Activity Report 2016

Team XPOP

statistical modelling for life sciences

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Saclay - Île-de-France

THEME
Computational Neuroscience and Medecine

Table of contents

Team XPOP

Creation of the Team: 2016 January 01

Keywords:

Computer Science and Digital Science:

3.1.1. - Modeling, representation
3.2.3. - Inference
3.3. - Data and knowledge analysis
3.3.1. - On-line analytical processing
3.3.2. - Data mining
3.3.3. - Big data analysis
3.4.1. - Supervised learning
3.4.2. - Unsupervised learning
3.4.4. - Optimization and learning
3.4.5. - Bayesian methods
3.4.6. - Neural networks
3.4.7. - Kernel methods
3.4.8. - Deep learning
5.9.2. - Estimation, modeling
6.1.1. - Continuous Modeling (PDE, ODE)
6.2.2. - Numerical probability
6.2.3. - Probabilistic methods
6.2.4. - Statistical methods
6.3.3. - Data processing
6.3.5. - Uncertainty Quantification

Other Research Topics and Application Domains:

1.1.5. - Genetics
1.1.6. - Genomics
1.1.9. - Bioinformatics
1.1.11. - Systems biology
2.2.3. - Cancer
2.2.4. - Infectious diseases, Virology
2.4.1. - Pharmaco kinetics and dynamics
9.1.1. - E-learning, MOOC

1. Members

Research Scientist

Marc Lavielle [Team leader, Inria, Senior Researcher]

Faculty Members

Julie Josse [Ecole Polytechnique, Associate Professor]
Erwan Le Pennec [Ecole Polytechnique, Associate Professor]

Eric Moulines [Ecole Polytechnique, Professor]

Engineer

François-Marie Floch [Inria, until Oct 2016]

PhD Students

Nicolas Brosse [Ecole Polytechnique]
Mohammed Karimi [Inria, from Oct 2016]
Genevieve Robin [Ecole Polytechnique, from Sep 2016]
Marine Zulian [Dassault Systemes, from Sep 2016, granted by CIFRE]

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 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, 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. A strong partnership with Lixoft (the spin-off company who continue developing MONOLIX) is indispensable to maintaining this positioning.

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. Population pharmacometrics

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.

4.2. Precision medicine and pharmacogenomics

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.

4.3. Biology - Intracellular processes

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.

5. Highlights of the Year

5.1. Highlights of the Year

R Foundation

Julie Josse has been elected member of the R Foundation for Statistical Computing.

mlxR 3.1

mlxR 3.1 available on CRAN

6. New Software and Platforms

6.1. mlxR

A R package for the simulation and visualization of longitudinal data. 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.

6.2. FactoMineR

A R package dedicated to principal component methods (PCA, Correspondence Analysis for contingency tables, Multiple Correspondence Analysis for categorical data, MFA for multi-blocks data). Google users group and you-tube videos available.

6.3. missMDA

A R package to perform principal component methods (PCA, MCA, MFA) with missing values and to impute continuous, categorical and mixed data. Multiple imputation is available.

6.4. denoiseR

A R package that approximates a low-rank matrix from noisy data (Gaussian and Poisson Noise). Singular values shrinkage methods are implemented.

7. New Results

7.1. Identifiability in mixed effects models

We considered the question of model identifiability within the context of nonlinear mixed effects models. Although there has been extensive research in the area of fixed effects models, much less attention has been paid to random effects models.

In this context we distinguish between theoretical identifiability, in which different parameter values lead to non-identical probability distributions, structural identifiability which concerns the algebraic properties of the structural model, and practical identifiability, whereby the model may be theoretically identifiable but the design of the experiment may make parameter estimation difficult and imprecise.

We have explored a number of pharmacokinetic models which are known to be non-identifiable at an individual level but can become identifiable at the population level if a number of specific assumptions on the probabilistic model hold. Essentially if the probabilistic models are different, even though the structural models are non-identifiable, then they will lead to different likelihoods. The findings are supported through simulations.

7.2. Enhanced Method for Diagnosing Pharmacometric Models

For nonlinear mixed-effects pharmacometric models, diagnostic approaches often rely on individual parameters, also called empirical Bayes estimates (EBEs), estimated through maximizing conditional distributions. When individual data are sparse, the distribution of EBEs can "shrink" towards the same population value, and as a direct consequence, resulting diagnostics can be misleading.

Instead of maximizing each individual conditional distribution of individual parameters, we propose to randomly sample them in order to obtain values better spread out over the marginal distribution of individual parameters.

We have evaluated, through diagnostic plots and statistical tests, hypothesis related to the distribution of the individual parameters and shown that the proposed method leads to more reliable results than using the EBEs. In particular, diagnostic plots are more meaningful, the rate of type I error is correctly controlled and its power increases when the degree of misspecification increases. An application to the warfarin pharmacokinetic data confirms the interest of the approach for practical applications.

