Bayesian analysis of growth curves using mixed models defined by stochastic differential equations

Sophie Donnet[1]

Ceremade, Universite Dauphine, France

**email:* sophie.donnet@ceremade.dauphine.fr

and

Jean-Louis Foulley[2]

INRA-GABI, PSGen, CR Jouy, France

**email:* jean-louis.foulley@jouy.inra.fr

and

Adeline Samson[3]

Laboratoire MAP5, Universite Paris Descartes, France

**email:* adeline.samson@parisdescartes.fr

SUMMARY: Growth curve data consist of repeated measurements of a continuous growth process over time in a population of individuals. These data are classically analyzed by nonlinear mixed models. However, 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) that are deduced from the standard deterministic growth function by adding random variations to the growth dynamics. A Bayesian inference of the parameters of these SDE mixed models is developed. In the case when the SDE has an explicit solution, we describe an easily implemented Gibbs algorithm. When the conditional distribution of the diffusion process has no explicit form, we propose to approximate it using the Euler-Maruyama scheme. Finally, we suggest to validate the SDE approach via criteria based on the predictive posterior distribution. We illustrate the efficiency of our method using the Gompertz function to model data on chicken growth, the modeling being improved by the SDE approach.

Biometrics –, 1–??

DOI: –

– –

Key words: Bayesian estimation; Euler-Maruyama scheme; Gompertz model; Growth curves; Mixed models; Predictive posterior distribution; Stochastic differential equation

1. Introduction

Growth curve data consist of repeated measurements of a growth process over time among a population of individuals. In agronomy, growth data allow differentiating animal or vegetal phenotypes by characterizing the dynamics of the underlying biological process. In gynecology or pediatrics, height and weight of children are regularly recorded to control their development. The parametric statistical approach used to analyze these longitudinal data is mixed model methodology (Huggins and Loesch, 1998). The regression function of this mixed model is a parametric growth function, such as the Gompertz, logistic, Richards or Weibull functions (Zimmerman and Núnez-Antón, 2001) which prescribe monotone increasing growth, whatever the parameter values. These models have proved their efficiency in animal genetics (Hou et al., 2005; Jaffrézic et al., 2006, *e.g.)* and in pediactrics (Spyrides et al., 2008, *e.g.)*. However, as pointed out by Davidian and Giltinan (2003), the used function may not capture the exact process, as responses for some individuals may display some local fluctuations such as weight decreases or growth slow down. These phenomena are not due to error measurements but are induced by an underlying biological process that is still unknown today. In animal genetics, a wrong modeling of these curves could affect the genetic analysis. In fetal growth, the detection of growth slow down is a crucial indicator of fetal development problems. This paper aims to model these variations in growth rate using a stochastic differential equation (SDE) whose solution is the regression term of the mixed model. More precisely, each growth function is defined as the solution of an ordinary differential equation (ODE). We suggest to add a random perturbation to the ODE, resulting in an SDE. Thus, the growth rate varies randomly around the mean dynamics. In this paper, we propose and study Bayesian estimators for mixed models defined by SDEs.

Parametric estimation by maximum likelihood of SDE with random parameters (without measurement noise) has been studied by Ditlevsen and De Gaetano (2005). However, es-

timation of SDE mixed models (including the measurement noise modeling) has received little attention. Overgaard et al. (2005) and Tornoe et al. (2005) proposed estimators based on an extended Kalman filter, but the algorithm convergence was not proved. Donnet and Samson (2008) proposed an EM-based estimator and proved the convergence of their algorithm. Whereas the Bayesian point of view is widely used on standard growth curves, Bayesian estimation of SDE mixed models has not been much investigated. Cano et al. (2006) computed the posterior distribution by approximating the diffusion process by an Euler scheme. Oravecz et al. (in press) studied the Bayesian estimation of an Ornstein-Uhlenbeck process with random parameters. In this paper, we propose either to use a judicious transformation of the SDE to compute the exact conditional distribution of the diffusion process, or, if it is not possible, to approximate the diffusion by the Euler-Maruyama scheme. Then we propose a Gibbs algorithm to simulate the exact or the approximate posterior distributions. In the case of approximation by the Euler scheme, we control the error induced by this scheme on the posterior distributions. Finally, we adapt the computation of the posterior predictive distributions to validate the SDE mixed model (Meng, 1994).

Section 2 presents the classical mixed model and the mixed model defined by SDEs. We discuss the choice of the volatility in the SDEs. In Section 3, we suggest some prior specifications and posterior computation and present the Euler-Maruyama scheme. Section 4 shows how to validate the SDE mixed model using posterior predictive distributions. In Section 5, the particular case of the Gompertz function is applied on chicken growth data.

2. Models and notations

2.1 *Nonlinear mixed models*

Let $\boldsymbol{y} = (y_i)_{1\leqslant i\leqslant n} = (y_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant n_i}$ denote the data, where y_{ij} is the noisy measurement of the observed biological process for individual i at time t_{ij}, for $i = 1, \ldots, n$, $j = 0, \ldots, n_i$.

In classical mixed models, the process is modeled by a deterministic function, depending on individual random parameters. Formally, the classical nonlinear mixed model is defined as:

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

with f being a parametric deterministic function and $\boldsymbol{\phi} = (\phi_i)_{1 \leqslant i \leqslant n}$ the p-vectors of individual parameter vectors. The ϕ_i are assumed to be independently and identically normally distributed with expectation μ and variance Ω. The ε_{ij} are the residual errors, assumed to be independently and identically normally distributed with null mean and variance σ^2.

