Instituts thématiques Inserm

Institut national
de la santé et de la recherche médicale

Bordeaux, November 7th 2025.

Development of machine learning based inference methods for high dimensional mechanistic models with population data: application to vaccine trial data analysis.

One internship position is available to work on the development of new inference methods for possibly high dimensional ordinary differential equation models in population setting to analyse data harvested during vaccine trials. It will take place at the Bordeaux Population Health Center, Statistics in the Systems and Translational Medicine team (SISTM) in Bordeaux (France) for a minimum period of 5 months.

The internship will be under the supervision of Quentin Clairon with the SISTM team directed by Dr. Mélanie Prague. The SISTM is a team belonging to INSERM U1219, Bordeaux Population Health and INRIA Bordeaux Sud-Ouest research institutes. The group is dedicated to the analysis and the modelling of the data generated in epidemiology and medicine with a special focus on vaccines and immune interventions in, among others, HIV, EBOLA and more recently SARS-CoV-2. Its expertise is mainly in biostatistics with a special emphasis on dynamical models based on ordinary differential equations (ODEs) and statistical learning using moderately high dimensional data.

Context

For their predictive ability for temporal/counterfactual forecasting and missing variables reconstruction, as well their biological interpretability, mechanistic models are widely used to analyse the evolution of a set of interacting variables. For formally, such models describe the joint evolution of d interacting variables $X := (X_1, \dots X_d)$ (representing for example antibody, cells population, transcriptomic activity etc…) by an ODE: $\dot{X} = f_\theta(X)$ through the specification of the vector field f_θ depending on the d_θ −dimensional parameter θ, often representing biological mechanisms of interest thanks to model interpretability. The statistical analysis of X behaviour in then turned into the problem of θ estimation. In the classical statistical setting, it is done from noisy and discrete observations: $Y_j = X(t_j) + \epsilon_j$ harvested at times $0 \leq t_1 .. \leq t_J$ from the longitudinal follow up of a given subject on an interval $[0, T]$, where the $\{\epsilon_j\}_{j=1,\dots,J}$'s represent the measurement noise, generally assumed to follows centered Gaussian law i.e. $\epsilon_j \sim \mathcal{N}(0, \Sigma_\epsilon)$.
However, two features of vaccine trial data drives us away from this classic statistical setting: 1/**inter-subject variability**, such data aggregate the follow up of many subjects presenting a wide range of response for the different measured variables; 2/**partial measurements**, despite a large number of measured variables, some biological components of particular interest are not accessible; it is the case of several cell populations playing a crucial role in the sustainability and quality of the induced humoral response after vaccination and their evolutions can be at best inferred from surrogate quantities. To account for that, classic ODE models have been integrated into the non-linear mixed effect (NLME) setting, giving rise to so-called **NLME-ODE models**,

Inria inventeurs du monde numérique

which describes the evolution of a whole population of N subjects where the dynamic of the i^{th} one is given by:

$$\begin{cases} Y_{i,j} = h_{\theta_i}\left(X_{\theta_i}(t_j)\right) + \varepsilon_{i,j} \\ \dot{X}_{\theta_i} = f_{\theta_i}(X_{\theta_i}) \\ \theta_i = \theta + b_i \; with \; b_i \sim \mathcal{N}(0, \Sigma_b) \end{cases}$$

embedding the **observation model** $Y_{i,j} = h_{\theta_i}\left(X_{\theta_i}(t_j)\right) + \epsilon_{i,j}$ linking the d_y-dimensional raw data available at the j^{th} measurement time for the i^{th} subject to the d-dimensional variable X_{θ_i} which has an evolution ruled by the **dynamic model** $\dot{X}_{\theta_i} = f_{\theta_i}\left(X_{\theta_i}\right)$ in which the d_θ-dimensional parameter θ_i is prescribed by the **parametric model** $\theta_i = \theta + b_i$ with $b_i \sim \mathcal{N}(0, \Sigma_b)$. In this model, the function $h_{\theta_i}: \mathbb{R}^d \to \mathbb{R}^{d_y}$ emphasize this partial observed setting and the parameter structure account for subject variability thanks to the addition of the so-called random effects b_i. Still, these b_i's are assumed to be the realizations of a common law $b_i \sim \mathcal{N}(0, \Sigma_b)$, to retain some homogeneity at the population level where the variance Σ_b quantifies the allowed discrepancy between subject specific dynamic and the mean population evolution [1-3].

