RESEARCH CENTRE

Saclay - Île-de-France

IN PARTNERSHIP WITH:

Ecole Polytechnique

2021 ACTIVITY REPORT

Project-Team XPOP

Statistical modelling for life sciences

IN COLLABORATION WITH: Centre de Mathématiques Appliquées (CMAP)

DOMAIN

Digital Health, Biology and Earth

THEME

Modeling and Control for Life Sciences

Contents

Project-Team XPOP

Creation of the Project-Team: 2017 July 01

Keywords

Computer sciences and digital sciences

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)

A6.1.2. – Stochastic Modeling

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

B2.3. – Epidemiology

B2.4.1. – Pharmaco kinetics and dynamics

B9.1.1. – E-learning, MOOC

1 Team members, visitors, external collaborators

Research Scientists

- Marc Lavielle [Team leader, Inria, Senior Researcher]
- Erwan Le Pennec [École polytechnique, Researcher]

Faculty Member

- Eric Moulines [École polytechnique, Professor, HDR]

Post-Doctoral Fellow

- Angie Pineda Centeno [Inria, until Nov 2021]

PhD Students

- Pablo Jimenez [Institut Polytechnique de Paris]
- Achille Thin [École polytechnique]

Administrative Assistant

- Hanadi Dib [Inria]

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:

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

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 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.

4.2 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.

4.3 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.

- 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.

4.4 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.

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

4.5 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 Social and environmental responsibility

Marc Lavielle is member of the Scientific Committee of *RESPIRE*, a French organization for the improvement of air quality.

6 New results

6.1 Efficient automatic incomplete data model building

Participants: Marc Lavielle.

We have proposed in [11] an extension of the EM algorithm and its stochastic versions for the construction of incomplete data models when the selected model minimizes a penalized likelihood criterion. This optimization problem is particularly challenging in the context of incomplete data, even when the model is relatively simple. However, by completing the data, the E-step of the algorithm allows us to simplify this problem of complete model selection into a classical problem of complete model selection that does not pose any major difficulties. We then have shown that the criterion to be minimized decreases with each iteration of the algorithm. Examples of the use of these algorithms include the identification of regression mixture models and the construction of nonlinear mixed-effects models.

The SAMBA (Stochastic Approximation for Model Building Algorithm) procedure was developed specifically for the construction of mixed-effects models. We have shown in [5] how this algorithm can be used to speed up this process of model building by identifying at each step how best to improve some of the model components. The principle of this algorithm basically consists in learning something about the best model, even when a poor model is used to fit the data. A comparison study of the SAMBA procedure with SCM and COSSAC show similar performances on several real data examples but with a much-reduced computing time. This algorithm is now implemented in Monolix and in the R package Rsmlx.

6.2 Modelling the COVID-19 dynamics

Participants: Marc Lavielle.

Short-term forecasting of the COVID-19 pandemic is required to facilitate the planning of COVID-19 healthcare demand in hospitals. We have shown in [12] how daily hospital data can be used to track the evolution of the COVID-19 epidemic in France. A piecewise defined dynamic model allows a very good fit of the available data on hospital admissions, deaths and discharges. The change-points detected correspond to moments when the dynamics of the epidemic changed abruptly. Although the proposed model is relatively simple, it can serve several purposes: It is an analytical tool to better understand what has happened so far by relating observed changes to changes in health policy or the evolution of the virus. It is also a surveillance tool that can be used effectively to warn of a resurgence of epidemic activity, and finally a short-term forecasting tool if conditions remain unchanged. The model, data and fits are implemented in an interactive web application.

In collaboration with Institut Pasteur (and other groups), we have evaluated in [13] the performance of 12 individual models and 19 predictors to anticipate French COVID-19 related healthcare needs from September 7th 2020 to March 6th 2021. We then built an ensemble model by combining the individual forecasts and tesedt this model from March 7th to July 6th 2021. We found that inclusion of early predictors (epidemiological, mobility and meteorological predictors) can halve the root mean square error for 14-day ahead forecasts, with epidemiological and mobility predictors contributing the most to the improvement.

We built in [4] an SIR-type compartmental model with two additional compartments: D (deceased patients); L (individuals who will die but who will not infect anybody due to social or medical isolation) and integration of a time-dependent transmission rate and a periodical weekly component linked to the way in which cases and deaths are reported. The model was implemented in a web application (as of 2 June 2020). It was shown to be able to accurately capture the changes in the dynamics of the pandemic for 20 countries whatever the type of pandemic spread or containment measures: for instance, the model explains 97% of the variance of US data (daily cases) and predicts the number of deaths at a 2-week horizon with an error of 1%. In early performance evaluation, our model showed a high level of accuracy between prediction and observed data.

6.3 Variance Reduction for Dependent Sequences with Applications to Stochastic Gradient MCMC

Participants: Eric Moulines.

