Chan Zuckerberg Intitiative Proposal: Saemix: Open source R package implementing the SAEM for mixed-effects models

Belhal Karimi[1] and Emmanuelle Comets[2]
[1]CMAP, Ecole Polytechnique, Universite Paris-Saclay, 91128 Palaiseau, France
[2]INSERM, IAME, UMR 1137, U. Paris Diderot, Paris, France
belhal.karimi@polytechnique.edu

January 21, 2020

1 Introduction

The `saemix` package for R provides maximum likelihood estimates of parameters in nonlinear mixed effect models (NLMEM), using a modern and efficient estimation algorithm, the Stochastic Approximation of the Expectation-Maximisation (SAEM).

Longitudinal data arise in many fields, such as agronomy, spatial analysis, imagery, clinical trials, and have been particularly prominent in the field of pharmacokinetics (PK) and pharmacodynamics (PD), where increasingly complex models involving mechanistic and empirical processes have been developed to describe the time course of and responses to drugs. Nonlinear models pose unique challenges in terms of estimation methods, and have driven the research to provide better estimation of parameters as well as the associated uncertainty, diagnostics of model misspecification and more informative designs. The SAEM algorithm, based on two highly cited publications by one of our project members Marc Lavielle, see (Delyon et al., 1999) and (Kuhn and Lavielle, 2004), was implemented in R in 2011 in the `saemix` R package (Comets et al., 2017). Several applications of SAEM in agronomy, animal breeding and PKPD analysis have been published using `saemix`.

PK/PD analyses are now a fundamental element of the registration file submitted to health authority for the approval of new drugs, but NLMEM are also increasingly applied to other areas. In clinical trials, they complement the point analyses by offering a unique understanding of the evolution of disease or treatment action. In cohort studies, they allow to model trajectories such as growth or cognitive decline. Joint models are now routinely used to link the evolution of markers with the occurrence of an event. Making use of S4 classes and methods to provide user-friendly interaction, `saemix` provides a new estimation tool with a powerful exact algorithm to the R community. The `saemix`

package for R provides maximum likelihood estimates of parameters in nonlinear mixed effect (NLME) models, using a modern and efficient estimation algorithm, the Stochastic Approximation of the Expectation-Maximisation (SAEM) algorithm based on two highly cited publication by one of our project member Marc Lavielle, see (Delyon et al., 1999) and (Kuhn and Lavielle, 2004). Making use of S4 classes and methods to provide user-friendly interaction, this package provides a new estimation tool to the R community.

2 Goals of this proposal (Not included)

The goal of this proposal is to create a fully functional `saemix` to overcome the limitations of current packages for NLMEM. Specifically, we aim to:

- increase the number of contributors and users of `saemix` R package (**Target**: 3 new contributors and developers, double the number of users of the package).
- bridge the gap between the commercial version of SAEM found in Monolix (developed in parallel by one of our core members Marc Lavielle) and the R open source package in terms of model scope (**Target**: implement several key features and extensions, as detailed next section)
- continue the implementation of the research done by Belhal Karimi during his PhD (Karimi et al., 2020) and published in *Computational Statistics and Data Analysis*.

We also wish to intensify our work in

- improving the usability of the package through its documentation and case studies (**Target**: user guide and collection of case studies using the R package showcased on a new website).
- the interoperability with other packages also developed by members of our group (**Target**: one new package built on the interoperability of our saemix package and an existing ODE solver R package. Interoperability with the `npde` package for model evaluation and visualization).
- increasing its visibility through training and communication (**Target**: trimestrial seminars with project team, training workshops at one statistical and one pharmacometric conference)

An exhaustive list of new extensions, features and measures that rely on this funding are detailed in the next section (**Work Plan**).

Completion of each goal regarding software development and features enhancement of the package will be assessed through a checklist principle. Regarding the visibility and userbase growth objectives, the applicants of this grant will track metrics yielding the evolution of the userbase, the number of reported bugs and the number of citations.

3 Work Plan

The requested funding would be devoted to enhancing the versatility of the package in dealing with general NLMEM.

3.1 Extensions

- The current CRAN version of `saemix` only handles continuous responses, for example drug concentrations in the blood of a patient. Following our work in (Karimi et al., 2020), we would like to include models for **non continuous data**, for example when the outcome is a count variable (eg the number of seizures per week) or a categorical variable (eg pain scores). This extension is under development and is made available on github (`https://github.com/belhal/saemixextension`).

