probability that a value is missing may depend on the observed data but does not depend on the missing value itself. In real data and in particular in data coming from clinical studies, both "Missing Non at Random" (MNAR) and MAR values occur. Taking into account in a proper way both types of missing values is extremely challenging but is worth investigating since the applications are extremely broad.

It is important to stress that missing data models are part of the general incomplete data models addressed by XPOP. Indeed, models with latent variables (i.e. non observed variables such as random effects in a mixed effects model), models with censored data (e.g. data below some limit of quantification) or models with dropout mechanism (e.g. when a subject in a clinical trial fails to continue in the study) can be seen as missing data models.

4. Application Domains

4.1. Surface Enhanced Raman Spectroscopy

(joint project with HEGP, AP-HP, and Lip(Sys)2, Université Paris-Saclay)

The objective of this work is to evaluate the feasibility of an evolving technique, surface enhanced Raman spectroscopy (SERS) for the analysis of cytotoxic drug concentration. This technique using silver nanoparticles was applied for quantitative analysis of 5-fluorouracil, one of the most widely used molecules in oncology [8].

In view of the high spectral variability observed between the various repetitions of the experiment, and the observed nonlinear interaction between signal concentration and intensity, nonlinear regression methods that take these variabilities into account have been developed.

4.2. Management of severe trauma

(joint project with the Traumabase group, AP-HP)

Major trauma is defined as any injury that endangers the life or the functional integrity of a person. It has been shown that management of major trauma based on standardized and protocol based care improves prognosis of patients especially for the two main causes of death in major trauma i.e., hemorrhage and traumatic brain injury.

However, evidence shows that patient management even in mature trauma systems often exceeds acceptable time frames, and despite existing guidelines deviations from protocol-based care are often observed. These deviations lead to a high variability in care and are associated with bad outcome such as inadequate hemorrhage control or delayed transfusion. Two main factors explain these observations. First, decision-making in trauma care is particularly demanding, because it requires rapid and complex decisions under time pressure in a very dynamic and multi-player environment characterized by high levels of uncertainty and stress. Second, being a complex and multiplayer process, trauma care is affected by fragmentation. Fragmentation is often the result of loss or deformation of information.

This disruptive influence prevents providers to engage with each other and commit to the care process.In order to respond to this challenge, our program has set the ambitious goal to develop a trauma decision support tool, the TraumaMatrix. The program aims to provide an integrative decision support and information management solution to clinicians for the first 24 hours of major trauma management. This program is divided into three steps.

Based on a detailed and high quality trauma database, Step 1 consists in developing the mathematical tools and models to predict trauma specific outcomes and decisions. This step raises considerable scientific and methodological challenges.

Step 2 will use these methods to apply them to develop in close cooperation with trauma care experts the decision support tool and develop a user friendly and ergonomic interface to be used by clinicians.

Step 3 will further develop the tool and interface and test in real-time its impact on clinician decision making and patient outcome.

4.3. Precision medicine and pharmacogenomics

(joint project with Dassault Systèmes)

Pharmacogenomics involves using an individual's genome to determine whether or not a particular therapy, or dose of therapy, will be effective. Indeed, people's reaction to a given drug depends on their physiological state and environmental factors, but also to their individual genetic make-up.

Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. While some advances in precision medicine have been made, the practice is not currently in use for most diseases.

Currently, in the traditional population approach, inter-individual variability in the reaction to drugs is modeled using covariates such as weight, age, sex, ethnic origin, etc. Genetic polymorphisms susceptible to modify pharmacokinetic or pharmacodynamic parameters are much harder to include, especially as there are millions of possible polymorphisms (and thus covariates) per patient.

The challenge is to determine which genetic covariates are associated to some PKPD parameters and/or implicated in patient responses to a given drug.

Another problem encountered is the dependence of genes, as indeed, gene expression is a highly regulated process. In cases where the explanatory variables (genomic variants) are correlated, Lasso-type methods for model selection are thwarted.

There is therefore a clear need for new methods and algorithms for the estimation, validation and selection of mixed effects models adapted to the problems of genomic medicine.

A target application of this project concerns the lung cancer.

EGFR (Epidermal Growth Factor Receptor) is a cell surface protein that binds to epidermal growth factor. We know that deregulation of the downstream signaling pathway of EGFR is involved in the development of lung cancers and several gene mutations responsible for this deregulation are known.