7.3. A Shrinkage-Thresholding Metropolis Adjusted Langevin Algorithm for Bayesian Variable Selection

We have introduced a new Markov Chain Monte Carlo method for Bayesian variable selection in high dimensional settings. The algorithm is a Hastings-Metropolis sampler with a proposal mechanism which combines a Metropolis Adjusted Langevin (MALA) step to propose local moves associated with a shrinkage-thresholding step allowing to propose new models.

The geometric ergodicity of this new trans-dimensional Markov Chain Monte Carlo sampler was established. An extensive numerical experiment, on simulated and real data, illustrates the performance of the proposed algorithm in comparison with some more classical trans-dimensional algorithms

7.4. Maximum likelihood estimation of a low-order building model

Our objective was to investigate the accuracy of the estimates learned with an open loop model of a building whereas the data is actually collected in closed loop, which corresponds to the true exploitation of buildings. We have proposed a simple model based on an equivalent RC network whose parameters are physically interpretable. We also described the maximum likelihood estimation of these parameters by the EM algorithm, and derived their statistical properties.

The numerical experiments clearly show the potential of the method, in terms of accuracy and robustness. We have emphasized the fact that the estimations are linked to the generating process for the observations, which includes the command system. For instance, the features of the building are correctly estimated if there is a significant gap between the heating and cooling setpoint.

7.5. LP-convergence of a Girsanov theorem based particle filter

We have analyzed the LP-convergence of a previously proposed Girsanov theorem based particle filter for discretely observed stochastic differential equation (SDE) models. We proved the convergence of the algorithm with the number of particles tending to infinity by requiring a moment condition and a step-wise initial condition boundedness for the stochastic exponential process giving the likelihood ratio of the SDEs. The practical implications of the condition are illustrated with an Ornstein–Uhlenbeck model and with a non-linear Bene's model.

7.6. Adaptive estimation in the nonparametric random coefficients binary choice model by needlet thresholding

In the random coefficients binary choice model, a binary variable equals 1 iff an index $X^T\beta$ is positive. The vectors X and β are independent and belong to the sphere S^{d-1} in R^d. We have proven lower bounds on the minimax risk for estimation of the density f_β over Besov bodies where the loss is a power of the $L^p(S^{d-1})$ norm for $1 \leq p \leq \infty$. We have shown that a hard thresholding estimator based on a needlet expansion with data-driven thresholds achieves these lower bounds up to logarithmic factors.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contract with Industry

Contract with Lixoft

9. Partnerships and Cooperations

9.1. European Initiatives

9.1.1. FP7 & H2020 Projects

The Drug Disease Model Resources (DDMoRe) consortium will build and maintain a universally applicable, open source, model-based framework, intended as the gold standard for future collaborative drug and disease modeling and simulation.

The DDMoRe project is supported by the Innovative Medicines Initiative (IMI), a large-scale public-private partnership between the European Union and the pharmaceutical industry association EFPIA.

Marc Lavielle was leader of WP6: "New tools for Model Based Drug Development".

DDMoRe website: http://www.ddmore.eu

Duration: 2010 - 2016

Project members: Uppsala Universitet, Sweden; University of Navarra, Spain; Universiteit Leiden, Netherlands; Université Paris Diderot, France; Universita degli Studi di Pavia, Italy; UCB Pharma , Belgium; Simcyp, UK; Pfizer, UK; Optimata , Israel; Novo Nordisk , Denmark; Novartis, Switzerland; Merck Serono, Switzerland; Takeda, Switzerland; Mango Business Solutions , UK; Lixoft, France; Interface Europe, Belgium; Institut de Recherches Internationales Servier, France; Inria, France; GlaxoSmithKline Research and Development, UK; Freie Universitat Berlin, Germany; F. Hoffmann - La Roche , Switzerland; EMBL - European Bioinformatics Institute, UK; Eli Lilly , UK; Cyprotex Discovery, UK; Consiglio Nazionale delle Ricerche, Italy; AstraZeneca, Sweden.

9.2. International Initiatives

9.2.1. Informal International Partners

Marc Lavielle is Adjunct Professor at the Faculty of Pharmacy of Florida University.

Marc Lavielle is Adjunct Professor at the Faculty of Pharmacy of Buffalo University.

Julie Josse collaborates with Susan Holmes, Stanford University.

Eric Moulines regularly collaborates with Sean P. Meyn, University of Florida.

9.3. International Research Visitors

9.3.1. Visits of International Scientists

Ricardo Rios, Universidad Central de Venezuela, Caracas: September 2016.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

Julie josse was chair of useR!2016, Stanford, CA, USA, July 2016. http://user2016.org.