For growth curve data, f is classically one of the four most famous parametric functions modeling growth curves, namely the logistic, the Gompertz, the Richards and the Weibull functions. Each of them can be written as the solution of an ordinary differential equation (ODE) describing the evolution of growth rate, which are respectively:

$$f'(t) = Cf(t)\left[1 - \frac{1}{A}f(t)\right], \quad f(0) = \tfrac{A}{1+B} \quad \text{(Logistic)} \tag{2}$$

$$f'(t) = BCe^{-Ct}f(t), \quad f(0) = Ae^{-B} \quad \text{(Gompertz)} \tag{3}$$

$$f'(t) = \frac{BCDe^{-Ct}}{1 + Be^{-Ct}}f(t), \quad f(0) = \tfrac{A}{(1+B)^D} \quad \text{(Richards)} \tag{4}$$

$$f'(t) = DCt^{D-1}(A - f(t)), \quad f(0) = A - B \quad \text{(Weibull)} \tag{5}$$

where A, B, C, D are non-negative parameters. A is the upper asymptote, C and D are growth rate parameters. All four models prescribe monotone increasing curves. More generally, if ϕ denotes the parameter vector (either (A, B, C), (A, B, C, D) or a well-chosen parametrization), f is the solution of the following ODE:

$$\frac{\partial f(\phi, t)}{\partial t} = F(f, t, \phi), \quad f(\phi, 0) = f_0(\phi) \tag{6}$$

2.2 *Nonlinear mixed models defined by stochastic differential equations*

The classical nonlinear mixed model is extended by replacing the regression function by a stochastic process. We propose to introduce a stochastic term in the ODE (6) to take into

account individuals whose growth curve suffers from an unexpected growth rate change. Growth curve is thus described by a random process, denoted (Z_t), solution of the SDE:

$$dZ_t = F(Z_t, t, \phi)dt + \Gamma(Z_t, \phi, \gamma^2)dW_t, \quad Z(t=0) = Z_0(\phi) \tag{7}$$

where W_t is a Brownian motion. $\Gamma(Z_t, \phi, \gamma^2)$ is the volatility function depending on the unknown parameter γ^2. The nonlinear mixed model defined by an SDE is thus:

$$\begin{aligned} y_{ij} &= Z_{t_{ij}}(\phi_i) + \varepsilon_{ij},\ \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2) \\ dZ_t(\phi_i) &= F(Z_t, t, \phi_i)dt + \Gamma(Z_t, \phi, \gamma^2)dW_t \\ \phi_i &\sim \mathcal{N}(\mu, \Omega) \end{aligned} \tag{8}$$

In model (8), three fundamentally different noises are distinguished: the inter-subject variability Ω, the dynamic noise γ^2, reflecting the random fluctuations around the corresponding theoretical dynamic model, and the measurement noise σ^2 representing the uncorrelated part of the residual variability associated with assay or sampling errors.

Many types of volatility functions can be proposed to extend an ODE into an SDE (*e.g.* constant, square root or polynomial volatility). This choice depends on several considerations. If the observed biological process is non-negative, a volatility function ensuring the positivity of (Z_t) will be chosen. If biological reasons imply that a model parameter fluctuates along the experiment record, the volatility can be derived by adding a random perturbation to this parameter. If heteroscedastic variances have been used in an ODE modeling approach, a polynomial volatility can be chosen. Finally, algorithmic and computational constraints have to be considered: an SDE with explicit solution implies a simpler estimation scheme leading to good estimation properties (convergence of the algorithm to the true posterior distribution) whereas an SDE without explicit solution implies additional computational difficulties (use of an approximation scheme). As an example, we propose to use an affine

volatility function $\Gamma(Z_t, \phi, \gamma^2) = \gamma Z_t$, for the logistic (2), Gompertz (3) and Richards (4) models: the process (Z_t) is then a log-Gaussian process (see Section 5.2 for more details).

3. Bayesian estimation

3.1 *Prior specification*

The Bayesian approach consists in the evaluation of the posterior distribution of the population parameters (μ, Ω), σ^2 and the volatility γ^2 for the SDE model. The first step is thus the choice of the prior distributions. Usual diffuse prior distributions can be chosen but the resulting posterior distributions may not be proper. Therefore, we suggest to use standard prior distributions proposed, among others, by De la Cruz-Mesia and Marshall (2006) for expectation or variance parameters in hierarchical models:

$$\begin{array}{rcll} \mu_k & \sim & \mathcal{N}(m_k^{prior}, v_k^{prior}), & k = 1, \ldots, p \\ \Omega^{-1} & \sim & W(R, p+1), & 1/\sigma^2 \sim \Gamma(\alpha_\sigma^{prior}, \beta_\sigma^{prior}) \end{array} \tag{9}$$

where W and Γ are respectively the Wishart and Gamma disributions. The γ^2 parameter controls the variance of the random perturbation. Many prior distributions can be used such as uniform, inverse-Gamma or Jeffreys. A sensitivity analysis is performed on the real data set (Section 5.5). In pratice the specification of hyperparameters $m_k^{prior}, v_k^{prior}, R, \alpha_\sigma^{prior}, \beta_\sigma^{prior}$ may be difficult. We choose the values of hyperparameters to obtain non-informative priors.

3.2 *Posterior computation*

Since models (1) and (8) are non-linear, posterior distributions are not explicit and iterative estimation procedures have to be used. For the ODE model (1), Gibbs sampling algorithms including the sampling of the auxiliary random variables ϕ_i under their conditional distributions have been proposed in the literature (Carlin and Louis, 2000, *e.g.)*. These algorithms do not present any particular difficulties and are not detailed here. For the SDE model (8), we propose to use a Gibbs algorithm, including the sampling of the auxiliary random variables

ϕ_i and the vectors Z_i of realizations of process (Z_t) for each individual at each observation time. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n) \in \boldsymbol{R}^{(n_1+1)+\ldots(n_n+1)}$ denote the vector of the n realizations. Hence the Gibbs sampling algorithm for the SDE model is outlined as follows:

- STEP 1: initialize the iteration counter of the chain $k = 1$ and start with initial values $\sigma^{-2(0)}, \gamma^{2(0)}, \mu^{(0)}, \boldsymbol{\phi}^{(0)}, \boldsymbol{Z}^{(0)}$.
- STEP 2: obtain $\sigma^{-2(k)}, \gamma^{2(k)}, \mu^{(k)}, \boldsymbol{\phi}^{(k)}, \boldsymbol{Z}^{(k)}$ from $\sigma^{-2(k-1)}$, $\gamma^{2(k-1)}$, $\mu^{(k-1)}$, $\boldsymbol{\phi}^{(k-1)}$, $\boldsymbol{Z}^{(k-1)}$ through successive generations of

(1) $\boldsymbol{Z}^{(k)} \sim p(\boldsymbol{Z}|\boldsymbol{\phi}^{(k-1)}, \gamma^{-2(k-1)}, \sigma^{-2(k-1)}, \boldsymbol{y})$

(2) $\boldsymbol{\phi}^{(k)} \sim p(\boldsymbol{\phi}|\sigma^{-2(k-1)}, \gamma^{-2(k-1)}, \mu^{(k-1)}, \Omega^{(k-1)}, \boldsymbol{Z}^{(k)}, \boldsymbol{y}_0)$ where $\boldsymbol{y}_0 = (y_{i0})_{i=1\ldots n}$

(3) $\mu^{(k)} \sim p(\mu|\boldsymbol{\phi}^{(k)})$ and $\Omega^{(k)} \sim p(\Omega|\boldsymbol{\phi}^{(k)})$

(4) $\sigma^{-2(k)} \sim p(\sigma^{-2}|\boldsymbol{Z}^{(k)}, \boldsymbol{\phi}^{(k)}, \boldsymbol{y})$ and $\gamma^{-2(k)} \sim p(\gamma^{-2}|\boldsymbol{Z}^{(k)}, \boldsymbol{\phi}^{(k)})$

- STEP 3: change k to $k+1$ and return to STEP 2 until convergence is reached.

Some conditional distributions are explicit. A Gamma prior distribution on σ^{-2} implies that $p(\sigma^{-2}|\boldsymbol{Z}^{(k)}, \boldsymbol{\phi}^{(k)}, \boldsymbol{y})$ is a Gamma density. The prior distribution of $p(\boldsymbol{\phi}|\mu, \Omega)$ being Gaussian, the conditional distribution of μ is Gaussian and the conditional distribution of Ω is inverse Wishart. The conditional distributions on $\boldsymbol{\phi}$, $\boldsymbol{Z}$ and γ^2 depend on the specific form of the SDE and are detailed in the particular example of the Gompertz model in Section 5. Depending on the model complexity, we may have to resort to Metropolis-Hastings algorithms. Moreover, for SDEs without explicit solution, the conditional distribution on $\boldsymbol{Z}$ has generally no closed form. In this case, we suggest to approximate the diffusion by the Euler-Maruyama scheme, which leads to Gaussian approximations of the transition densities. An approximate statistical model is introduced on which the posterior distributions are computed.

3.3 *Posterior distribution using Euler-Maruyama approximation*

The Euler-Maruyama scheme is presented for subject i. If the time intervals between the observation instants are too great to obtain a good approximation of the transition density, a natural approach is to introduce a set of auxiliary latent data points between every pair of observations, as first proposed by Pedersen (1995). Let $t_{i0} = \tau_0 < \ldots < \tau_m < \ldots < \tau_{M_i} = t_{i,n_i}$ denote the equally spaced discretization of the time interval $[t_{i0}, t_{i,n_i}]$ and h be the step size of the discretization. We assume that, for all $j = 0 \ldots n_i$, there exists an integer m_j verifying $t_{ij} = \tau_{m_j}$ ($m_0 = 0$ by definition). Using the approximated diffusion process from the Euler-Maruyama scheme of step size h, an approximate statistical model is defined as:

$$\begin{aligned} y_{ij} &= \tilde{Z}^h_{m_j}(\phi_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2) \\ \tilde{Z}^h_m(\phi_i) &= \tilde{Z}^h_{m-1}(\phi_i) + h\, F(\tilde{Z}^h_{m-1}, \tau_{m-1}, \phi_i) + \Gamma(\tilde{Z}^h_{m-1}, \phi_i, \gamma^2)\sqrt{h}\, \xi_m \,, 1 \leqslant m \leqslant M_i, \\ \xi_m &\sim_{i.i.d} \mathcal{N}(0,1), \quad \phi_i \sim_{i.i.d} \mathcal{N}(\mu, \Omega) \end{aligned} \tag{10}$$

For model (10), the conditional distribution of the approximate diffusion $\tilde{Z}^h$ is Gaussian, allowing to implement the previously presented Gibbs algorithm. The convergence of this Gibbs algorithm is ensured by classical results (Carlin and Louis, 2000). However, this Gibbs algorithm is performed on the approximate model (10), and computes the posterior distribution $p^h(\theta|\boldsymbol{y})$ of model (10), with $\theta = (\mu, \Omega, \sigma^2, \gamma^2)$, instead of the original posterior distribution $p(\theta|\boldsymbol{y})$. But, the error induced by the Euler scheme on the posterior distributions can be controlled, as shown in the Supplementary materials.

4. Model validation

The goal in model checking is to monitor the quality of the proposed model, i.e. to determine whether the observed data are representative of the type of data we might expect under this model. Posterior predictive checks set this up by generating replicated data sets from the estimated posterior predictive distribution. These replicated data sets are then compared

with the observed data. The function used to compare observed and replicated datasets is the discrepancy function; it depends on data and model parameters and is denoted $T(\boldsymbol{y}, \eta)$, η being used as generic notation for a function of the model parameters. It quantifies incompatibility of the model with the observed data. In our case, we consider for T the χ^2 discrepancy function, computed for the whole population at each age as:

$$T_j(\boldsymbol{y}, \eta) = \sum_i \frac{(y_{ij} - \eta_{ij})^2}{Var(y_{ij} - \eta_{ij})}$$

For the observation at time t_{ij}, we choose $\eta_{ij} = f(\phi_i, t_{ij})$ for the ODE model and $\eta_{ij} = Z_{ij}(\phi_i)$ for the SDE model. Consequently, for both models, $Var(y_{ij} - \eta_{ij}) = \sigma^2$.