Our objective is to identify the variants responsible for the disruption of this pathway using a modelling approach. The data that should be available for developing such model are ERK (Extracellular signal–regulated kinases) phosphorylation time series, obtained from different genetic profiles.

The model that we aim to develop will describe the relationship between the parameters of the pathway and the genomic covariates, i.e. the genetic profile. Variants related to the pathway include: variants that modify the affinity binding of ligands to receptors, variants that modify the total amount of protein, variants that affect the catalytic site,...

4.4. Oncology

(joint project with the Biochemistry lab of Ecole Polytechnique and Institut Curie)

In cancer, the most dreadful event is the formation of metastases that disseminate tumor cells throughout the organism. Cutaneous melanoma is a cancer, where the primary tumor can easily be removed by surgery. However, this cancer is of poor prognosis; because melanomas metastasize often and rapidly. Many melanomas arise from excessive exposure to mutagenic UV from the sun or sunbeds. As a consequence, the mutational burden of melanomas is generally high

RAC1 encodes a small GTPase that induces cell cycle progression and migration of melanoblasts during embryonic development. Patients with the recurrent P29S mutation of RAC1 have 3-fold increased odds at having regional lymph nodes invaded at the time of diagnosis. RAC1 is unlikely to be a good therapeutic target, since a potential inhibitor that would block its catalytic activity, would also lock it into the active GTP-bound state. This project thus investigates the possibility of targeting the signaling pathway downstream of RAC1.

XPOP is mainly involved in Task 1 of the project: *Identifying deregulations and mutations of the ARP2/3 pathway in melanoma patients*.

Association of over-expression or down-regulation of each marker with poor prognosis in terms of invasion of regional lymph nodes, metastases and survival, will be examined using classical univariate and multivariate analysis. We will then develop specific statistical models for survival analysis in order to associate prognosis factors to each composition of complexes. Indeed, one has to implement the further constraint that each subunit has to be contributed by one of several paralogous subunits. An original method previously developed by XPOP has already been successfully applied to WAVE complex data in breast cancer.

The developed models will be rendered user-friendly though a dedicated Rsoftware package.

This project can represent a significant step forward in precision medicine of the cutaneous melanoma.

4.5. Anesthesiology

(joint project with AP-HP Lariboisière and M3DISIM)

Two hundred million general anaesthesias are performed worldwide every year. Low blood pressure during anaesthesia is common and has been identified as a major factor in morbidity and mortality. These events require great reactivity in order to correct them as quickly as possible and impose constraints of reliability and reactivity to monitoring and treatment.

Recently, studies have demonstrated the usefulness of noradrelanine in preventing and treating intraoperative hypotension. The handling of this drug requires great vigilance with regard to the correct dosage. Currently, these drugs are administered manually by the healthcare staff in bolus and/or continuous infusion. This represents a heavy workload and suffers from a great deal of variability in order to find the right dosage for the desired effect on blood pressure.

The objective of this project is to automate the administration of noradrelanine with a closed-loop system that makes it possible to control the treatment in real time to an instantaneous blood pressure measurement.

4.6. Intracellular processes

(joint project with the InBio and IBIS inria teams and the MSC lab, UMR 7057)

Significant cell-to-cell heterogeneity is ubiquitously-observed in isogenic cell populations. Cells respond differently to a same stimulation. For example, accounting for such heterogeneity is essential to quantitatively understand why some bacteria survive antibiotic treatments, some cancer cells escape drug-induced suicide, stem cell do not differentiate, or some cells are not infected by pathogens.

The origins of the variability of biological processes and phenotypes are multifarious. Indeed, the observed heterogeneity of cell responses to a common stimulus can originate from differences in cell phenotypes (age, cell size, ribosome and transcription factor concentrations, etc), from spatio-temporal variations of the cell environments and from the intrinsic randomness of biochemical reactions. From systems and synthetic biology perspectives, understanding the exact contributions of these different sources of heterogeneity on the variability of cell responses is a central question.

The main ambition of this project is to propose a paradigm change in the quantitative modelling of cellular processes by shifting from mean-cell models to single-cell and population models. The main contribution of XPOP focuses on methodological developments for mixed-effects model identification in the context of growing cell populations [9].