In this context, analyzing the population evolution turns into the problem of estimating the population parameter $\phi := (\theta, \Sigma_b)$. As a classic statistical inference problem, a straightforward approach can be to rely on the most widespread frequentist approach, the maximum likelihood estimator $\hat{\phi}^{ML} := \text{argmax}_\phi \ln\mathbb{P}(Y|\phi)$ where in this case:

$$\ln\mathbb{P}(Y|\phi) = \sum_i \ln \int \mathbb{P}(Y_i|b_i, \phi)\mathcal{N}(0, \Sigma_b)db_i := \sum_i \ln \mathbb{E}_{b_i \sim \mathcal{N}(0,\Sigma_b)}[\mathbb{P}(Y_i|b_i, \phi)]$$

by considering the b_i as unobserved latent variables. So, in the mixed effect setting, the direct maximization of $ln\mathbb{P}(Y|\phi)$ requires repeated computations of d_θ −dimensional integrals which in turn call for computationally efficient approximations for this task [4-5]. Still, they can suffer from accuracy loss depending on the complexity of the mapping $b_i \to X_{\theta+b_i}$. To avoid this, methods based stochastic approximation of EM (SAEM) algorithm get rid of direct $ln\mathbb{P}(Y|\phi)$ maximization to obtain $\hat{\phi}^{ML}$ by iteratively constructing and maximizing lower bounds: $Q^{(l)}(\phi) \leq ln\mathbb{P}(Y|\phi)$ for $l \geq 0$ where $Q^{(l)} := Q\left(\phi\middle|\phi^{(l)}\right)$ and $\phi^{(l)} := \text{argmax}_\phi Q$. Computationally efficient and user-friendly implementations of such algorithms have been proposed which gives more than satisfactory results for model and dataset of reasonable dimensions [4-5-6]. Nonetheless, even with SAEM, likelihood-based inference requires 1/repeated numerical ODE integrations and 2/the estimation of initial conditions $\{X_{\theta_i}(0)\}_{i=1,\dots,N}$ as nuisance parameters. Moreover, 3/the complex nonlinear relationship $(\theta, b_i) \to X_{\theta+b_i}$ leads the likelihood function to have a complex structure, a cause of local minima presence jeopardizing the convergence of used minimization algorithms. This can limit their use to simple models because of the subsequent computational and stability issues.

ODE inference in the one subject setting also suffers from points 1/-2/-3/ and alternative to likelihood have been proposed to circumvent them. Indeed, when ignoring subject variability, (i.e. $b_i = 0$) in a totally observed setting (i.e. $h_\theta(x) = x$), machine learning/statistical methods have been proposed to infer θ for possibly large d and d_θ values. Assuming the access to densely sampled observations $\tilde{X} := (X_\theta(t_1), \dots, X_\theta(t_J))$ as well as their derivatives $\dot{\tilde{X}} = (\dot{X}_\theta(t_1), \dots, \dot{X}_\theta(t_J))$, the SINDy algorithm consider $\hat{\theta} := \text{argmin}_\theta \left\{\left\|\dot{\tilde{X}} - f_\theta(\tilde{X})\right\|_2^2\right\}$ as θ estimator [7]. This method does not require to approximate the ODE solution via numerical procedures (point /1), initial condition estimation (point /2) and the complexity of the inverse problem amount to the one of $\theta \to f_\theta$ rather than $\theta \to X_\theta$ (point /3). More particularly, when considering specific additive vector field structure of the form $f_\theta(x) = \Theta(x)\theta$ (where $\Theta: \mathbb{R}^d \to \mathbb{R}^{d \times d_\theta}$ is a pre-specified so-called "dictionary" function), the optimization problem inherits the same structure as the linear least square criterion and provides a closed-form estimator. Moreover, by integrating θ inference within the linear regression setting, SINDy can incorporate