We proposed in [2] a novel and practical variance reduction approach for additive functionals of dependent sequences. Our approach combines the use of control variates with the minimization of an empirical variance estimate. We analyzed finite sample properties of the proposed method and derive finite-time bounds of the excess asymptotic variance to zero. We applied our methodology to stochastic gradient Markov chain Monte Carlo (SGMCMC) methods for Bayesian inference on large data sets and combined it with existing variance reduction methods for SGMCMC. We have presented empirical results carried out on a number of benchmark examples showing that our variance reduction method achieves significant improvement as compared to state-of-the-art methods at the expense of a moderate increase of computational overhead.

6.4 Stochastic gradient Langevin dynamics with dependent data streams in the log-concave case

Participants: Eric Moulines.

We have studied in [1] the problem of sampling from a probability distribution on R which has a density w.r.t. the Lebesgue measure known up to a normalization factor We analyzed a sampling method based on the Euler discretization of the Langevin stochastic differential equations under the assumptions that the potential U is continuously differentiable, ∇U is Lipschitz, and U is strongly concave. We focused on the case where the gradient of the log-density cannot be directly computed but unbiased estimates of the gradient from possibly dependent observations are available. This setting can be seen as a combination of a stochastic approximation (here stochastic gradient) type algorithms with discretized Langevin dynamics. We obtained an upper bound of the Wasserstein-2 distance between the law of the iterates of this algorithm and the target distribution π with constants depending explicitly on the Lipschitz and strong convexity constants of the potential and the dimension of the space. Finally, under weaker assumptions on U and its gradient but in the presence of independent observations, we obtain analogous results in Wasserstein-2 distance.

6.5 The Perturbed Prox-Preconditioned SPIDER algorithm

Participants: Eric Moulines.

Incremental Expectation Maximization (EM) algorithms were introduced to design EM for the large scale learning framework by avoiding the full data set to be processed at each iteration. Nevertheless, these algorithms all assume that the conditional expectations of the sufficient statistics are explicit. We then proposed in [9] a novel algorithm named Perturbed Prox-Preconditioned SPIDER (3P-SPIDER), which builds on the Stochastic Path Integral Differential EstimatoR EM (SPIDER-EM) algorithm. The 3P-SPIDER algorithm addresses many intractabilities of the E-step of EM; it also deals with non-smooth regularization and convex constraint set. Numerical experiments show that 3P-SPIDER outperforms other incremental EM methods.

Studying the convergence in expectation, we have shown in [8] that 3P-SPIDER achieves a near-optimal oracle inequality even when the gradient is estimated by Monte Carlo methods.

7 Bilateral contracts and grants with industry

7.1 Bilateral contracts with industry

Participant: Marc Lavielle.

Xpop has a contract with ADELIS, a French analytical instrumentation company. The goal of this collaboration is to develop a method to analyze the physiological size profile of circulating DNA to detect different types of pathological abnormalities.

8 Partnerships and cooperations

Participant: Marc Lavielle.

8.1 National initiatives

8.1.1 ANR

Mixed-Effects Models of Intracellular Processes: Methods, Tools and Applications (MEMIP)
Period: from 2017-01-01 to 2021-06-30
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.)

8.1.2 Institut National du Cancer (INCa)

Targeting Rac-dependent actin polymerization in cutaneous melanoma - Institut National du Cancer
Period: from 2018-01-01 to 2021-11-20
Coordinator: Alexis Gautreau (Ecole Polytechnique)
Other partners: Laboratoire de Biochimie (Polytechnique), Institut Curie, INSERM.

8.2 National partnerships

Marc Lavielle is co-leader of the program *Numerical Innovation and Data Science for Healthcare*, created to address technological developments and data issues in the medical field. This sponsorship program with Ecole polytechnique and Sanofi was inaugurated in 2020.

9 Dissemination

Participants: Marc Lavielle, Erwan Le Pennec, Eric Moulines.

9.1 Promoting scientific activities

9.1.1 Journal

Member of the editorial boards

- Stochastic Processes and their Applications

- Journal of Statistical Planning and Inference
- Electronic Journal of Statistics
- Comptes Rendus de l'Académie des Sciences

Reviewer - reviewing activities : diverse and varied

9.1.2 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 is associate researcher of the Alan Turing Institute
- Eric Moulines is scientific director Hi! Paris, Paris Center for Artificial Intelligence for Business and Society
- Eric Moulines is elected member of the French Académie des Sciences.

9.1.3 Scientific expertise

- Marc Lavielle was member of the Scientific Committee of the High Council for Biotechnologies from 2012 to 2021.
- 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".