- Mixed-effects models usually consider an homogeneous population of independent individuals. This assumption does not hold when the population of cells (i.e. the statistical individuals) consists of several generations of dividing cells. We then need to account for inheritance of single-cell parameters in this population. More precisely, the problem is to attribute the new state and parameter values to newborn cells given (the current estimated values for) the mother.
- The mixed-effects modelling framework corresponds to a strong assumption: differences between cells are static in time (ie, cell-specific parameters have fixed values). However, it is likely that for any given cell, ribosome levels slowly vary across time, since like any other protein, ribosomes are produced in a stochastic manner. We will therefore extend our modelling framework so as to account for the possible random fluctuations of parameter values in individual cells. Extensions based on stochastic differential equations will be investigated.
- Identifiability is a fundamental prerequisite for model identification and is also closely connected to optimal experimental design. We will derive criteria for theoretical identifiability, in which different parameter values lead to non-identical probability distributions, and for structural identifiability, which concerns the algebraic properties of the structural model, i.e. the ODE system. We will then address the problem of practical identifiability, whereby the model may be theoretically identifiable but the design of the experiment may make parameter estimation difficult and imprecise. An interesting problem is whether accounting for lineage effects can help practical identifiability of the parameters of the individuals in presence of measurement and biological noise.

4.7. Population pharmacometrics

(joint project with Lixoft)

Pharmacometrics involves the analysis and interpretation of data produced in pre-clinical and clinical trials. Population pharmacokinetics studies the variability in drug exposure for clinically safe and effective doses by focusing on identification of patient characteristics which significantly affect or are highly correlated with this variability. Disease progress modeling uses mathematical models to describe, explain, investigate and predict the changes in disease status as a function of time. A disease progress model incorporates functions describing natural disease progression and drug action.

The model based drug development (MBDD) approach establishes quantitative targets for each development step and optimizes the design of each study to meet the target. Optimizing study design requires simulations, which in turn require models. In order to arrive at a meaningful design, mechanisms need to be understood and correctly represented in the mathematical model. Furthermore, the model has to be predictive for future studies. This requirement precludes all purely empirical modeling; instead, models have to be mechanistic.

In particular, physiologically based pharmacokinetic models attempt to mathematically transcribe anatomical, physiological, physical, and chemical descriptions of phenomena involved in the ADME (Absorption - Distribution - Metabolism - Elimination) processes. A system of ordinary differential equations for the quantity of substance in each compartment involves parameters representing blood flow, pulmonary ventilation rate, organ volume, etc.

The ability to describe variability in pharmacometrics model is essential. The nonlinear mixed-effects modeling approach does this by combining the structural model component (the ODE system) with a statistical model, describing the distribution of the parameters between subjects and within subjects, as well as quantifying the unexplained or residual variability within subjects.

The objective of XPOP is to develop new methods for models defined by a very large ODE system, a large number of parameters and a large number of covariates. Contributions of XPOP in this domain are mainly methodological and there is no privileged therapeutic application at this stage [7], [21], [14].

However, it is expected that these new methods will be implemented in software tools, including MONOLIX and Rpackages for practical use.

4.8. Mass spectrometry

(joint project with the Molecular Chemistry Laboratory, LCM, of Ecole Polytechnique)

One of the main recent developments in analytical chemistry is the rapid democratization of high-resolution mass spectrometers. These instruments produce extremely complex mass spectra, which can include several hundred thousand ions when analyzing complex samples. The analysis of complex matrices (biological, agri-food, cosmetic, pharmaceutical, environmental, etc.) is precisely one of the major analytical challenges of this new century. Academic and industrial researchers are particularly interested in trying to quickly and effectively establish the chemical consequences of an event on a complex matrix. This may include, for example, searching for pesticide degradation products and metabolites in fruits and vegetables, photoproducts of active ingredients in a cosmetic emulsion exposed to UV rays or chlorination products of biocides in hospital effluents. The main difficulty of this type of analysis is based on the high spatial and temporal variability of the samples, which is in addition to the experimental uncertainties inherent in any measurement and requires a large number of samples and analyses to be carried out and computerized data processing (up to 16 million per mass spectrum).

A collaboration between XPOP and the Molecular Chemistry Laboratory (LCM) of the Ecole Polytechnique began in 2018. Our objective is to develop new methods for the statistical analysis of mass spectrometry data.

These methods are implemented in the SPIX software.

5. Highlights of the Year

5.1. Highlights of the Year

Version 2.0 of the SPIX software was available in June 2019.

