Analyse bayésienne de courbes de croissance par des modèles à effets mixtes définis par équations différentielles stochastiques.

Sophie Donnet, Jean-Louis Foulley, Adeline Samson

ANALYSE BAYÉSIENNE DE COURBES DE CROISSANCE PAR DES MODÈLES À EFFETS MIXTES DÉFINIS PAR ÉQUATIONS DIFFÉRENTIELLES STOCHASTIQUES.

Sophie Donnet[1] & Jean-Louis Foulley[2] & Adeline Samson[3]

[1] *sophie.donnet@ceremade.dauphine.fr, Cérémade, Université Paris Dauphine, France*
[2] *jean-louis.foulley@jouy.inra.fr, INRA, UMR GABI, CR Jouy, France,*
[3] *adeline.samson@parisdescartes.fr, Laboratoire MAP5, Université Paris Descartes, France*

Résumé
On désigne par courbes de croissance des mesures répétées au cours du temps d'un processus continu de croissance sur une population d'individus. Ces données longitudinales sont classiquement analysées par des modèles non-linéaires à effets mixtes dont la fonction de régression impose une évolution monotone croissante du phénomène. Ces modèles de croissance ne permettent pas de modéliser des modifications inattendues du taux de croissance. Nous proposons de prendre en compte ces éventuelles variations à l'aide d'équations différentielles stochastiques déduites du modèle de croissance standard par ajout d'une composante stochastique. Nous développons une méthode d'inférence Bayésienne de ces modèles reposant sur un algorithme de Gibbs et validons ce nouveau modèle en utilisant et en adaptant des critères basés sur la distribution prédictive a posteriori. Nous illustrons la pertinence de notre approche dans le cas d'un modèle de Gompertz sur un jeu de données réelles de croissance de poulets.

Mot-Clefs
Courbes de croissance; Distribution prédictive a posteriori; Equation différentielle stochastique; Estimation bayésienne; Modèle de Gompertz; Modèles mixtes;

Abstract
Growth curve data consist of repeated measurements of a continuous growth process over time among a population of individuals. These data are classicaly analysed by nonlinear mixed models. The standard growth functions used in this context prescribe monotone increasing growth and can fail to model unexpected changes in growth rates. We propose to model these variations using stochastic differential equations (SDEs) which are deduced from the standard deterministic growth function by adding random variations to the growth dynamics. A Bayesian inference of parameters of these SDE mixed models is developed, relying on a Gibbs algorithm. We suggest to validate the SDE approach via criteria based on the predictive posterior distribution of replicate samples and a chisquare discrepancy function. We illustrate the efficiency of our method in the particular case of the Gompertz function to model data on chichen growth, the modeling being improved by the SDE approach.

Key-words
Bayesian estimation; Gompertz model; Growth curves; Mixed models; Predictive Posterior distribution; Stochastic differential equation

1 Introduction

On désigne par courbes de croissance des mesures répétées au cours du temps d'un processus continu de croissance sur une population d'individus. En agronomie ces données de croissance permettent par exemple de différencier des phénotypes animaux ou végétaux en fonction de la dynamique du processus biologique sous-jacent. Ces données sont classiquement analysées par des modèles non-linéaires à effets mixtes dont la fonction de régression appartient à la famille classique des fonctions de croissance, telles que les fonctions de Gompertz, logistique, de Richards ou de Weibull (Zimmerman & Núnez-Antón (2001)). Ces modèles imposent une évolution croissante du phénomène observé, quelle que soit la valeur des paramètres. Bien que ces modèles aient prouvé leur pertinence, ils ne permettent pas de prendre en compte les variations inattendues du taux de croissance telles que le ralentissement de la croissance ou même des décroissances. Ces phénomènes observés ne sont pas dus à des erreurs de mesures mais bien à des phénomènes biologiques sous-jacents non-expliqués.
Dans ce papier, nous cherchons à modéliser ces variations de processus de croissance au moyen d'équations différentielles stochastiques (EDS). En effet, les courbes de croissance standard sont solutions d'une équation différentielle ordinaire (EDO). Nous proposons d'introduire une composante stochastique dans cette équation. Au final, le processus de croissance varie aléatoirement autour d'une dynamique moyenne.
L'estimation bayésienne de modèles mixtes définis par EDS a été peu abordée dans la littérature. Cano et al. (2006) calculent la loi a posteriori des paramètres en approchant la solution de l'EDS par un schéma d'Euler-Maruyama alors que Oravecz et al. (2008) se concentrent sur le cas du processus d'Ornstein-Uhlenbeck avec paramètres aléatoires. Dans ce papier, nous proposons une méthode d'inférence bayésienne dans le cas des courbes de croissance. Dans la partie 2, nous présentons le modèle de croissance à effets mixtes classique ainsi que le modèle défini par EDS. Nous discutons notamment du choix du terme de volatilité. Au paragraphe 3, nous spécifions les lois a priori sur les paramètres et développons un algorithme de Gibbs pour estimer les lois a posteriori des paramètres. Une démarche de validation bayésienne du modèle est en outre proposée. Finalement (partie 4), nous illustrons la pertinence de notre modèle sur des données de croissance de poulets dans le cas d'une fonction de croissance de Gompertz.