- A major request given the increasing model complexity is the ability to handle **multi-response** models, allowing simultaneous modelling of several biomarkers or joint modelling of longitudinal data and time to an event of interest. For survival models we also need to treat **censored** data.

- Finally, **latent variables** are a natural extension of the missing data framework in which the EM algorithms operate so efficiently, so we would like to add Hidden Markov Models (HMM), following on the work of Maud Delattre on the analysis of an antiepileptic agent ((Delattre et al., 2012) and (Delattre and Lavielle, 2012)).

Note: The current project members are aware of the technical expertise to achieve state-of-the-art implementation of such development and pursue this grant to onboard and supervise experts software engineers for those tasks whereas current members will keep on focussing on the core statistical engine.

3.2 New Features

Besides extending the current scope of the package to allow new users to use our tool on a wide array of applications, we also plan to develop and implement methodological extensions.

- **Uncertainty:** quantifying estimation error on model parameters (SE) is essential to propagate uncertainty and make informed decisions. We plan to extend asymptotic estimates of SE for non-continuous data, which require stochastic or numerical integration approaches, as well as implement finite distance uncertainty estimates such as the SIR.

- **Model Selection:** we plan to develop several automatic covariate search methods for covariate selection given a base model and a set of candidate variables. Furthermore, one request we have had for several months is to add a powerful tool for automatic PK model building, in line with similar R tools for regression.

- **Visualization:** In the current version of the `saemix` R package, a number of diagnostic graphs, aimed at evaluating model properties and guiding model building, are produced automatically when running `saemix()`, and are output by a call to the `plot()` function applied to the object resulting from the fit (see Figures 1 and 2 in Section 6). Those plots are functional for current models but would need to be extended to new models after delivering all of the above.

- **Interoperability:** Monolix, as the commercial product using the core engine found in our `saemix` package, is allowing users to access enhanced features and to use more complex models to deal with sophisticated tasks such as modelling the response to treatments for Hepatitis C. This kind of task often involves complex ordinary differential equation (ODE) PK models, such as viral kinetics in the case of Hepatitis.

  - We would like to provide interoperability between our package `saemix` and the `lixoftConnectors` package, found on CRAN as well, that allows R users to fit complex ODE-based models and thus allowing us to reach a broader audience. A package under development achieving this goal can be found here: https://github.com/belhal/saemixmlxconnectors and would need further coding and improvements for publication on CRAN. A portion of the requested fund would be dedicated to hire a part-time developer to secure this goal.
  - We also plan on including functionalities of the `npde` R package (http://www.npde.biostat.fr/), developed by one of the applicant Emmanuelle Comets, in our `saemix` package to take advantage of its model evaluation process.

3.3 Visibility and Education

One important aspect of this project that we would like to improve is the visibility and the awareness around the R package to attract more and more users. With `saemix`, our aim is to address the general statistical community, much as the `nlme` package which is distributed directly with R, contrasting with current existing packages or solutions such as `nlmixR` (Fidler et al., 2019) or Monolix which are tailored to a specific audience of pharmacometricians.

We plan to improve on that domain thanks to the following actions:

- **Website**: We will hire a freelancer to redesign the current outdated project website http://www.saemix.biostat.fr/. Influenced by a current CZI project proposal called PlotLy https://plot.ly/, we would like to improve the ergonomy and design of our current webpage. This is an important resource for communication and assistance to our userbase.

- **Reproducibility**: Similarly to another package (`simulx`, developed by one of our project members Marc Lavielle and dealing with data simulation) website http://simulx.webpopix.org/case-studies/, we plan on enhancing the content

of ours in terms of case studies and illustrative coding lines in order to increase our reproducibility standards.

We would also like to develop a simple landing page for users to upload their own datasets attached to their analysis notebooks. This will allow the community to update their benchmarks with modern datasets.

- **Conferences**: Attending conferences is important to showcase our newest studies and to talk with prominent researchers of the domain. The main annual conference we attend is the **Population Approach Group in Europe** event https://www.page-meeting.org/ where the `saemix` tool and the PKPD analyses we perform are often presented. Another important conference is **useR!** where newly developed R packages are presented each year, and which has a broader spectrum of statisticians than the pharmacometrics community.

  The requested fund will only be used to finance new members to attend this conference while current project members will use their own research funding.