10.1.2. Scientific Expertise

Marc Lavielle is member of the Scientific Committee of the High Council for Biotechnologies

10.1.3. Research administration

Marc Lavielle is member of

- the Scientific Programming Committee (CPS) of the Institute Henri Poincaré (IHP),
- the Executive Board (CA) of SMAI.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

Master : Eric Moulines, Regresson 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, Time Series, 24, 3A, X

10.2.2. Supervision

PhD in progress : Nicolas Brosse, September 2016, Eric Moulines

PhD in progress : Geneviève Robin, September 2016, Julie Josse

PhD in progress : Belhal Karimi, October 2016, Marc Lavielle and Eric Moulines

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

11. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] H. BRAHAM, S. B. JEMAA, G. FORT, E. MOULINES, B. SAYRAC. *Fixed Rank Kriging for Cellular Coverage Analysis*, in "IEEE Transactions on Vehicular Technology", 2016, 11 p. [*DOI :* 10.1109/TVT.2016.2599842], https://hal.inria.fr/hal-01418961

[2] H. BRAHAM, S. B. JEMAA, G. FORT, E. MOULINES, B. SAYRAC. *Spatial Prediction Under Location Uncertainty in Cellular Networks*, in "IEEE Transactions on Wireless Communications", 2016, vol. 15, pp. 7633 - 7643 [*DOI :* 10.1109/TWC.2016.2605676], https://hal.inria.fr/hal-01419051

[3] M. JALA, C. C. LEVY LEDUC, E. MOULINES, E. CONIL, J. WIART. *Sequential design of computer experiments for the assessment of fetus exposure to electromagnetic fields*, in "Technometrics", January 2016, vol. 58, n° 1, pp. 30–42 [*DOI :* 10.1080/00401706.2014.979951], https://hal.archives-ouvertes.fr/hal-01418859

[4] M. LAVIELLE, L. AARONS. *What do we mean by identiability in mixed effects models?*, in "Journal of Pharmacokinetics and Pharmacodynamics", September 2016, https://hal.archives-ouvertes.fr/hal-01365535

[5] M. LAVIELLE, B. RIBBA. *Enhanced Method for Diagnosing Pharmacometric Models: Random Sampling from Conditional Distributions*, in "Pharmaceutical Research / Pharmaceutical Research (Dordrecht)", 2016 [*DOI :* 10.1007/s11095-016-2020-3], https://hal.archives-ouvertes.fr/hal-01365532

[6] A. SCHRECK, G. FORT, S. LE CORFF, E. MOULINES. *A Shrinkage-Thresholding Metropolis Adjusted Langevin Algorithm for Bayesian Variable Selection*, in "IEEE Journal of Selected Topics in Signal Processing", 2016, vol. 10, pp. 366 - 375 [*DOI :* 10.1109/JSTSP.2015.2496546], https://hal.inria.fr/hal-01418960

International Conferences with Proceedings

[7] P. ILHE, F. ROUEFF, E. MOULINES, A. SOULOUMIAC. *Nonparametric estimation of a shot-noise process*, in "SSP 16 - Statistical Signal Processing Workshop", Palma de Mallorca, Spain, IEEE Signal Processing Society, June 2016 [*DOI :* 10.1109/SSP.2016.7551709], https://hal.inria.fr/hal-01418963

[8] J. LAFOND, H.-T. WAI, E. MOULINES. *D-FW: Communication efficient distributed algorithms for high-dimensional sparse optimization*, in "International Conference on Acoustics, Speech and Signal Processing", Shangai, China, March 2016, pp. 4144 - 4148 [*DOI :* 10.1109/ICASSP.2016.7472457], https://hal.inria.fr/hal-01419048

[9] T. NABIL, E. MOULINES, F. ROUEFF, J.-M. JICQUEL, A. GIRARD. *Maximum likelihood estimation of a low-order building model*, in "EUSIPCO 2016: 2016 24th European Signal Processing Conference", Budapest, Hungary, August 2016 [*DOI :* 10.1109/EUSIPCO.2016.7760339], https://hal.inria.fr/hal-01419050

[10] S. SÄRKKÄ, E. MOULINES. *On the L<inf>P</inf>-convergence of a Girsanov theorem based particle filter*, in "ICASSP 2016 - International Conference on Acoustics, Speech and Signal Processing", Shangai, China, March 2016, pp. 3989 - 3993 [*DOI :* 10.1109/ICASSP.2016.7472426], https://hal.inria.fr/hal-01419046

Other Publications

[11] E. GAUTIER, E. LE PENNEC. *Adaptive estimation in the nonparametric random coefficients binary choice model by needlet thresholding*, September 2016, working paper or preprint, https://hal.inria.fr/inria-00601274

[12] M. LAVIELLE, L. AARONS. *What do we mean by identifiability in mixed effects models?*, January 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01251986

[13] L. MONTUELLE, E. LE PENNEC. *PAC-Bayesian aggregation of affine estimators*, October 2016, working paper or preprint, https://hal.inria.fr/hal-01070805