5.1.1. Awards

Geneviève Robin was awarded: “Prix L’Oréal-UNESCO Pour les Femmes et la Science (Jeunes Talents France 2019)”

6. New Software and Platforms

6.1. mlxR

KEYWORDS: Simulation - Data visualization - Clinical trial simulator

FUNCTIONAL DESCRIPTION: The models are encoded using the model coding language 'Mlxtran', automatically converted into C++ codes, compiled on the fly and linked to R using the 'Rcpp' package. That allows one to implement very easily complex ODE-based models and complex statistical models, including mixed effects models, for continuous, count, categorical, and time-to-event data.

- Contact: Marc Lavielle
- URL: http://simulx.webpopix.org/

6.2. Rsmlx

R speaks Monolix

KEYWORDS: Data modeling - Nonlinear mixed effects models - Statistical modeling

FUNCTIONAL DESCRIPTION: Among other tasks, 'Rsmlx' provides a powerful tool for automatic PK model building, performs statistical tests for model assessment, bootstrap simulation and likelihood profiling for computing confidence intervals. 'Rsmlx' also proposes several automatic covariate search methods for mixed effects models.

- Partner: Lixoft
- Contact: Marc Lavielle
- URL: http://rsmlx.webpopix.org/

6.3. SPIX

KEYWORDS: Data modeling - Mass spectrometry - Chemistry

FUNCTIONAL DESCRIPTION: SPIX allows you to - automatically identify, on the basis of statistical approaches, small but significant differences in spectra measured under different conditions, - model the kinetics of entities that evolve over time

- Authors: Marc Lavielle and Yao Xu
- Partner: Laboratoire de Chimie Moléculaire - Ecole Polytechnique
- Contact: Marc Lavielle
- URL: http://spix.webpopix.org/

7. New Results

7.1. Modelling inheritance and variability of kinetic gene expression parameters in microbial cells

Modern experimental technologies enable monitoring of gene expression dynamics in individual cells and quantification of its variability in isogenic microbial populations. Among the sources of this variability is the randomness that affects inheritance of gene expression factors at cell division. Known parental relationships among individually observed cells provide invaluable information for the characterization of this extrinsic source of gene expression noise. Despite this fact, most existing methods to infer stochastic gene expression models from single-cell data dedicate little attention to the reconstruction of mother-daughter inheritance dynamics. Starting from a transcription and translation model of gene expression, we proposed a stochastic model for the evolution of gene expression dynamics in a population of dividing cells. Based on this model, we developed a method for the direct quantification of inheritance and variability of kinetic gene expression parameters from single-cell gene expression and lineage data. We demonstrated that our approach provides unbiased estimates of mother-daughter inheritance parameters, whereas indirect approaches using lineage information only in the post-processing of individual-cell parameters underestimate inheritance. Finally, we have shown on yeast osmotic shock response data that daughter cell parameters are largely determined by the mother, thus confirming the relevance of our method for the correct assessment of the onset of gene expression variability and the study of the transmission of regulatory factors [9].

7.2. Main effects and interactions in mixed and incomplete data frames

A mixed data frame (MDF) is a table collecting categorical, numerical and count observations. The use of MDF is widespread in statistics and the applications are numerous from abundance data in ecology to recommender systems. In many cases, an MDF exhibits simultaneously main effects, such as row, column or group effects and interactions, for which a low-rank model has often been suggested. Although the literature on low-rank approximations is very substantial, with few exceptions, existing methods do not allow to incorporate main effects and interactions while providing statistical guarantees. We proposed a new method that fills this gap [11], [3].

7.3. Quantification of gemcitabine intravenous drugs

This aim of this study was to assess the ability of Raman spectroscopy to quantify antineoplastic drugs directly in the finished product in plastic bags using a handled Raman spectrometer. Gemcitabine diluted in 0.9% sodium chloride was analyzed at various concentrations ranging from 1 to 20mg/mL directly through plastic bags using a handled 785nm Raman spectrometer.In accordance with EMA guidelines, quantitative models were developed to predict gemcitabine concentration in bag using partial least squares (PLS) regression. In order to evaluate the transposability of the developed Raman method and the routine method (flow injection analysis with UV detection), independent samples were analyzed using both techniques. The impact of the plastic bag was also evaluated by analysis samples through two different bags.The best model was obtained after standard normal variates preprocessing (SNV) for 15 latent variables. This model presented an excellent correlation between predicted and theoretical concentration values (R^2 of 0.9938 from the calibration set), a low limit of quantification (LLOQ) of 3.68mg/mL and acceptable repeatability and intermediate precision lower than the expected acceptance limit of 5% over the entire concentration range tested (except for the average concentration of 5.73mg/mL). For the 48 preparations higher than the LLOQ, the Bland-Altman approach showed the interchangeability of the wo methods with a difference bias of 2%. Moreover, no significant difference of predicted concentrations between the two containers tested (p = 0.189) was observed.Despite some limitations for low concentrations, this study clearly shows promising results for real-time monitoring of gemcitabine infusion preparations without removing samples. The non-invasive nature of this method should ensure the correct dose before administration to patients and with heightened safety for operators [8].