We aim at comparing the posterior distribution $p(T_j(\boldsymbol{y}, \eta)|\boldsymbol{y})$ of the observed data $\boldsymbol{y}$ with the posterior distribution $p(T_j(\boldsymbol{y}_{rep}, \eta)|\boldsymbol{y})$ where $\boldsymbol{y}_{rep}$ denotes the replicated data drawn from the posterior predictive distribution $p(\boldsymbol{y}_{rep}|\boldsymbol{y})$. A short version of that posterior predictive distribution is the posterior predictive p-value:

$$p_{pp,j} = P\left[T_j(\boldsymbol{y}_{rep}, \eta) \geqslant T_j(\boldsymbol{y}, \eta)|\boldsymbol{y}\right] = \int P\left[T_j(\boldsymbol{y}_{rep}, \eta) > T_j(\boldsymbol{y}, \eta)|\boldsymbol{y}, \eta\right] p(\eta|\boldsymbol{y}) d\eta \qquad (11)$$

Since this quantity has no closed form, the idea is to approximate it by the Monte Carlo method. For each estimated model (ODE and SDE), the Gibbs algorithm used to estimate the posterior distribution provides a set of vectors η^l $(l = 1 \dots L)$ drawn from the posterior distribution $p(\eta|\boldsymbol{y})$. For each of this draw, a replicated data set y^l_{rep} is simulated from the posterior predictive distribution of the data $p(\boldsymbol{y}_{rep}, \eta^l)$. Finally, the posterior predictive p-value (11) is estimated by the Monte Carlo method as $\frac{1}{L}\sum_{l=1}^{L} 1_{T_j(\boldsymbol{y}^l_{rep}, \eta^l) > T_j(\boldsymbol{y}, \eta^l)}$.

5. An example: chicken growth modeling with the Gompertz function

We focus on the modeling of chicken growth. Data $\boldsymbol{y}$ are noisy weight measurements of $n = 50$ chickens at weeks t =0, 4, 6, 8, 12, 16, 20, 24, 28, 32, 36, 40 after birth: see the corresponding curves on Figure 1. Such a data set has been previously analyzed by Mignon-Grasteau et al. (1999), Jaffrézic et al. (2006) and Meza et al. (2007), who conluded that,

among the standard growth models, the monotonic mixed Gompertz model is the most appropriate one. This model is adapted to the most subjects, however it fails to model the unexpected variations of growth rate for some individuals (see Figure 1).

5.1 *The classical Gompertz nonlinear mixed model*

Jaffrézic and Foulley (2006) underline that a heteroscedastic error model is required to obtain satisfactory results. For simplicity's sake, we consider modeling the logarithm of the data $\boldsymbol{y}$ by adding an additive measurement error with a constant variance:

$$\left\{\begin{array}{rcl} \log y_{ij} & = & \log A_i - B_i e^{-C_i t_{ij}} + \varepsilon_{ij},\ \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0,\sigma^2),\ \forall i = 1,\ldots,n, j = 0,\ldots,n_i \\ \phi_i & = & (\log A_i, B_i, \log C_i) \sim_{i.i.d.} \mathcal{N}(\mu,\Omega),\ \forall i = 1,\ldots,n \end{array}\right. \tag{12}$$

We use the log-parametrization for parameters A_i and C_i. This parametrization has two advantages: it simplifies the computation of the posterior distributions and it ensures the positivity of the parameters. We set $\mu = (\log(a), b, \log(c))$.

5.2 *Extension to the Gompertz stochastic nonlinear mixed model*

We now deduce the SDE model from the Gompertz equation (3). Given the heteroscedasticity of the process, the volatility function is set to be equal to $\Gamma(Z_t,\phi,\gamma^2) = \gamma Z_t$:

$$dZ_t = BCe^{-Ct}Z_t dt + \gamma Z_t dW_t,\ Z_0 = Ae^{-B} \tag{13}$$

This means that the standard error of the random perturbations of the growth rate is proportional to weight. This choice of volatility has two main advantages. First, SDE (13) has an explicit solution. Indeed, set $X_t = \log(Z_t)$. By the Ito's formula, for $h > 0$, the conditional distribution of X_{t+h} given $(X_s), s \leqslant t$ is:

$$X_{t+h}|(X_s)_{s\leqslant t} \sim \mathcal{N}(X_t - Be^{-Ct}(e^{-Ch}-1) - \frac{1}{2}\gamma^2 h, \gamma^2 h),\ \ X_0 = \log(A) - B$$

Thus, $\forall t > 0$, we have $Z_t = Ae^{-Be^{-Ct}}e^{-\frac{1}{2}\gamma^2 t + \eta_t} = f(t)e^{-\frac{1}{2}\gamma^2 t+\eta_t}$ with $\eta_t \sim \mathcal{N}(0,\gamma^2 t)$ and $Z_0 = Ae^{-B}$. As a consequence, Z_t is a multiplicative random perturbation of the solution of

the Gompertz model. Second, due to the assumption of the non-negativity of A, Z_t is almost surely non-negative, which is a natural constraint to model weight records.