2 Modèles et notations

Soient $\boldsymbol{y} = (y_i)_{1\leq i\leq n} = (y_{ij})_{1\leq i\leq n, 1\leq j\leq n_i}$ les observations du processus de croissance où y_{ij} est l'observation bruitée du sujet i à l'instant t_{ij} ($i = 1 \dots n$, $j = 0 \dots n_i$).

Modèles de croissance à effets mixtes

Dans les modèles mixtes classiques, l'évolution du processus est décrite par une fonction déterministe dépendant de paramètres aléatoires propres à l'individu. Plus précisément,

$$\begin{aligned} y_{ij} &= f(\phi_i, t_{ij}) + \varepsilon_{ij},\ \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2) \\ \phi_i &\sim \mathcal{N}(\mu, \Omega) \end{aligned} \tag{1}$$

où f est une fonction déterministe paramétrique; $\phi = (\phi_i)_{1\leq i\leq n} \in \mathbb{R}^p$ sont les paramètres individuels; les ε_{ij} sont les erreurs résiduelles.
Dans le cas des courbes de croissance, f est classiquement l'une des quatre fonctions suivantes: fonctions logistique, de Gompertz, de Richards et de Weibull. Chacune d'entre elles peut-être écrite comme la solution d'une EDO dont la solution est strictement croissante (voir Zimmerman & Núnez-Antón (2001)).
Exemple. La fonction de Gompertz est solution de l'EDO $f'_{\mathcal{G}}(t) = BCe^{-Ct} f_{\mathcal{G}}(t)$ avec pour condition initiale $f_{\mathcal{G}}(0) = Ae^{-B}$. Ainsi $f_{\mathcal{G}}(t) = A\exp\left[-Be^{-Ct}\right]$.
Plus généralement, soit ϕ l'ensemble des paramètres (A,B,C) ou une re-paramétrisation judicieuse des paramètres, alors f est solution de l'EDO générale suivante:

$$\frac{\partial f(\phi,t)}{\partial t} = F(f,t,\phi), \quad f(\phi,0) = f_0(\phi) \tag{2}$$

Modèles de croissance à effets mixtes définis par EDS
Afin de prendre en compte les variations aléatoires du processus de croissance, nous introduisons un terme de volatilité stochastique dans l'EDO (2). Le processus de croissance est alors décrit par l'EDS suivante:

$$dZ_t = F(Z_t,t,\phi)dt + \Gamma(Z_t,\phi,\gamma^2)dW_t, \quad Z(t=0) = Z_0(\phi)$$

où W_t est un processus browien et $\Gamma(Z_t,\phi,\gamma^2)$ est la fonction de volatilité paramétrée par γ^2 inconnu. Finalement, le modèle de croissance à effets mixtes défini par EDS s'écrit:

$$\begin{aligned} y_{ij} &= Z_{t_{ij}}(\phi_i) + \varepsilon_{ij},\ \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0,\sigma^2) \\ dZ_t(\phi_i) &= F(Z_t,t,\phi_i)dt + \Gamma(Z_t,\phi_i,\gamma^2)dW_t \\ \phi_i &\sim \mathcal{N}(\mu,\Omega) \end{aligned} \tag{3}$$