7.4. Low-rank model with covariates for count data analysis

Count data are collected in many scientific and engineering tasks including image processing, single-cell RNA sequencing and ecological studies. Such data sets often contain missing values, for example because some ecological sites cannot be reached in a certain year. In addition, in many instances, side information is also available, for example covariates about ecological sites or species. Low-rank methods are popular to denoise and impute count data, and benefit from a substantial theoretical background. Extensions accounting for covariates have been proposed, but to the best of our knowledge their theoretical and empirical properties have not been thoroughly studied, and few softwares are available for practitioners. We propose a complete methodology called LORI (Low-Rank Interaction), including a Poisson model, an algorithm, and automatic selection of the regularization parameter, to analyze count tables with covariates. We also derive an upper bound on the estimation error. We provide a simulation study with synthetic data, revealing empirically that LORI improves on state of the art methods in terms of estimation and imputation of the missing values. We illustrate how the method can be interpreted through visual displays with the analysis of a well-know plant abundance data set, and show that the LORI outputs are consistent with known results. Finally we demonstrate the relevance of the methodology by analyzing a waterbirds abundance table from the French national agency for wildlife and hunting management (ONCFS). The method is available in the R package lori on the Comprehensive Archive Network (CRAN), [10].

7.5. Imputation and low-rank estimation with Missing Non At Random data

Missing values challenge data analysis because many supervised and unsupervised learning methods cannot be applied directly to incomplete data. Matrix completion based on low-rank assumptions are very powerful solution for dealing with missing values. However, existing methods do not consider the case of informative missing values which are widely encountered in practice. We propose matrix completion methods to recover Missing Not At Random (MNAR) data. Our first contribution is to suggest a model-based estimation strategy by modelling the missing mechanism distribution. An EM algorithm is then implemented, involving a Fast Iterative Soft-Thresholding Algorithm (FISTA). Our second contribution is to suggest a computationally efficient surrogate estimation by implicitly taking into account the joint distribution of the data and the missing mechanism: the data matrix is concatenated with the mask coding for the missing values ; a low-rank structure for exponential family is assumed on this new matrix, in order to encode links between variables and missing mechanisms. The methodology that has the great advantage of handling different missing value mechanisms is robust to model specification errors, [22].

7.6. A mathematical model to predict BNP levels in hemodialysis patients

Clinical interpretation of B-Type Natriuretic Peptide (BNP) levels in hemodialysis patients (HD) for fluid management remains elusive. We conducted a retrospective observational monocentric study. We built a mathematical model to predict BNP levels, using multiple linear regressions, [12].

7.7. Analysis of the global convergence of (fast) incremental EM methods

The EM algorithm is one of the most popular algorithm for inference in latent data models. The original formulation of the EM algorithm does not scale to large data set, because the whole data set is required at each iteration of the algorithm. To alleviate this problem, Neal and Hinton (1998) have proposed an incremental version of the EM (iEM) in which at each iteration the conditional expectation of the latent data (E-step) is updated only for a mini-batch of observations. Another approach has been proposed by Cappé and Moulines (2009) in which the E-step is replaced by a stochastic approximation step, closely related to stochastic gradient. In this study, we analyzed incremental and stochastic version of the EM algorithm in a common unifying framework. We also introduced a new version incremental version, inspired by the SAGA algorithm by Defazio et al. (2014). We established non-asymptotic convergence bounds for global convergence, [15].