- **Education**: Educating our members and interested researchers is crucial for improvement of the tool and to grow our userbase. After having updated the documentation that comes with R package on CRAN, with new examples and new features explanation, we plan on organising 'demo days' where several research groups would be gathered to watch a presentation on how the package works, have the opportunity to try it on demo or their own data. We would start with project-teams from our organization such as INRIA or INSERM. Indeed, we regularly receive requests for debugging or implementation on specific problems from researchers within our institution, sometimes dealing with completely different subjects than life sciences. Doing a day full of workshops and presentationx would help many teams and would increase the awareness around our R package. The requested funding would be used for transportation and the organisation of the workshops, using faculty rooms to limit the budget.

4 Existing Support

Applicants resources in addition to the requested funding: We, as applicants, will provide desks and offices next to our principal project members for future staff members. The current project members are willing to provide their expertise in their own domain ranging from software engineering in R to statistical modelling and especially NLME modelling. Networks and colleagues of each member are also at the disposal of the members giving access to a variety of experts from top national research institutes such as INSERM or INRIA.
Current and recent financial support: Currently, Github contributions mainly stems from Masters or PhD students that use the package for their own research and sometimes need to improve it. All members are also maintaining it during their freetime. Most are

being funded through their direct supervisors (public funding or research budget of the advisor).

Nevertheless there is no central and constant funding available at any time, and in particular no funding available for key points such as technical developments, code profiling, website design and communication material. This lack of constant available funding for new projects related to the package is a reason the applicants apply to the requested CZI grant.

5 Landscape Analysis (250 words)

R contains several packages for non-linear mixed effects modelling. The `nlme` package, distributed with R base, has a companion package `nlmeODE` allowing to specify models in ODE form. This package has strong statistical base, but uses an outdated linearisation-based algorithm which has been shown both theoretically and practically to provide biased parameter estimates. It is very limited in the scope of models that can be run and in practice convergence is often tricky.

Our main competitor within R is the new `nlmixr` package, which is going through rapid and highly publicised development through communication through workshops, conferences and social media. The main limitation of this package is the emphasis on pharmacometrics models and the complex syntax derived from `NONMEM`, whereas `saemix` can use MLXTRAN syntax (which can be seen as a Stan-like syntax). Our focus on `saemix` is to keep the syntax simple and flexible enough to allow statistical developments and fit simple models without the steep learning curve entailed by `nlmixr`, as well as to provide a general tool with less emphasis on pharmacometrics. This grant would allow us to bridge this financial and human gap with `nlmixr`.

Outside of R, software like `Monolix` and `NONMEM` have many features and a solid user-base, however they require licenses to run and cannot be easily tweaked to allow methodological developments to be implemented easily.

6 Figures

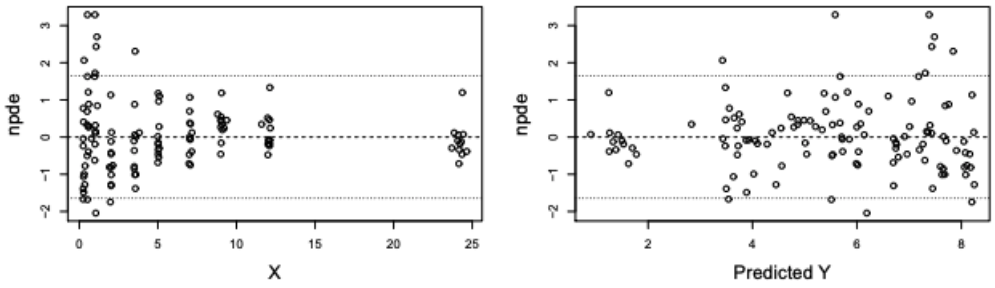

Figure 1: Individual plots for the first 4 subjects in the theophylline study (`plot(saemix.fit, plot.type = "npde")`)