We then discretize the SDE. The discrete realization $(X_{t_{ij}})$ of the SDE is Markovian:

$$X_{i,t_{ij}}|X_{i,t_{ij-1}} \sim \mathcal{N}\left(X_{i,t_{ij-1}} - B_i e^{-C_i t_{ij-1}}(e^{-C_i(t_{ij}-t_{ij-1})} - 1) - \frac{1}{2}\gamma^2(t_{ij} - t_{ij-1}), \gamma^2(t_{ij} - t_{ij-1})\right)$$

with $X_{i,0} = \log(A_i) - B_i$. The SDE model (8) on the logarithm of data is thus defined as:

$$\begin{cases} (\log y_{i0}, \log y_{i1}, \ldots, \log y_{in_i})' = \left(\log(A_i) - B_i, X_{t_{i1}}, \ldots, X_{t_{in_i}}\right)' + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}_{n_i+1}\right) \\ \left(X_{t_{i1}}, \ldots, X_{t_{in_i}}\right)' = \left(\log(A_i) - B_i e^{-C_i t_{i1}}, \ldots, \log(A_i) - B_i e^{-C_i t_{in_i}}\right)' - \gamma^2 \left(t_{i1}, \ldots, t_{in_i}\right)' + \eta_i \\ \eta_i \sim_{i.i.d} \mathcal{N}\left(0_J, \gamma^2 T_i\right), \; T_i = (\min(t_{ij}, t_{ij'}))_{1 \leqslant j,j' \leqslant n_i} \\ (\log A_i, B_i, \log C_i) \sim_{i.i.d.} \mathcal{N}(\mu, \Omega) \end{cases} \tag{14}$$

5.3 *Posterior computation and inference in the Gompertz model*

Conditional distribution computation for the ODE mixed model is standard. We detail the computation under the SDE mixed model. Let $m_a^{prior}, m_b^{prior}, m_c^{prior}, v_a^{prior}, v_b^{prior} v_c^{prior}$ denote the prior parameters of the 3 components of μ. The conditional distribution of $\boldsymbol{X}_i = (X_{ij})_{1 \leqslant j \leqslant n_i}$ given $(\phi_i, \gamma^{-2}, \boldsymbol{y_i}, \sigma^2)$ is Gaussian with mean $m_{\boldsymbol{X}_i}^{post}$ and variance $V_{\boldsymbol{X}_i}^{post}$:

$$V_{\boldsymbol{X}_i}^{post} = (\sigma^{-2} I_{n_i - 1} + \gamma^{-2} T_i^{-1})^{-1}, \;\; m_{\boldsymbol{X}_i}^{post} = V_{\boldsymbol{X}_i}^{post} \left[\sigma^{-2}(\log y_{i1} \ldots \log y_{in_i})' + \gamma^{-2} T_i^{-1} u_{\boldsymbol{X}_i}\right]$$

with $u_{\boldsymbol{X}_i} = \log A_i - B_i \left(e^{-C_i t_{i1}} \ldots e^{-C_i t_{in_i}}\right)' - \frac{1}{2}\gamma^2 \left(t_{i1} \ldots t_{in_i}\right)'$. Let $\omega^2_{\log A}, \omega^2_B, \omega^2_{\log C}$ denote the diagonal elements of Ω. Let $\Omega_{k,(j,j')}$ denote the two-vector composed of the elements on the k-th row and (j, j') columns of Ω and $\Omega_{(j,j'),(j,j')}$ the two-symetric-matrix composed of the elements on the (j, j')-th rows and (j, j')-th columns of Ω. Set the $(n_i + 1) \times (n_i + 1)$-matrix:

$$G_i = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \gamma^2 T_i \end{pmatrix} \tag{15}$$

The conditional distributions for the individual parameters $\log A_i$ and B_i are $\mathcal{N}(m_{A_i}^{post}, V_{A_i}^{post})$ and $\mathcal{N}(m_{B_i}^{post}, V_{B_i}^{post})$, respectively, with:

$$V_{A_i}^{post} = \left[(1 \ldots 1) G_i^{-1} (1 \ldots 1)' + \frac{1}{\omega^2_{\log A \,|B, \log C}}\right]^{-1}$$

$$
\begin{aligned}
m_{A_i}^{post} &= V_{A_i}^{post}\left[(1\ \dots\ 1)G_i^{-1}u_{Ai} + \frac{\mu_{\log A\,|B,\log C}}{\omega^2_{\log A\,|B,\log C}}\right]\\
V_{B_i}^{post} &= \left[(e^{-C_it_{i0}}\ \dots\ e^{-C_it_{in_i}})G_i^{-1}(e^{-C_it_{i0}}\ \dots\ e^{-C_it_{in_i}})' + \frac{1}{\omega^2_{B\,|\log A,\log C}}\right]^{-1}\\
m_{B_i}^{post} &= V_{B_i}^{post}\left[(e^{-C_it_{i0}}\ \dots\ e^{-C_it_{in_i}})G_i^{-1}u_{Bi} + \frac{\mu_{B\,|\log A,\log C}}{\omega^2_{B\,|\log A,\log C}}\right]
\end{aligned}
$$

where

$$
\begin{aligned}
u_{Ai} &= (\log y_{i0}\ X_{i1}\dots X_{in_i})' + B_i\left(e^{-C_it_{i0}}\dots e^{-C_it_{in_i}}\right)' - \frac{1}{2}\gamma^2\,(t_{i0}\dots t_{in_i})'\\
\omega^2_{\log A|B,\log C} &= \omega^2_{\log A} - \Omega_{\log A,(B,\log C)}\Omega^{-1}_{(B,\log C),(B,\log C)}\Omega'_{\log A,(B,\log C)}\\
\mu_{\log A\,|B,\log C} &= \log a + \Omega_{\log A,(B,\log C)}\Omega^{-1}_{(B,\log C),(B,\log C)}\left((B_i,\log C_i)' - (b,\log c)'\right)\\
u_{Bi} &= (\log y_{i0},\ X_{i1}\dots X_{in_i})' + \log A_i - \frac{1}{2}\gamma^2\,(t_{i0}\dots t_{in_i})\\
\omega^2_{B|\log A,\log C} &= \omega^2_B - \Omega_{B,(\log A,\log C)}\Omega^{-1}_{(\log A,\log C),(\log A,\log C)}\Omega'_{B,(\log A,\log C)}\\
\mu_{B\,|\log A,\log C} &= b + \Omega_{B,(\log A,\log C)}\Omega^{-1}_{(\log A,\log C),(\log A,\log C)}\left((\log A_i,\log C_i)' - (\log a,\log c)'\right)
\end{aligned}
$$