penalization terms of the form $\lambda\|\theta\|_{Pen}$ for robust estimation when $d_\theta > d$ by considering $\hat{\theta} \coloneqq \text{argmin}_\theta \left\{\left\|\dot{\tilde{X}} - \Theta(\tilde{X})\theta\right\|_2^2 + \lambda\|\theta\|_{Pen}\right\}$. By adopting a LASSO term, SINDy simultaneously provides θ estimation and model structure selection by adaptively enforcing the nullity of some components of θ corresponding to some of entries of $\Theta(x)$. Recent SINDy extensions aim to better handle noise measurements by constructing statistical criteria avoiding the use of $\dot{\tilde{X}}$ by relying on the notion of weak ODE solution stating that: X_θ is a solution of $\dot{X} = \Theta(X)\theta$ on $[0,T]$ in the weak sense if $G_\theta(X_\theta, g) \coloneqq \int_0^T \dot{g}(t)X_\theta(t)dt + \int_0^T g(t)f_\theta(X_\theta(t))dt = 0$ for all differentiable functions g with $g(0) = g(T) = 0$. The original SINDY criterion is then transformed into: $\hat{\theta} \coloneqq \text{argmin}_\theta \left\{\sum_l \left\|G_\theta(\tilde{X}, g_l)\right\|_2^2 + \lambda\|\theta\|_{Pen}\right\}$, based on discretized approximation of $G_\theta(\tilde{X}, g_l)$ and well-chosen candidates functions $\{g_l\}_{l=1,..,L}$ [8-9]. Such methods find their origin in previous statistical works [10-12], in which a non-parametric estimator $\hat{X}$ of X_θ was firstly derived from $\tilde{X}$ to counteract data sparsity and measurement noise before proceeding to θ estimation via the minimization of criteria similar to the previous ones.

Despite sharing some of their goals, our setting does not permit us the direct use of such approaches. They are highly sensitive to sparsity and measurement noise and require a totally observed setting. Moreover, these methods ignore inter-subject variability during θ estimation.

Job Description

The goal of this internship is to extend the previous machine learning approaches to the population setting to infer complex NLME-ODE models and/or for which classic approaches face the difficulties mentioned. For they are more robust to measurement noise and can deal with some partially observed setting, we focus on weak ODE based procedure.

Similarly as the likelihood $\ln\mathbb{P}(Y|\phi)$ replacement by $\sum_l \left\|G_\theta(\tilde{X}, g_l)\right\|_2^2$ as new criterion to optimize in the one subject setting, we can replace $\mathbb{P}(Y_i|b_i, \phi)$ by $\sum_{l=1}^{L} \left\|G_{\theta_i}(\hat{X}, \varphi_l)\right\|_2^2$ in $\mathbb{P}(Y|\phi)$ expression to derive the following alternative objective function: $G_L(\phi) = \sum_i \mathbb{E}_{b_i \sim \mathcal{N}(0,\Sigma_b)} \left[\sum_{l=1}^{L} \left\|G_{\theta+b_i}(\hat{X}_i, \varphi_l)\right\|_2^2\right]$ in the population case where $\hat{X}_i$ is the curve estimator for the i^{th} subject. This allows us to bypass both the need for computations of d_θ −dimensional integrals and ODE numerical approximation, making this inference method a well-suited approach for high dimensional models. Moreover, as in the one subject setting, vector field of the form $f_\theta(x) = \Theta(x)\theta$ may provide a quadratic form for $G_L(\phi)$ with respect to θ for well-chosen random effect structure. This once again ensures a closed form expression for $\hat{\theta}$ but also the possibility to incorporate classic regularization techniques in a high-dimensional setting such as LASSO [13] or sparse partial least squares [14] to account for the effect of correlation between the different dimensions of $\hat{X}_i$ on θ inference.

Two applications cases will be considered:

1) Parameter estimation of NLME-ODE model used for the analysis of humoral response evolution triggered by multi-dose vaccine strategy, notably against SARS-CoV-2 [1]. Its SAEM-based inference may suffer from local minima issues. Since its vector field has a linear dependence with respect to parameters, we aim to use the developed inference procedure to provide a first guess in a computationally efficient manner to bypass such convergence problems.
2) Data driven vector field structure inference for transcriptomic data analysis. Vaccination induces a complex immune response which can be partially analyzed via the longitudinal measurement of transcriptomic activity. Still, such data are often high dimensional and with unknown causal links, such as regulatory activities, between them which makes the construction of structural model in a

priori way for their analysis impractical. We aim to use the developed penalized inference methods to provide an ODE model in a data driven way accounting for inter-subject variability.