Le modèle (3) intègre trois sources de variabilités distinctes: la variabilité inter-sujets quantifiée par Ω, la variabilité γ^2 reflétant les variations aléatoires du processus autour du modèle théorique f et l'erreur de mesure σ^2. Par ailleurs, le choix de la fonction de volatilité est crucial et doit découler de multiples considérations: positivité ou non du phénomène observé, type d'aléa à introduire, existence ou non d'une solution explicite de l'EDS, etc.

Exemple. Dans le cas de la Gompertz – prenant en compte le fait qu'une analyse antérieure de nos données réelles a montré qu'un modèle de bruit hétéroscédastique était pertinent– nous proposons d'introduire une fonction de volatilité polynomiale:

$$dZ_t = BCe^{-Ct}Z_t dt + \gamma Z_t dW_t,\ Z_0 = Ae^{-B} \tag{4}$$

De cette façon, l'équation (4) a une solution explicite qui s'écrit comme une perturbation aléatoire multiplicative de la fonction de Gompertz standard: $Z_t = f_{\mathcal{G}}(t)e^{-\frac{1}{2}\gamma^2 t+\eta_t}$ où $\eta_t \sim$

$\mathcal{N}(0, \gamma^2 t)$. De plus, nous garantissons entre autre la positivité presque sûre du phénomène. Nous obtenons alors le modèle mixte global suivant: $\forall i = 1 \dots n$,

$$\begin{aligned}
(\log y_{i0}, \log y_{i1}, \dots, \log y_{in_i})' &= \left(\log(A_i) - B_i, \log Z_{t_{i1}}, \dots, \log Z_{t_{in_i}}\right)' + \varepsilon_i, \\
\varepsilon_i &\sim_{i.i.d.} \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}_{n_i+1}\right) \\
\left(\log Z_{t_{i1}}, \dots, \log Z_{t_{in_i}}\right)' &= \log(A_i) - B_i \left(e^{-C_i t_{i1}}, \dots, e^{-C_i t_{in_i}}\right)' - \gamma^2 \left(t_{i1}, \dots, t_{in_i}\right)' + \eta_i, \\
\eta_i &\sim_{i.i.d} \mathcal{N}\left(0_{n_i}, \gamma^2 T_i\right), \text{ avec } T_i = (\min(t_{ij}, t_{ij'}))_{1 \leq j, j' \leq n_i} \\
(\log A_i, B_i, \log C_i) &\sim_{i.i.d.} \mathcal{N}(\mu, \Omega)
\end{aligned}$$

3 Inférence bayésienne

Dans le cadre d'une approche bayésienne, nous cherchons à évaluer la distribution a posteriori des paramètres de population μ, Ω ainsi que de σ^2 et enfin de γ^2 dans le cas du modèle par EDS.

Spécification des lois a priori: Nous suggérons d'utiliser des loi a priori standard (c.f. De la Cruz-Mesia & Marshall (2006)) pour les espérances et variances dans les modèles hiérarchiques, i.e. $\forall k = 1 \dots p, \mu_k \sim \mathcal{N}(m_k^{prior}, v_k^{prior})$, $\Omega^{-1} \sim W(R, p+1)$ (distribution de Wishart) $1/\sigma^2 \sim \Gamma(\alpha_\sigma^{prior}, \beta_\sigma^{prior})$. γ^2 étant un paramètre contrôlant la variance des perturbations aléatoires du processus, il est raisonnable de choisir aussi une loi a priori inverse-Gamma pour cette quantité: $1/\gamma^2 \sim \Gamma(\alpha_\gamma^{prior}, \beta_\gamma^{prior})$. Le choix des paramètres de ces lois a priori peut s'avérer délicat d'où notre souci de les calibrer notamment dans une optique de robustesse des résultats obtenus sur les paramètres d'intérêt.