7.8. Efficient Metropolis-Hastings sampling for nonlinear mixed effects models

The ability to generate samples of the random effects from their conditional distributions is fundamental for inference in mixed effects models. Random walk Metropolis is widely used to conduct such sampling, but such a method can converge slowly for high dimension problems, or when the joint structure of the distributions to sample is complex. We proposed a Metropolis-Hastings (MH) algorithm based on a multidimensional Gaussian proposal that takes into account the joint conditional distribution of the random effects and does not require any tuning, in contrast with more sophisticated samplers such as the Metropolis Adjusted Langevin Algorithm or the No-U-Turn Sampler that involve costly tuning runs or intensive computation. Indeed, this distribution is automatically obtained thanks to a Laplace approximation of the original model. We have shown that such approximation is equivalent to linearizing the model in the case of continuous data, [14], [2].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

Contract with Dassault Systèmes

Contract with Lixoft

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR

Mixed-Effects Models of Intracellular Processes: Methods, Tools and Applications (MEMIP)

Coordinator: Gregory Batt (InBio Inria team)

Other partners: InBio and IBIS Inria teams, Laboratoire Matière et Systèmes Complexes (UMR 7057; CNRS and Paris Diderot Univ.)

9.1.2. Institut National du Cancer (INCa)

Targeting Rac-dependent actin polymerization in cutaneous melanoma - Institut National du Cancer

Coordinator: Alexis Gautreau (Ecole Polytechnique)

Other partners: Laboratoire de Biochimie (Polytechnique), Institut Curie, INSERM.

9.2. International Initiatives

9.2.1. International Initiatives

SaSMoTiDep

Title: Statistical and Stochastic modeling for time-dependent data

International Partners (Institution - Laboratory - Researcher):

Universidad de Valparaiso (Chile) - Centro de Investigación y Modelamiento de Fenómenos Aleatorios Valparaíso (CIMFAV) - Cristian Meza Becerra

Universidad Nacional de Colombia (Colombia) - Department of Statistics - Viswanathan Arunachalam

Duration: 01/01/2018 - 31/12/2019

Start year: 2018

See also: https://sasmotidep.uv.cl

In many applications, multiple measurements are made on one or several experimental units over a period of time. Such data could be called time-dependent data. From a statistical point of view, if we consider only one experimental unit, we can use a time series analysis. In the other hand, if we consider experimental designs (or observational studies) for several experimental units (or subjects) where each subject is measured at several points in time, we can use the term longitudinal data. In this project, we propose to study several statistical and stochastic models for repeated measures using parametric and non-parametric approaches. In particular, we will study the inference in complex mixed effects models, we will propose novel segmentation models for multiple series, non-parametric methods in dependent models and stochastic models. We will apply these methods to real data from several fields as biometrics, reliability, population dynamics and finance.

9.3. International Research Visitors

9.3.1. Visits of International Scientists

Ricardo Rios, Universidad Central de Venezuela, Caracas: September 2019.

Cristian Meza, Universidad de Valparaiso, Chile, September 2019.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Selection

10.1.1.1. Member of the Conference Program Committees

- ALT 2019
- CLAPEM 2019, Merida, Mexico
- useR!2019

10.1.2. Journal

10.1.2.1. Member of the Editorial Boards

- Stochastic Processes and their Applications
- Journal of Statistical Planning and Inference
- Journal of Computational and Graphical Statistics

10.1.3. Invited Talks

- International Federation of Classification Societies, Thessaloniki, Greece, August 2019.
- useR!2019, Toulouse, France, July 2019.
- R in Official Statistics, Bucharest, Romania, May 2019.
- French Statistical Society, France, Nancy, June 2019.
- Statlearn 2019, Grenoble, France.
- The Big Sick Conference, Zermatt, Switzerland, February, 2019.
- REDIF 2019, La Habana, Cuba, October 2019.
- 4th Worshop on Virus Dynamics, Paris, October 2019.
- ModelBio 2019, Salamnca, Spain, November 2019.

10.1.4. Leadership within the Scientific Community

- Eric Moulines is in charge of the academic supervision of the International Laboratory of Stochastic Algorithms and High-dimensional inference, National Research University, Higher School of Economics, funded by the Russian Academic Excellence Project.
- Eric Moulines is associate researcher of the Alan Turing Institute
- Eric Moulines is elected member of the French Académie des Sciences.
- Julie Josse is visiting researcher at Google France

10.1.5. Scientific Expertise

- Marc Lavielle is member of the Scientific Committee of the High Council for Biotechnologies.
- Marc Lavielle is member of the evaluation committee of the International Center for Mathematics (CIMAT), Guanajuato, Mexico.
- Eric Moulines is member of the award committee of foundation "Charles Defforey".