Qualifications and Personal Skills:

The candidate seeks for a master's degree in mathematics, physics, computer science or statistics or an engineer school diploma in these fields. We are looking for a highly motivated candidate with an outstanding potential and a strong background in at least dynamical systems or statistics (the formation in the lacking area of expertise will be provided during the internship) and a deep interest in immunology and biological application. Continuation as a PhD student is encouraged.

Proven experience in at least one scientific programming language is required, preferably R and/or Python. The ideal candidates are able to work effectively as part of a team, but also to develop and pursue independent ideas. The successful candidates are expected to conduct innovative research at the highest international level.

Experience in mechanistic modelling and/or applications in biology is recommended and previous work in immunology/vaccinology will be highly appreciated.

The expected starting date can be as soon as possible starting from February 2026. Salary will follow Inserm rates and can be negotiated to be higher depending on previous experience and skills.

The application must include:

- CV summarizing education, positions held, details of academic work, pedagogical and administrative experience and other qualifying activities
- Copies of educational certificates and transcripts of records
- Names and contact details of 2 referees stating relation to candidate, e-mail and telephone number

The application with attachments should be sent to quentin.clairon@u-bordeaux.fr. Foreign applicants are advised to attach an explanation of their University's grading system. Please remember that **all** documents should be in English or French

Applicants may be called in for an interview. In accordance with the University of Bordeaux´s equal opportunities policy, we invite applications from all interested individuals regardless of gender or ethnicity.

Bibliography:

[1] Clairon Q et al. "Modeling the kinetics of the neutralizing antibody response against SARS-CoV-2 variants after several administrations of Bnt162b2." PLoS Computational Biology, 2023.
[2] Thiébaut R et al. "Quantifying and predicting the effect of exogenous interleukin-7 on CD4+ T cells in HIV-1 infection." 2014
[3] Perelson A.S., Ribeiro R.M. "Modeling the within-host dynamics of HIV infection." BMC biology, 2013.
[4] Rackauckas C et al. "Accelerated predictive healthcare analytics with pumas, a high performance pharmaceutical modeling and simulation platform." BioRxiv, 2020.

[5] Keizer R.J, Karlsson M.O, Hooker A. “Modeling and simulation workbench for NONMEM: tutorial on Pirana, PsN, and Xpos.” CPT: pharmacometrics & systems pharmacology, 2013.
[6] Lixoft. “Monolix version 2024R1.” Anthony, France, 2024.
[7] Brunton S.L, Proctor J.L, Kutz J.N. “Discovering governing equations from data by sparse identification of nonlinear dynamical systems.” Proceedings of the national academy of sciences, 2016.
[8] Bortz D.M, Messenger D.A, Dukic V. “Direct estimation of parameters in ODE models using WENDy: Weak-form estimation of nonlinear dynamics.” Bulletin of Mathematical Biology, 2023.
[9] Qian Z, Kacprzyk K., van der Schaar M.. “D-code: Discovering closed-form odes from observed trajectories.” International Conference on Learning Representations, 2022.
[10] Henderson J, Michailidis G. “Network reconstruction using nonparametric additive ODE models.” PloS one, 2014.
[11] Wu H et al. “Sparse additive ordinary differential equations for dynamic gene regulatory network modeling.” Journal of the American Statistical Association, 2014.
[12] Brunel N, Clairon Q, d’Alché-Buc F. “Parametric estimation of ordinary differential equations with orthogonality conditions.” Journal of the American Statistical Association, 2014.
[13] Tibshirani, R. «Regression shrinkage and selection via the lasso.» Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996.
[14] Chun H, Sündüz K. «Sparse partial least squares regression for simultaneous dimension reduction and variable selection.» Journal of the Royal Statistical Society Series B: Statistical Methodology, 2010.

Region:

Bordeaux, Aquitaine, France

Working hours:
Full-time internship

Location:
Centre de recherche Inserm U1219
Université de Bordeaux, ISPED case 11
146 Rue Léo Saignat, 33076 Bordeaux, France
https://www.google.fr/maps/@44.8268226,-0.6030823,16z