The conditional distributions of $\log(a)$ and b are Gaussian with parameters:

$$
V_a^{post} = \left[n\omega^{-2}_{\log A} + (v_a^{prior})^{-1}\right]^{-1} \quad \text{and} \quad m_a^{post} = V_a^{post}\left[\omega^{-2}_{\log A}\frac{1}{n}\sum_{i=1}^{n}\log A_i + \frac{m_a^{prior}}{v_a^{prior}}\right]
$$

$$
V_b^{post} = \left[n\omega_B^{-2} + (v_b^{prior})^{-1}\right]^{-1} \quad \text{and} \quad m_b^{post} = V_b^{post}\left[\omega_B^{-2}\frac{1}{n}\sum_{i=1}^{n}B_i + \frac{m_b^{prior}}{v_b^{prior}}\right]
$$

The conditional distribution of Ω^{-1} is $W(R + (\boldsymbol{\phi}-\mu)(\boldsymbol{\phi}-\mu)', n+p+1)$ where $\boldsymbol{\phi}-\mu = [(\phi_1-\mu)\dots(\phi_n-\mu)] \in \boldsymbol{R}^{3\times n}$. The conditional distribution of σ^2 is $\Gamma(\alpha_\sigma^{post}, \beta_\sigma^{post})$ with:

$$
\alpha_\sigma^{post} = \alpha_\sigma^{prior} + \sum_{i=1} n\frac{n_i+1}{2} \quad \text{and} \quad \beta_\sigma^{post} = \left[\frac{1}{\beta_\sigma^{prior}} + \frac{1}{2}\sum_{i=1,j=0}^{n,n_i}(\log y_{ij} - X_{ij})^2\right]^{-1}
$$

The posterior distributions of $\log C_i$, $\log c$ and γ^2 have no explicit form and we use the Metropolis-Hastings random-walks.

The Metropolis-Hastings and Gibbs algorithm convergences are ensured by the theorems proposed by Carlin and Louis (2000) and Mengersen and Tweedie (1996). The implementations of the ODE and SDE approaches have the same level of complexity. The convergence and the stability of the MCMC algorithms produced by these two algorithms are equivalent.

5.4 *Simulations*

We simulate datasets mimicking chicken growth with $n = 50$ individuals and $n_i = 9$ measurements obtained every 5 weeks after birth. The population parameters are $\log(a) = \log(3000)$, $b = 5$, $\log(c) = \log(14)$, Ω is assumed diagonal with diagonal elements equal to 100 and $\sigma^{-2} = 5$. A 100 datasets are simulated via the mixed model defined by the Gompertz model (12) and a 100 datasets with the mixed model defined by the Gompertz SDE (14), with $\gamma^2 = 1$. We estimate all the parameters under the ODE mixed model (12) and the SDE mixed model (14), successively. The two algorithms take 96s and 206s on a dataset with a Intel Core2 Duo CPU (2.4 GHz), respectively. Estimates are obtained as the expectation of the parameter posterior distribution. Means and standard errors computed for each parameter on the 100 estimation results obtained with both algorithms are presented in Table 1.

[Table 1 about here.]

When data are simulated under the ODE model, estimates obtained with the Bayesian ODE algorithm are very satisfactory. Those obtained by the Bayesian SDE algorithm are also satisfactory although the bias for the variance parameter ω_{lnA}^{-2} is larger. Note that, as expected, the estimation of the volatility parameter γ^2 is rather low (0.19). When data are simulated under the SDE model, estimates obtained with the Bayesian SDE model are very satisfactory, with small bias and standard error. On the contrary, the estimates obtained with the Bayesian ODE algorithm have larger bias, including the parameters of fixed effects. The parameter $\omega_{\log A}^{-2}$ is very badly estimated (8.63 to be compared to the true value 100).

5.5 *Application on chicken growth data*

The proposed models are applied on real data of chicken growth. The ODE and SDE models (12) and (14) are used to model the logarithm of the data. The influence of the ODE model priors has been validated in Jaffrézic et al. (2006). For the SDE model, the influence of the prior of γ^2 is studied using the Deviance Information Criterion (DIC) (Spiegelhalter et al.

(2002)). Only differences in DIC are meaningful. A inverse Gamma prior on γ^2, a Jeffreys prior on γ^{-2}, a log normal prior on $\log(\gamma)$, a uniform prior on γ^2 have been tested leading to DIC variations from the uniform prior choice equal to 24, 1.5, 1.4 and 0 respectively. The influence on the posterior distributions was very light. Results are presented with a uniform prior on γ^2. Posterior expectations of the parameters are presented in Table 2. Diagnostic tools to validate the models are applied to both ODE and SDE models: Figure 2 presents the posterior predictive distributions and p-values of both models computed for each time point. The estimate of γ^2 is strictly positive and its credibility interval puts the parameter of long way from zero (see the posterior distribution of γ^2 in the Supplementary Materials). This means that the dynamical process that most likely represents the growth is a stochastic process with non-negligible noise. The diagnostic tools also show a clear improvement from the ODE model to SDE model for the whole population, both at early and late ages. The reduction in DIC from the ODE to the SDE models is equal to 393, which clearly indicates the better predictive ability of the SDE model. The predictive abilities of models $M_1 = ODE$ and $M_2 = SDE$ can also be compared on the posterior expectation of squared errors using cross-validation techniques i.e. after dropping information at measurement j (the new data set is denoted $\boldsymbol{y}_{-j}$) to make prediction at the corresponding occasion:

$$r_j^k = \sum_{i=1}^{n} \boldsymbol{E}\left[(\log(y_{i,j}^{rep,k}) - \log(y_{i,j}))^2 | \boldsymbol{y}_{-j}\right], \quad k = 1, 2$$

with $y_{i,j}^{rep,k}$ drawn from the predictive distribution $y_{i,j}^{rep,k} \sim p(y_{i,j}^{rep,k}|\boldsymbol{y}_{-j})$. Averaging in r_j^k is with respect to the posterior uncertainty in the parameters of the model. We performed that comparison for the last observation $j = 12$ which is especially critical with respect to the growth pattern studied here. These quantities are $r_{12}^{ode} = 0.56$ and $r_{12}^{sde} = 0.48$ resulting in a reduction of the squared errors of prediction of 14% when using SDE vs ODE.