9.1.4 Research administration

- Marc Lavielle was member of the Scientific Programming Committee (CPS) of the Institute Henri Poincaré (IHP) from 2013 to 2021.
- Marc Lavielle is member of the research working group *Pharmacology* of the Institut thématique multi-organismes (ITMO) *Technologies pour la Santé*.
- Eric Moulines is member of the Evaluation Committee of the Swiss nationale Science Foundations.
- Eric Moulines is member of the Evaluation Committee of IVADO and grant APOGEE, Canada.

9.2 Teaching - Supervision - Juries

9.2.1 Teaching

- Master : Eric Moulines, Regression models, 36, M2
- 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
- Master : Marc Lavielle, Mixed-effects models, 24, M2

9.2.2 Supervision

PhD in progress:

- Achille Thin, October 2020
- Pablo Jimenez, October 2020

9.3 Popularization

9.3.1 Platforms

Marc Lavielle developed and maintains the platform covidix for Covid-19 data visualization and modelling.

9.3.2 Education

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".

10 Scientific production

10.1 Publications of the year

International journals

[1] M. Barkhagen, N. Chau, É. Moulines, M. Rásonyi, S. Sabanis and Y. Zhang. 'On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case'. In: *Bernoulli* 27.1 (1st Feb. 2021). DOI: 10.3150/19-BEJ1187. URL: https://hal.inria.fr/hal-03529653.

[2] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov and S. Samsonov. 'Variance Reduction for Dependent Sequences with Applications to Stochastic Gradient MCMC'. In: *SIAM/ASA Journal on Uncertainty Quantification* 9 (Jan. 2021), pp. 507–535. DOI: 10.1137/19m1301199. URL: https://hal.inria.fr/hal-03529417.

[3] G. Fort, P. Gach and E. Moulines. 'Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence'. In: *Statistics and Computing* 31.48 (2021). URL: https://hal.archives-ouvertes.fr/hal-02617725.

[4] M. Lavielle, M. Faron, J. H. Lefevre and J.-D. Zeitoun. 'Predicting the propagation of COVID-19 at an international scale: extension of an SIR model'. In: *BMJ Open* 11.5 (May 2021), e041472. DOI: 10.1136/bmjopen-2020-041472. URL: https://hal.sorbonne-universite.fr/hal-03239518.

[5] M. Prague and M. Lavielle. 'SAMBA: a Novel Method for Fast Automatic Model Building in Nonlinear Mixed-Effects Models'. In: *CPT: Pharmacometrics and Systems Pharmacology* (2021). URL: https://hal.inria.fr/hal-03410025.

[6] A. Shokry, S. Medina-González, P. Baraldi, E. Zio, E. Moulines and A. Espuña. 'A machine learning-based methodology for multi-parametric solution of chemical processes operation optimization under uncertainty'. In: *Chemical Engineering Journal* 425 (Dec. 2021), p. 131632. DOI: 10.1016/j.cej.2021.131632. URL: https://hal-mines-paristech.archives-ouvertes.fr/hal-03480576.

[7] J.-D. Zeitoun, M. Faron, S. Manternach, J. Fourquet, M. Lavielle and J. Lefevre. 'Reciprocal association between participation to a national election and the epidemic spread of COVID-19 in France: nationwide observational and dynamic modeling study.' In: *European Journal of Public Health* (2021). DOI: 10.1101/2020.05.14.20090100. URL: https://hal.archives-ouvertes.fr/hal-02995300.

International peer-reviewed conferences

[8] G. Fort and E. Moulines. 'The perturbed prox-preconditioned spider algorithm: non-asymptotic convergence bounds'. In: SSP 2021 - IEEE Statistical Signal Processing Workshop. IEEE Proceedings of SSP 2021. Rio de Janeiro, Brazil, 11th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03183775.

[9] G. Fort and E. Moulines. 'The Perturbed Prox-Preconditioned SPIDER algorithm for EM-based large scale learning'. In: SSP 2021 - IEEE Statistical Signal Processing Workshop. IEEE Proceedings of SSP 2021. Rio de Janeiro, Brazil, 11th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03183774.

[10] G. Fort, E. Moulines and H.-T. Wai. 'Geom-SPIDER-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization'. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada, Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03021394.

Reports & preprints

[11] M. Lavielle. *Some EM-type algorithms for incomplete data model building*. 2021. URL: https://hal.archives-ouvertes.fr/hal-03512130.

[12] M. Lavielle. *Using hospital data for monitoring the dynamics of COVID 19 in France*. 2021. URL: https://hal.archives-ouvertes.fr/hal-03321804.

[13] J. Paireau, A. Andronico, N. Hozé, M. Layan, P. Crepey, A. Roumagnac, M. Lavielle, P.-Y. Boëlle and S. Cauchemez. *An ensemble model based on early predictors to forecast COVID-19 healthcare demand in France*. Nov. 2021. URL: https://hal-pasteur.archives-ouvertes.fr/pasteur-03149082.