10.1.6. Research administration

- Marc Lavielle is member of the Scientific Programming Committee (CPS) of the Institute Henri Poincaré (IHP).
- Eric Moulines is a board member of the Institut de Convergence DataIA.
- Julie Josse is elected member of the R foundation.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Master : Julie Josse, Statistics with R, 48, M2, X-HEC

Master : Eric Moulines, Regression models, 36, M2, X-HEC

Engineering School : Eric Moulines, Statistics, 36, 2A, X

Engineering School : Eric Moulines, Markov Chains, 36, 3A, X

Engineering School : Erwan Le Pennec, Statistics, 36, 2A, X

Engineering School : Erwan Le Pennec, Statistical Learning, 36, 3A, X

Engineering School : Marc Lavielle, Statistics in Action, 48, 3A, X

10.2.2. Supervision

PhD defended : Nicolas Brosse, June 2019, Eric Moulines

PhD defended : Geneviève Robin, June 2019, Julie Josse and Eric Moulines

PhD defended : Belhal Karimi, September 2019, Marc Lavielle and Eric Moulines

PhD in progress : Marine Zulian, October 2016, Marc Lavielle

PhD in progress : Wei Jiang , October 2017, Julie Josse and Marc Lavielle

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

Eric Moulines was member of the Evaluation Committee of Inria.

10.3.2. Creation of media or tools for science outreach

Marc Lavielle developed and maintains the learning platform Statistics in Action. The purpose of this online learning platform is to show how statistics (and biostatistics) may be efficiently used in practice using R. It is specifically geared towards teaching statistical modelling concepts and applications for self-study. Indeed, most of the available teaching material tends to be quite "static" while statistical modelling is very much subject to "learning by doing".

Julie Josse (with Nicholas Tierney and Nathalie Vialaneix) developed and maintains the website R-miss-tastic, A resource website on missing data.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] N. BROSSE. *Around the Langevin Monte Carlo algorithm : extensions and applications*, Université Paris Saclay, June 2019, https://hal.inria.fr/tel-02430579

[2] B. KARIMI. *Non-Convex Optimization for Latent Data Models : Algorithms, Analysis and Applications*, Université Paris-Saclay, September 2019, https://tel.archives-ouvertes.fr/tel-02319140

[3] G. ROBIN. *Low-rank methods for heterogeneous and multi-source data*, Université Paris-Saclay, June 2019, https://tel.archives-ouvertes.fr/tel-02168204

Articles in International Peer-Reviewed Journals

[4] A. HAVET, M. LERASLE, É. MOULINES. *Density estimation for RWRE*, in "Mathematical Methods of Statistics", March 2019, https://arxiv.org/abs/1806.05839 [*DOI :* 10.3103/S1066530719010022], https://hal.archives-ouvertes.fr/hal-01815990

[5] F. HUSSON, J. JOSSE, B. NARASIMHAN, G. ROBIN. *Imputation of mixed data with multilevel singular value decomposition*, in "Journal of Computational and Graphical Statistics", 2019, https://arxiv.org/abs/1804.11087 , forthcoming [*DOI :* 10.1080/10618600.2019.1585261], https://hal.archives-ouvertes.fr/hal-01781291

[6] W. JIANG, J. JOSSE, M. LAVIELLE. *Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction*, in "Computational Statistics and Data Analysis", December 2019, 106907 p. , https://arxiv.org/abs/1805.04602 , forthcoming [*DOI :* 10.1016/J.CSDA.2019.106907], https://hal.archives-ouvertes.fr/hal-01958835

[7] B. KARIMI, M. LAVIELLE, É. MOULINES. *f-SAEM: A fast Stochastic Approximation of the EM algorithm for nonlinear mixed effects models*, in "Computational Statistics and Data Analysis", July 2019, forthcoming [*DOI :* 10.1016/J.CSDA.2019.07.001], https://hal.inria.fr/hal-01958248

[8] L. LÊ, M. BERGE, A. TFAYLI, A. BAILLET-GUFFROY, P. PROGNON, A. DOWEK, E. CAUDRON. *Quantification of gemcitabine intravenous drugs by direct measurement in chemotherapy plastic bags using a handheld Raman spectrometer*, in "Talanta", May 2019, vol. 196, pp. 376-380 [*DOI :* 10.1016/J.TALANTA.2018.11.062], https://hal.archives-ouvertes.fr/hal-01970020

[9] A. MARGUET, M. LAVIELLE, E. CINQUEMANI. *Inheritance and variability of kinetic gene expression parameters in microbial cells: modeling and inference from lineage tree data*, in "Bioinformatics", 2019, vol. 35, n^{o} 14, pp. i586-i595 [*DOI :* 10.1093/BIOINFORMATICS/BTZ378], https://hal.archives-ouvertes.fr/hal-02317115