Figure 3 reports, for four subjects, the observed weights, the ODE prediction, the empirical mean of the last 1000 simulated trajectories of the SDE (14) generated during the Gibbs

algorithm, their empirical 95 % confidence limits (from the 2.5th percentile to the 97.5th percentile) and one simulated trajectory. Subjects 4 and 13 are examples of subjects with no growth slow down. Both ODE and SDE models satisfactorily fit the observations. Subject 14 has a small observed weight decrease. For subject 1, the weight decrease is more important. For both subjects, the ODE model fails to capture this phenomenon while the SDE model does. Furthermore, the SDE model provides different estimates for the individual parameters. For example for subject 1, the individual parameter A_1 (adult weight) is estimated at 3.922 kg and 3.484 kg by the ODE and SDE models, respectively.

[Figure 1 about here.]

[Table 2 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

6. Conclusion and discussion

We propose a Bayesian approach to nonlinear mixed models defined by stochastic differential equations. These models are an alternative to classical nonlinear mixed models whose deterministic regression function is too restrictive to model some unexplained biological processes such as growth rate changes. On the presented data set, the introduction of this SDE model leads to a clear validation of the model (Figure 2) which was not the case in the standard model, justifying the introduction of the new stochastic component. Moreover, the modelling of these fluctuations has a non-neglectible impact on the estimation of the parameters. This might have a great influence on the conclusions about genetic specifications.

We consider SDEs defined with a general volatility function. As discussed in Section 2, the study context can induce a natural choice of this volatility function (non-negativity of the process, fluctuation of a parameter, heteroscedastic variances,...). This volatility function

choice has obviously some consequences on the complexity of the Bayesian posterior computation. As detailed in Section 3, a volatility function which induces an explicit solution of the SDE will imply a very easy implementation of the algorithm. The example detailed in this work belongs to this case. On the contrary, when the SDE has no explicit distribution, we propose to use the Euler-Maruyama scheme to approximate the diffusion: the conditional distribution is then Gaussian, implying an easy Bayesian implementation. We control the error induced by this Euler approximate scheme on the posterior distribution. In this context, auxiliary latent points are introduced to obtain a better approximation of the diffusion. The choice of the discrete grids $(\tau_0, \ldots, \tau_{M_i})$ is complex and has been evoked by Pedersen (1995) and Donnet and Samson (2008). In conclusion, generally speaking, there is no limitations to the choice of the volatility function of the SDE approach.

Our model differs from mixed models with continuous time autoregressive measurement errors, as proposed by De la Cruz-Mesia and Marshall (2006) or others. These authors assume that measurement errors have an auto-regressive structure ($Y_{t_{ij}} = f(t_{ij}, \phi_i) + \varepsilon_{t_{ij}}$ and $d\varepsilon_t = -a\varepsilon_t dt + \sigma dW_t$). We assume that the auto-regressive structure observed in residuals of classical nonlinear mixed models comes from a model failure: the regression function is too restrictive and rigid to model random variations of the biological process. Therefore, in our model, it is the regression process that has an auto-regressive structure, while the observation measurements are assumed to be independant. These two models are different. De la Cruz-Mesia and Marshall (2006) consider a stationnary CAR process, implying that the process of the observations has a homoscedastic variance. On the contrary, we do not assume any stationnarity for the process $(Z(t))$. Moreover, heteroscedastic error model can be considered with our approach, depending on the choice of the volatility function.

The proposed model should prove to be useful for other applications in which deterministic models are too restrictive to take into account different sources of variation that exist in real

life. For example, Picchini et al. (2006) propose a stochastic differential equation to model glucose/insulin dynamics, where sources of variability are various (anxiety, rest, etc). The extension of this work to mixed models using our approach should be of great interest.

An interesting area for future research is the development of model selection tools in this context. Indeed, the analysis of covariate effects and the comparison between the ODE and the SDE models require specific selection tools. The method of pseudo-priors proposed by Carlin and Chib (1995) and developed by others, which is very sensitive to the choice of priors and pseudo-priors, would be difficult to use in practice. Bayes factors are complex to compute in these models but could be an interesting alternative. Finally, the extension of this work to multidimensional SDEs would be of great interest in several biological applications.

7. Supplementary Materials

Web Appendix referenced in Sections 3 and 5.5 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgements

The authors are greatful to Dr F. Jaffrézic for helpful discussions on mixed model and to Mrs Brand Williams and Vincent-Arnaud for their English revision. This paper is dedicated to the memory of Prof Guy Lefort (1921-1979) for the 30th anniversary of his death.

References

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

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **57,** 473–484.

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.

Davidian, M. and Giltinan, D. (2003). Nonlinear models for repeated measurements: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics* **8,** 387–419.

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.

Ditlevsen, S. and De Gaetano, A. (2005). Stochastic vs. deterministic uptake of dodecanedioic acid by isolated rat livers. *Bulletin of Mathematical Biology,* **67,** 547–561.

Donnet, S. and Samson, A. (2008). Parametric inference for mixed models defined by stochastic differential equations. *ESAIM Probability & Statistics* **12,** 196–218.

Hou, W., Garvan, C., Zhao, W., Behnke, M., Eyler, F., and Wu, R. (2005). A general model for detecting genetic determinants underlying longitudinal traits with unequally spaced measurements and nonstationary covariance structure. *Biostatistics* **6,** 420–433.