Lois a posteriori: Les modèles (1) et (3) étant non-linéaires, nous avons recours à une procédure itérative du type algorithme Monte Carlo Markov Chain (MCMC) pour étudier la loi a posteriori des paramètres. Dans le cas du modèle de croissance standard (1), l'algorithme de Gibbs à mettre en place ne présente pas de difficulté particulière (voir par exemple Carlin & Louis (2000)). Dans le cas du modèle défini par EDS (3) nous proposons l'algorithme suivant:

- ETAPE 1: initialisation sur des valeurs initiales $\sigma^{-2(0)}, \gamma^{2(0)}, \mu^{(0)}, \boldsymbol{\phi}^{(0)}, \boldsymbol{Z}^{(0)}$.
- ETAPE 2: génération de $\sigma^{-2(k)}, \gamma^{2(k)}, \mu^{(k)}, \boldsymbol{\phi}^{(k)}, \boldsymbol{Z}^{(k)}$ à partir de $\sigma^{-2(k-1)}$, $\gamma^{2(k-1)}$, $\mu^{(k-1)}$, $\boldsymbol{\phi}^{(k-1)}$, $\boldsymbol{Z}^{(k-1)}$ au travers des étapes successives suivantes:
  1. $\boldsymbol{Z}^{(k)} \sim p(\boldsymbol{Z} | \boldsymbol{\phi}^{(k-1)}, \gamma^{-2(k-1)}, \sigma^{-2(k-1)}, \boldsymbol{y})$
  2. $\boldsymbol{\phi}^{(k)} \sim p(\boldsymbol{\phi} | \sigma^{-2(k-1)}, \gamma^{-2(k-1)}, \mu^{(k-1)}, \Omega^{(k-1)}, \boldsymbol{Z}^{(k)}, \boldsymbol{y}_0)$ où $\boldsymbol{y}_0 = (y_{i0})_{i=1\dots n}$
  3. $\mu^{(k)} \sim p(\mu | \boldsymbol{\phi}^{(k)})$ et $\Omega^{(k)} \sim p(\Omega | \boldsymbol{\phi}^{(k)})$
  4. $\sigma^{-2(k)} \sim p(\sigma^{-2} | \boldsymbol{Z}^{(k)}, \boldsymbol{\phi}^{(k)}, \boldsymbol{y})$ et $\gamma^{-2(k)} \sim p(\gamma^{-2} | \boldsymbol{Z}^{(k)}, \boldsymbol{\phi}^{(k)})$
- ETAPE 3: passer de k à $k+1$ et retourner à l'ETAPE 2 jusqu'à convergence.

où Z_i est une réalisation du processus (Z_t) pour chaque individu et à tous les instants d'obervation. $\boldsymbol{Z} = (Z_1, \ldots, Z_n) \in \boldsymbol{R}^{n_1+\cdots+n_n}$ représente le vecteur des n individus.
Certaines des lois a posteriori conditionnelles sont explicites. Ainsi, la loi a priori sur σ^{-2} étant une loi Gamma, la loi a posteriori $p(\sigma^{-2}|\boldsymbol{Z}^{(k)}, \boldsymbol{\phi}^{(k)}, \boldsymbol{y})$ l'est aussi. En outre, $p(\boldsymbol{\phi}|\mu, \Omega)$ étant gaussienne, la loi conditionnelle de μ est gaussienne et celle de Ω est une inverse Wishart. Pour les autres quantités, les lois conditionnelles sont propres au modèle.

Exemple: Dans le cas du modèle Gompertz, la loi conditionnelle de $\log \boldsymbol{Z_i} = (\log Z_{ij})_{1 \leq j \leq n_i}$ conditionnellement à $(\phi_i, \gamma^{-2}, \boldsymbol{y_i}, \sigma^2)$ est gaussienne: $\log \boldsymbol{Z}_i|\boldsymbol{y_i}, \sigma^2, \gamma^2, \phi_i \sim \mathcal{N}(m^{post}_{\log \boldsymbol{Z}_i}, V^{post}_{\log \boldsymbol{Z}_i})$ avec

