ESTIMATION PARAMÉTRIQUE POUR DES MODÈLES MIXTES COMPLEXES À L'AIDE DE MÉTA-MODÈLES

Pierre Barbillon ^{‡,◊} & Célia Barthélémy ^{*} & Adeline Leclercq-Samson [†]

[‡] *AgroParisTech / UMR INRA MIA, F-75005 Paris, France*

[◊] *INRA, UMR 518, F-75005 Paris, France, pierre.barbillon@agroparistech.fr*

^{*} *INRIA Saclay, Popix team, Orsay*

[†] *Laboratoire Jean Kuntzmann CNRS UMR 5224, Université Joseph Fourier, Grenoble, adeline.leclercq-samson@imag.fr*

Résumé. Les processus biologiques sont très souvent mesurés de manière répétée sur un même sujet. Cela donne naturellement lieu à des données longitudinales. Ces données sont généralement analysées grâce à des modèles mixtes qui permettent de discriminer la variabilité inter-sujet de la variabilité intra-sujet. Les fonctions de régression utilisées dans ces modèles intègrent des mécanismes biologiques complexes qui peuvent être des solutions d'équations différentielles ordinaires multi-dimensionnelles ou d'équations aux dérivées partielles. Lorsque ces solutions n'ont pas de forme analytique, on fait appel à des méthodes numériques qui se révèlent très coûteuses en temps de calcul. Les méthodes statistiques couramment utilisées pour estimer les paramètres des modèles mixtes se fondent sur l'utilisation d'une version stochastique de l'algorithme EM (SAEM) que l'on couple avec un algorithme MCMC. Cet algorithme demande un grand nombre d'évaluations des fonctions de régression du modèle ce qui n'est pas praticable si ces fonctions sont coûteuses. C'est pourquoi nous aurons recours à un méta-modèle fondé sur une approximation par processus gaussien des fonctions coûteuses. Nous intégrerons la nouvelle source d'incertitude due à cette approximation dans ce que nous appellerons un méta-modèle mixte. Nous établirons le lien entre la qualité d'approximation du méta-modèle et la proximité entre les estimations obtenues dans le modèle mixte exact et dans le "méta-modèle" mixte. Des simulations numériques seront également proposées pour illustrer la pertinence de cette approche.

Mots-clés. Modèle à effets mixtes, pharmacocinétique, méta-modèle par processus gaussien, algorithme SAEM-MCMC.

Abstract. Complex biological or physiological processes are more and more experimented along time among a collection of individuals. Longitudinal data are then available and the statistical challenge is to analyze these data to better understand the underlying biological mechanisms. The parametric statistical approach commonly used with longitudinal data is mixed-effects model methodology, which allows to discriminate between inter- and intra-subjects variabilities. In these models, the regression functions are now highly-developed to describe precisely biological processes. They may be solutions of

multi-dimensional ordinary differential equations or of partial differential equation. When there is no analytical solution, a classical approach for estimating the parameters relies on the coupling of a stochastic version of the EM algorithm (SAEM) with a MCMC algorithm. This procedure needs many evaluations of the regression function which is clearly prohibitive when a time-consuming solver is used for computing it. That is why we propose to replace this regression function with a meta-model based on a Gaussian process. The new source of uncertainty due to this approximation can be incorporated in the model which leads to what we call a mixed meta-model. We guarantee a control on the distance between the maximum likelihood estimates in this mixed meta-model and the maximum likelihood estimates obtained with the exact mixed model. Eventually, numerical simulations are performed to illustrate the efficiency of this approach.

Keywords. Mixed model, pharmacokinetics, Gaussian process meta-modeling, SAEM-MCMC algorithm.