[10] G. ROBIN, J. JOSSE, É. MOULINES, S. SARDY. *Low-rank model with covariates for count data analysis*, in "Journal of Multivariate Analysis", April 2019, vol. 173, https://arxiv.org/abs/1703.02296 , https://hal.archives-ouvertes.fr/hal-01482773

[11] G. ROBIN, O. KLOPP, J. JOSSE, É. MOULINES, R. TIBSHIRANI. *Main effects and interactions in mixed and incomplete data frames*, in "Journal of the American Statistical Association", June 2019 [*DOI :* 10.1080/01621459.2019.1623041], https://hal.archives-ouvertes.fr/hal-02423445

[12] M. TOUZOT, P. SERIS, C. MAHEAS, J. VANMASSENHOVE, A.-L. LANGLOIS, K. MOUBAKIR, S. LAPLANCHE, T. PETITCLERC, C. RIDEL, M. LAVIELLE. *A mathematical model to predict BNP levels in hemodialysis patients*, in "Nephrology", 2019 [*DOI :* 10.1111/NEP.13586], https://hal.archives-ouvertes.fr/hal-02127228

Invited Conferences

[13] B. KARIMI, B. MIASOJEDOW, É. MOULINES, H.-T. WAI. *Non-asymptotic Analysis of Biased Stochastic Approximation Scheme*, in "COLT 2019 - 32nd Annual Conference on Conference on Learning Theory", Phoenix, United States, 2019, pp. 1 - 33, https://hal.inria.fr/hal-02127750

International Conferences with Proceedings

[14] B. KARIMI, M. LAVIELLE. *Efficient Metropolis-Hastings sampling for nonlinear mixed effects models*, in "BAYSM 2018 - Bayesian Young Statisticians Meeting", Warwick, United Kingdom, Bayesian Statistics and New Generations - Proceedings of BAYSM, Springer, 2019, https://hal.inria.fr/hal-01958247

[15] B. KARIMI, H.-T. WAI, É. MOULINES, M. LAVIELLE. *On the Global Convergence of (Fast) Incremental Expectation Maximization Methods*, in "NeurIPS 2019 - 33th Annual Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, https://hal.inria.fr/hal-02334656

Conferences without Proceedings

[16] V. AUDIGIER, F. HUSSON, J. JOSSE, M. RESCHE-RIGON. *Imputation multiple pour données mixtes par analyse factorielle*, in "JdS2019 - 51es Journées de Statistique de la Société Française de Statistique", Vandœuvre-lès-Nancy, France, Société Française de Statistique, June 2019, https://hal-agrocampus-ouest.archives-ouvertes.fr/hal-02355840

[17] T. LEVENT, P. PREUX, E. LE PENNEC, J. BADOSA, G. HENRI, Y. BONNASSIEUX. *Energy Management for Microgrids: a Reinforcement Learning Approach*, in "ISGT-Europe 2019 - IEEE PES Innovative Smart Grid Technologies Europe", Bucharest, France, IEEE, September 2019, pp. 1-5 [*DOI :* 10.1109/ISGTEUROPE.2019.8905538], https://hal.archives-ouvertes.fr/hal-02382232

Other Publications

[18] F. CHOULY, J. LOUBANI, A. LOZINSKI, B. MÉJRI, K. MERITO, S. PASSOS, A. PINEDA. *Computing bi-tangents for transmission belts*, January 2020, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02429962

[19] W. JIANG, M. BOGDAN, J. JOSSE, B. MIASOJEDOW, V. ROCKOVA. *Adaptive Bayesian SLOPE—High-dimensional Model Selection with Missing Values*, January 2020, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02430600

[20] J. JOSSE, N. PROST, E. SCORNET, G. VAROQUAUX. *On the consistency of supervised learning with missing values*, March 2019, https://arxiv.org/abs/1902.06931 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02024202

[21] B. KARIMI, M. LAVIELLE, É. MOULINES. *On the Convergence Properties of the Mini-Batch EM and MCEM Algorithms*, October 2019, working paper or preprint, https://hal.inria.fr/hal-02334485

[22] A. SPORTISSE, C. BOYER, J. JOSSE. *Imputation and low-rank estimation with Missing Non At Random data*, January 2019, https://arxiv.org/abs/1812.11409 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01964720