$$\begin{aligned} V^{post}_{\log \boldsymbol{Z}_i} &= (\sigma^{-2} I_{n_i} + \gamma^{-2} T_i^{-1})^{-1}, \\ m^{post}_{\log \boldsymbol{Z}_i} &= V^{post}_{\log \boldsymbol{Z}_i} \left[\sigma^{-2} (\log y_{i1} \ldots \log y_{in_i})' + \gamma^{-2} T_i^{-1} u_{\log \boldsymbol{Z}_i}\right] \\ u_{\log \boldsymbol{Z}_i} &= \log A_i - B_i \left(e^{-C_i t_{i1}} \ldots e^{-C_i t_{in_i}}\right)' - \frac{1}{2}\gamma^2 \left(t_{i1} \ldots t_{in_i}\right)' \end{aligned}$$

Remarque. Dans le cas général, lorsque la distribution conditionnelle du processus $\boldsymbol{Z}$ n'est pas explicite, une solution envisageable est d'avoir recours à un schéma d'Euler-Maruyama, approchant les probabilités de transition par des lois gaussiennes.

Critère Bayésien de validation du modèle EDS: nous cherchons à valider le modèle défini par EDS en utilisant les distributions prédictives a posteriori. Cette méthode repose sur la génération de données répliquées sous la loi prédictive a posteriori. Ces données répliquées sont comparées aux données observées au travers de la fonction de discrépance que nous choisissons du type χ^2 dans notre cas: $T(\boldsymbol{y}, \eta) = \frac{(\boldsymbol{y}-\eta)^2}{Var(\boldsymbol{y}-\eta)}$ où $\eta = f(\phi, t_{ij})$ pour le modèle de croissance déterministe et $\eta = Z_{ij}(\phi_i)$ pour le modèle défini par EDS. Cette fonction vise à quantifier l'adéquation du modèle avec les données observées. Nous cherchons à comparer la distribution à posteriori de $p(T(\boldsymbol{y}, \eta)|\boldsymbol{y})$ avec celle de $p(T(\boldsymbol{y}^{rep}, \eta)|\boldsymbol{y})$ où $\boldsymbol{y}^{rep}$ représente les données répliquées sous la loi $p(\boldsymbol{y}^{rep}|\boldsymbol{y})$ au travers de la probabilité $p_{pp} = \int P\left[T(\boldsymbol{y}^{rep}, \eta) > T(\boldsymbol{y}, \eta)|\boldsymbol{y}, \eta\right] p(\eta|\boldsymbol{y}) d\eta$. En pratique, pour chaque modèle, l'algorithme de Gibbs fournit des η^l $(l = 1 \ldots L)$ générés sous la loi a posteriori $p(\eta|\boldsymbol{y})$. A partir de chaque η^l nous générons des données répliquées sous la loi $p(\boldsymbol{y}^{rep}|\eta^l)$. Finalement, la quantité p_{pp} est approchée par $\frac{1}{L}\sum_{l=1}^{L} 1_{T(\boldsymbol{y}^{rep\ l}, \eta^l) > T(\boldsymbol{y}, \eta^l)}$ et comparée à $\frac{1}{2}$.

4 Application sur données réelles

Nous nous intéressons à la modélisation de croissances de poulets, jeu de données précédemment analysé par Jaffrézic et al. (2006) et Meza et al (2007). Les données $\boldsymbol{y}$ sont des observations bruitées de poids de n =50 poulets, mesurés à t =0, 4, 6, 8, 12, 16, 20, 24, 28, 32, 36, 40 jours après la naissance (voir exemple sur la figure 1). Nous estimons les paramètres des deux modèles par les algorithmes de Gibbs précédemment décrits. Nous présentons –sur la figure (1)