1 Modèle à effets mixtes

Nous considérons les observations $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^t$ avec $y_{ij} \in \mathbb{R}^p$ étant la réponse pour l'individu i au temps t_{ij} , $i = 1, \dots, N$, $j = 1, \dots, n_i$, et nous notons $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ le vecteur contenant toutes ces observations, il est de taille $n_{tot} = \sum_{i=1}^N n_i$. Nous supposons qu'un vecteur correspondant aux observations longitudinales pour un individu $\mathbf{y}_i$ est décrit par le modèle non linéaire à effets mixtes suivant :

pour $j = 1, \dots, n_i$:

$$\begin{aligned} y_{ij} &= f(t_{ij}, \psi_i) + \sigma_\varepsilon \varepsilon_{ij}, & \varepsilon_{ij} &\sim_{iid} \mathcal{N}(0, 1), \\ \psi_i &\sim_{iid} \mathcal{N}(\mu, \Omega), \end{aligned} \tag{1}$$

avec $f(\cdot, \cdot) : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}^p$ la fonction de régression, ψ_i un vecteur de paramètres individuels de dimension d . Les résidus ε_i sont supposés indépendants des paramètres individuels ψ_i . Les paramètres individuels ne sont pas observés, notre but principal est d'estimer à partir des observations $\mathbf{y}$ les paramètres de leur loi : le vecteur μ d'espérance et la matrice de covariance Ω . Le paramètre σ_ε^2 pourra également être estimé. Nous notons $\theta = (\mu, \Omega, \sigma_\varepsilon^2)$.

Nous cherchons à estimer θ par l'estimateur du maximum de vraisemblance. Elle

s'écrit :

$$\begin{aligned}
p(\mathbf{y}, \theta) &= \int p(\mathbf{y}, \psi; \theta) d\psi = \prod_{i=1}^N \int p(\mathbf{y}_i | \psi_i; \theta) p(\psi_i; \theta) d\psi_i, \\
&= \prod_{i=1}^N \int \frac{1}{(2\pi\sigma_\varepsilon^2)^{n_i/2}} \exp\left(-\frac{1}{2} {}^t(\mathbf{y}_i - \mathbf{f}(\mathbf{t}_i, \psi_i))(\sigma_\varepsilon^2 I_{n_i})^{-1}(\mathbf{y}_i - \mathbf{f}(\mathbf{t}_i, \psi_i))\right) \\
&\quad \times \frac{1}{\sqrt{2\pi|\Omega|}} \exp\left(-\frac{1}{2} {}^t(\psi_i - \mu)\Omega^{-1}(\psi_i - \mu)\right) d\psi_i.
\end{aligned} \tag{2}$$

avec $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$ et $\mathbf{f}(\mathbf{t}_i, \psi_i) = (f(t_{i1}, \psi_i), \dots, f(t_{in_i}, \psi_i))$. Quand f n'est pas linéaire par rapport à ψ , l'estimateur du maximum de vraisemblance n'a pas de forme explicite. Dans ce cadre, une démarche classique est d'avoir recours à une version stochastique de l'algorithme EM à savoir SAEM dans notre cas (Delyon *et al.*, 1999). Les paramètres individuels non observés ψ_i sont simulés conditionnellement aux observations et à la valeur courante du paramètre θ grâce à un algorithme MCMC (Kuhn et Lavielle, 2005).

2 Méta-modèles mixtes

2.1 Méta-modèles par processus gaussien

Afin de construire une approximation pour la fonction f , nous la considérons comme la réalisation d'un processus gaussien F_λ (Sacks *et al.*, 1989). Dans cette partie, le couple d'entrées de $f(t, \psi)$ est noté x . Pour tout x , nous définissons

$$F_\lambda(x) = \sum_{j=1}^L \beta_j h_j(x) + \zeta(x) = {}^t H(x) \boldsymbol{\beta} + \zeta(x), \tag{3}$$

avec $h_1, \dots, h_L$ des fonctions basiques connues (des fonctions linéaires par exemple), $H = (h_1, \dots, h_L)$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_L)$ un vecteur de paramètres, ζ un processus gaussien ($\mathcal{GP}$) centré de noyau de covariance : $\text{Cov}(\zeta(x), \zeta(x')) = \sigma^2 K^\phi(x, x')$. Afin de simplifier les calculs, nous considérerons le vecteur $\lambda = (\boldsymbol{\beta}, \sigma, \phi)$ comme fixe bien qu'il soit en pratique estimé grâce aux données.