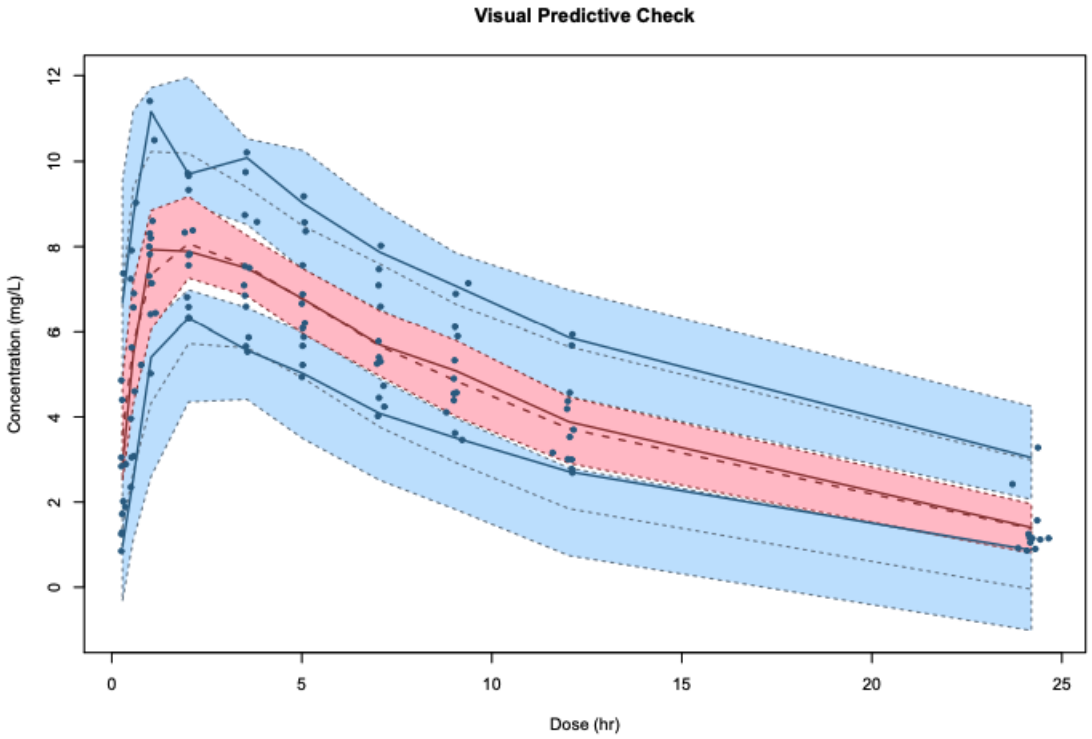

Figure 2: VPC for the theophylline data (`plot(saemix.fit, plot.type = "vpc")`)

References

Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. *Journal of Statistical Software, Articles*, 80(3):1–41.

Delattre, M. and Lavielle, M. (2012). Maximum likelihood estimation in discrete mixed hidden markov models using the saem algorithm. *Computational Statistics & Data Analysis*, 56(6):2073–2085.

Delattre, M., Savic, R. M., Miller, R., Karlsson, M. O., and Lavielle, M. (2012). Analysis of exposure–response of ci-945 in patients with epilepsy: application of novel mixed hidden markov modeling methodology. *Journal of pharmacokinetics and pharmacodynamics*, 39(3):263–271.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128.

Fidler, M., Xiong, Y., Schoemaker, R., Wilkins, J., Trame, M., Hooijmaijers, R., Post, T., and Wang, W. (2019). *nlmixr: Nonlinear Mixed Effects Models in Population Pharmacokinetics and Pharmacodynamics*. R package version 1.1.1-3.

Karimi, B., Lavielle, M., and Moulines, E. (2020). f-saem: A fast stochastic approximation of the em algorithm for nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 141:123–138.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131.

7 Milestones and Deliverables

7.1 Milestones

September 2020: Kick off of the project with all funded staff members.
September 2020: Onboard main new full time developer.
January - June 2021: Two or three seminars across INRIA teams at Ecole Polytechnique, INRIA Bordeaux and INRIA Grenoble.
March - April 2021: New documentation with modern and diverse case studies.
May-June 2021: Onboard intern in statistics and software engineering
June 2021: `saemix` v3.0 published on CRAN.
June 2021: PAGE Conference 2021 (https://www.page-meeting.org).
July 2021: useR! conference.
Aout 2021: 50 000 downloads on CRAN.

7.2 Deliverables

- **Features and Extensions**:
  - Evaluation scripts for new extensions (checking error estimations depending on the design of the study).
  - Full interoperability with the publication of a new package linking `saemix` and `lifoxtConnectors` on CRAN.
  - Implementation of Features and main extensions developed above for the v3.0 publication.
- **Documentation**:
  - Gallery of diversified case studies from PKPD modeling to broader biological processes.
  - Full instructive guidelines files for our package and associated updated reference manual.
- **Visibility and Education**:
  - Online platform for users to upload their datasets and attached notebooks showcasing their analysis using our R package.
  - Updated and easy to use project website.
  - Building a community of users and researchers of our own network with helpful seminars to present our tool.
  - Mixed Effects Modeling with R event through several well established meet-ups such as "Paris Machine Learning".
  - First contributors to the package, outside of the core members, along contribution guidelines.