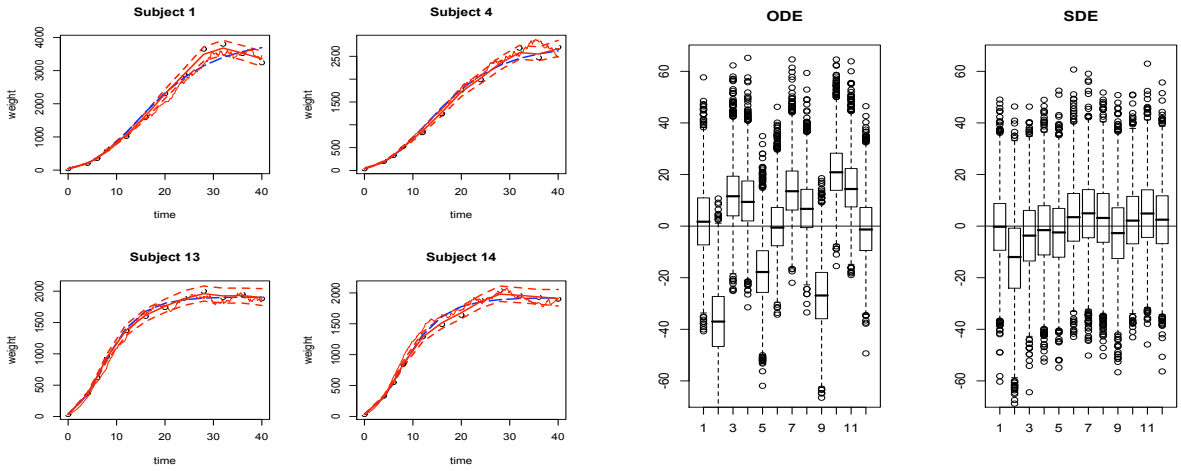

Figure 1: *A gauche*: Données et prédictions pour les sujets 1, 4 13 and 14. *Au milieu et à droite*: PPP pour les modèles standard et EDS

à gauche– les observations pour 4 individus ($\circ$), la trajectoire prédite par le modèle déterministe (tirets longs), la trajectoire moyenne (sur 1000 réalisations) fournie par le modèle EDS (trait plein), un exemple de trajectoire fourni par le modèle EDS ainsi qu'un intervalle de confiance prédictif à 95% (tirets courts). Les sujets 4 et 13 sont des individus sans ralentissement de croissance, sur lesquels les 2 modèles fournissent des résultats comparables. Le sujet 14 présente une légère décroissance alors que le sujet 1 présente une nette décroissance. Dans ces deux cas, le modèle EDS réussit à modéliser ces phénomènes, ce qui n'est pas le cas du modèle standard. Nous calculons ensuite le critère de validation de modèle. La figure (1) à droite présente un résumé des distributions prédictives a posteriori sous chacun des modèles. Ce graphique met clairement en évidence l'amélioration fournie par ce nouveau modèle.

Références

[1] Cano, J., Kessler, M., and SalmerÛn, D. (2006), *Approximation of the posterior density for diffusion processes*, Statistics & Probability Letters **76,** 39–44.

[2] Carlin, B. P. and Louis, T. A. (2000).*Bayes and empirical Bayes methods for data analysis*, volume 69 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

[3] De la Cruz-Mesia, R. and Marshall, G. (2006).*Non-linear random effects models with continuous time autoregressive errors: a Bayesian approach. Statistics in Medicine* **25,** 1471–1484.

[4] Donnet S., Foulley J.-L. and Samson A. (2009).*Bayesian analysis of growth curves using mixed models defined by stochastic differential equations Soumis*.

[5] Jaffrézic, F., Meza, C., Lavielle, M., and Foulley, J.L. (2006). *Genetic analysis of growth curves using the SAEM algorithm. Genetics Selection Evolution* **38,** 583–600.

[5] Meza, C., Jaffrézic, F., and Foulley J.-L. (2007). *REML estimation of variance parameters in nonlinear mixed effects models using the SAEM algorithm*, *Biometrical Journal*, **6**, 876-888

[6] Oravecz, Z., Tuerlinckx, F., and Vandekerckhove, J. (in press). *A hierarchical Ornstein-Uhlenbeck model for continuous repeated measurement data. Psychometrika*.

[7] Zimmerman, D. and Núnez-Antón, V. (2001). *Parametric modelling of growth curve data: an overview. Test* **10,** 1–73.