Nous supposons que nous pouvons effectuer une étape de pré-calculs où nous évaluons f en certaines entrées choisies selon un plan d'expérience numérique D (Fang *et al.*, 2005). Nous obtenons alors des évaluations non bruitées de f et nous conditionnons le processus gaussien F_λ à les interpoler. Le processus conditionné est encore un processus gaussien pour lequel nous avons une forme explicite pour l'espérance $m_D(\cdot)$ et la covariance $C_D(\cdot, \cdot)$. L'espérance m_D interpole aux entrées du plan d'expérience les évaluations de f et l'approche partout ailleurs. La variance permet d'encoder l'incertitude sur cette

approximation (nous pouvons remarquer que cette variance est nulle aux points où f a été évaluée). Une fois que cette étape a été réalisée, nous utilisons uniquement m_D et C_D dans l’algorithme SAEM donc nous n’aurons plus besoin de nouveaux appels à f ce qui rendra l’algorithme praticable. Nous faisons remarquer que cette approche a un parallèle en théorie de l’approximation. Dans ce cadre, f est considérée comme appartenant à un espace à noyau reproduisant (RKHS) et pour certains noyaux il est possible d’obtenir sur un domaine compact une borne uniforme sur l’erreur d’approximation (Schaback, 1995). Ce résultat sera utile pour montrer le contrôle sur l’écart entre les estimateurs du maximum de vraisemblance dans le modèle mixte exact et dans un “méta-modèle” mixte.

2.2 Trois méta-modèles mixtes

Nous pouvons alors introduire naturellement le méta-modèle dans le modèle mixte (1) en remplaçant f par:

$$f_D(x) = m_D(x) + r(x), \quad r(x) \sim \mathcal{GP}(0, C_D(x, x)).$$

Ainsi l’erreur d’approximation sera incluse dans ce modèle que nous nommons méta-modèle mixte complet. Il a cependant le défaut de ne plus appartenir à la famille exponentielle si le paramètre σ_ε^2 est inconnu. De plus, les erreurs induites par le terme $r(x)$ ne sont plus indépendantes à cause de la structure de processus gaussien. Cela empêche de séparer la vraisemblance en un produit sur les individus et de simuler indépendamment les uns des autres les paramètres individuels ψ_i . Le calcul de la vraisemblance nécessitera d’inverser une matrice de covariance non creuse ce qui alourdira considérablement l’étape de simulation par MCMC dans l’algorithme SAEM.

Afin de rendre la vraisemblance séparable par rapport aux individus et plus rapide à calculer, nous proposons le méta-modèle mixte intermédiaire qui consiste à imposer une matrice de covariance diagonale (on ne garde que les variances issues du processus gaussien $r(\cdot)$ et on fixe à 0 les covariances). Nous espérons cette démarche conservatrice car en négligeant des covariances positives, on augmente l’incertitude dans le modèle. Par contre, ce modèle n’est toujours pas dans la famille exponentielle dans le cas où σ_ε^2 est inconnue.

Finalement, nous proposons le méta-modèle simplifié qui est le plus simple à traiter car il néglige l’erreur d’approximation et remplace f par m_D dans le modèle (1). Ainsi ce modèle appartient bien à la famille exponentielle.

3 Convergence des estimations

Puisque nous utilisons l’algorithme SAEM dans le cadre d’un modèle approché, il n’est pas possible de montrer la convergence de cet algorithme vers l’estimateur du maximum de la

vraisemblance du modèle exact (2). Par contre, en nous plaçant sous les mêmes hypothèses que Kuhn et Lavielle (2005), nous pouvons montrer la convergence de l'algorithme SAEM vers l'estimateur du maximum de la vraisemblance correspondant au méta-modèle mixte considéré.

À défaut de pouvoir prouver la convergence vers l'estimateur du maximum de la vraisemblance (2), notre but est d'obtenir un contrôle uniforme sur la distance entre cette vraisemblance et la vraisemblance d'un méta-modèle mixte. Nous donnons la proposition pour le cas du méta-modèle simplifié où nous notons $\tilde{p}_D(\mathbf{y}; \boldsymbol{\theta})$ sa vraisemblance. Des résultats similaires peuvent être obtenus dans les deux autres cas. Ce contrôle est fonction de la qualité du méta-modèle qui dépend de la couverture par le plan d'expérience du domaine des paramètres individuels $\boldsymbol{\psi}$.

Proposition 1 *Considérons les vraisemblances $p(\mathbf{y}; \boldsymbol{\theta})$ du modèle mixte exact (1) et $\tilde{p}_D(\mathbf{y}; \boldsymbol{\theta})$ du méta-modèle mixte simplifié où le méta-modèle a été construit à partir d'un plan d'expérience minimax (Johnson et al., 1990) et d'un noyau K vérifiant les propriétés données dans Schaback (1995). Si on suppose la loi de $\boldsymbol{\psi}$ à support compact et que les fonctions f et m_D sont uniformément bornées sur ce support, alors il existe une constante $\tilde{C}_y$ dépendant seulement de $\mathbf{y}$ telle que*

$$|p(\mathbf{y}; \boldsymbol{\theta}) - \tilde{p}_D(\mathbf{y}; \boldsymbol{\theta})| \leq \tilde{C}_y \frac{n_{tot}}{\sigma_\varepsilon^{n_{tot}+2}} G_K(a_D),$$

où $G_K(a)$ tend vers 0 quand $a \rightarrow 0$ et la constante a_D est la distance de recouvrement par D du support de la loi de $\boldsymbol{\psi}$.

Nous avons bien besoin de supposer que le support de la loi de $\boldsymbol{\psi}$ est borné pour assurer qu'un plan d'expérience suffisamment riche en nombre de points implique une distance de recouvrement aussi petite que souhaité. Ainsi nous avons bel et bien un contrôle uniforme sur la distance entre les deux vraisemblances.

Ensuite avec des hypothèses de régularité sur les matrices hessiennes associées aux vraisemblances de chaque modèle, nous pouvons garantir des résultats similaires à ceux de Donnet et Samson (2007).

4 Simulations

Nous comparerons les performances des algorithmes SAEM-MCMC sur les différents méta-modèles mixtes à partir de simulations sur des modèles de pharmacocinétique. Nous utiliserons des modèles où le coût computationnel n'est pas encore rédhibitoire afin d'avoir une référence à laquelle se rapporter. Nous testerons différentes tailles de plan d'expériences et distinguerons les cas où le paramètre de bruit résiduel σ_ε^2 est connu et fixé des cas où il doit être estimé.

Bibliographie

- DELYON, B., LAVIELLE, M. et MOULINES, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27:94–128.
- DONNET, S. et SAMSON, A. (2007). Estimation of parameters in incomplete data models defined by dynamical systems. *J. Statist. Plann. Inference*, 137:2815–2831.
- FANG, K., LI, R. et SUDJIANTO, A. (2005). *Design and Modeling for Computer Experiments (Computer Science & Data Analysis)*. Chapman & Hall/CRC.
- JOHNSON, M. E., MOORE, L. M. et YLVIKAKER, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131 – 148.
- KUHN, E. et LAVIELLE, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, 49:1020–1038.
- SACKS, J., SCHILLER, S. B. et WELCH, W. J. (1989). Designs for computer experiments. *Technometrics*, 31(1):41–47.
- SCHABACK, R. (1995). Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264.