Huggins, R. and Loesch, D. (1998). On the analysis of mixed longitudinal growth data. *Biometrics* **54,** 583–595.

Jaffrézic, F. and Foulley, J. (2006). Modelling variances with random effects in non linear mixed models with an example in growth curve analysis. *XXIII International Biometric Conference, Montreal, 16-21 juillet* .

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

Meng, X. (1994). Posterior predictive p-values. *Annals of Statistics* **22,** 1142–1160.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics* **24,** 101–121.

Meza, C., Jaffrézic, F., and Foulley, J.-L. (2007). REML estimation of variance parameters in nonlinear mixed effects models using the SAEM algorithm. *Biom. J.* **49,** 876–888.

Mignon-Grasteau, S., Beaumont, C., LeBihan, E., Poivey, J., de Rochambeau, H., and Ricard, F. (1999). Genetic parameters of growth curve parameters in male and female chickens. *British Poultry Science* **40,** 44–51.

Overgaard, R., Jonsson, N., Tornoe, C., and Madsen, H. (2005). Non-linear mixed-effects models with stochastic differential equations: Implementation of an estimation algorithm. *Journal of Pharmacokinetics and Pharmacodynamics* **32,** 85–107.

Pedersen, A. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics* **22,** 55–71.

Picchini, U., Ditlevsen, S., and De Gaetano, A. (2006). Modeling the euglycemic hyperinsulinemic clamp by stochastic differential equations. *Journal of Mathematical Biology* **53,** 771–796.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B Statistical Methodology* **64,** 583–639.

Spyrides, M., Struchiner, C., Barbosa, M., and Kac, G. (2008). Effect of predominant breastfeeding duration on infant growth: a prospective study using nonlinear mixed effect models. *Journal of Pediatrics* **84,** 237–243.

Tornoe, C., Overgaard, R., Agerso, H., Nielson, H., and Madsen, H. Jonsson, E. (2005). Stochastic differential equation in NONMEM: Implementation, Application and Comparison with Ordinary Differential Equations. *Pharmaceutical Research* **22,** 1247–1258.

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

Received. Revised.

Accepted.

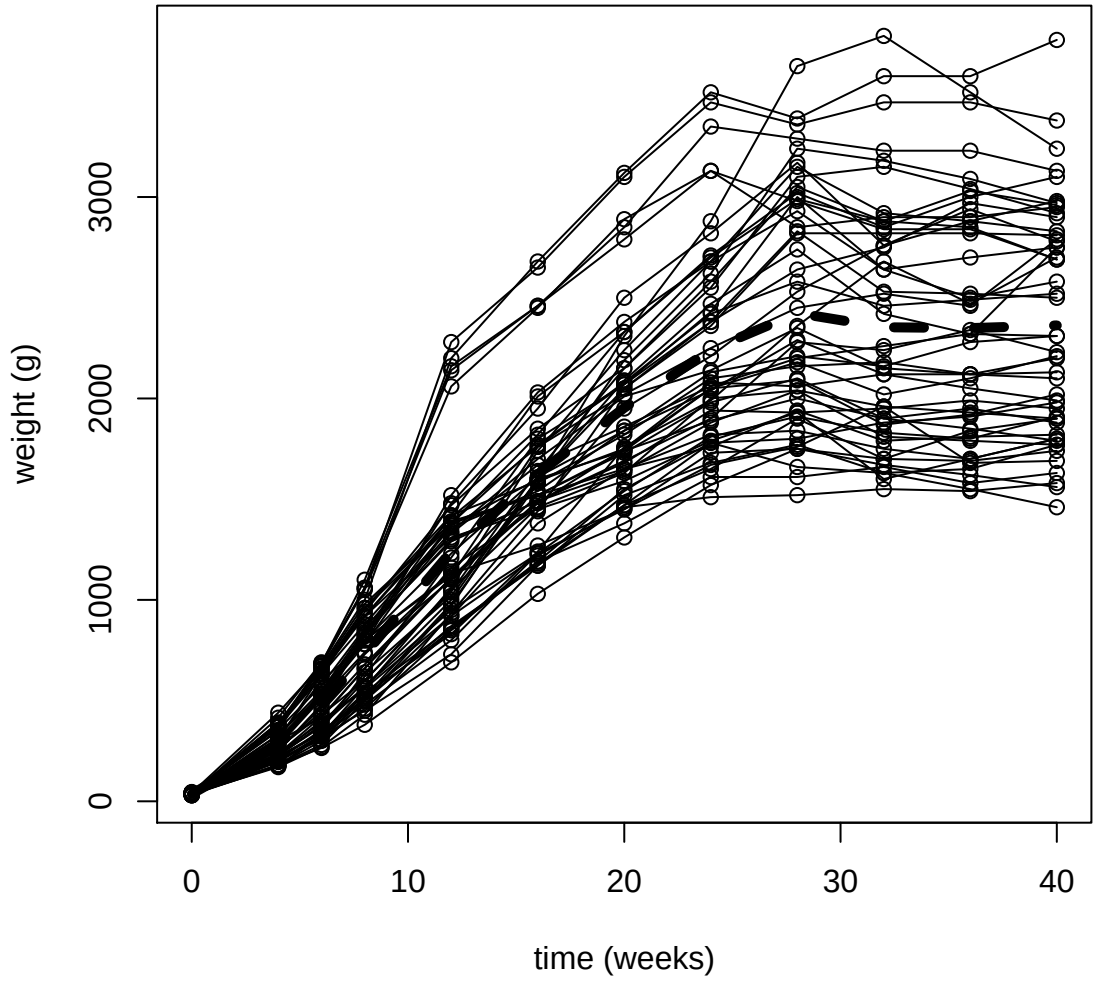

Figure 1. Growth curves of the 50 chickens and mean growth curve in dashed bold line.

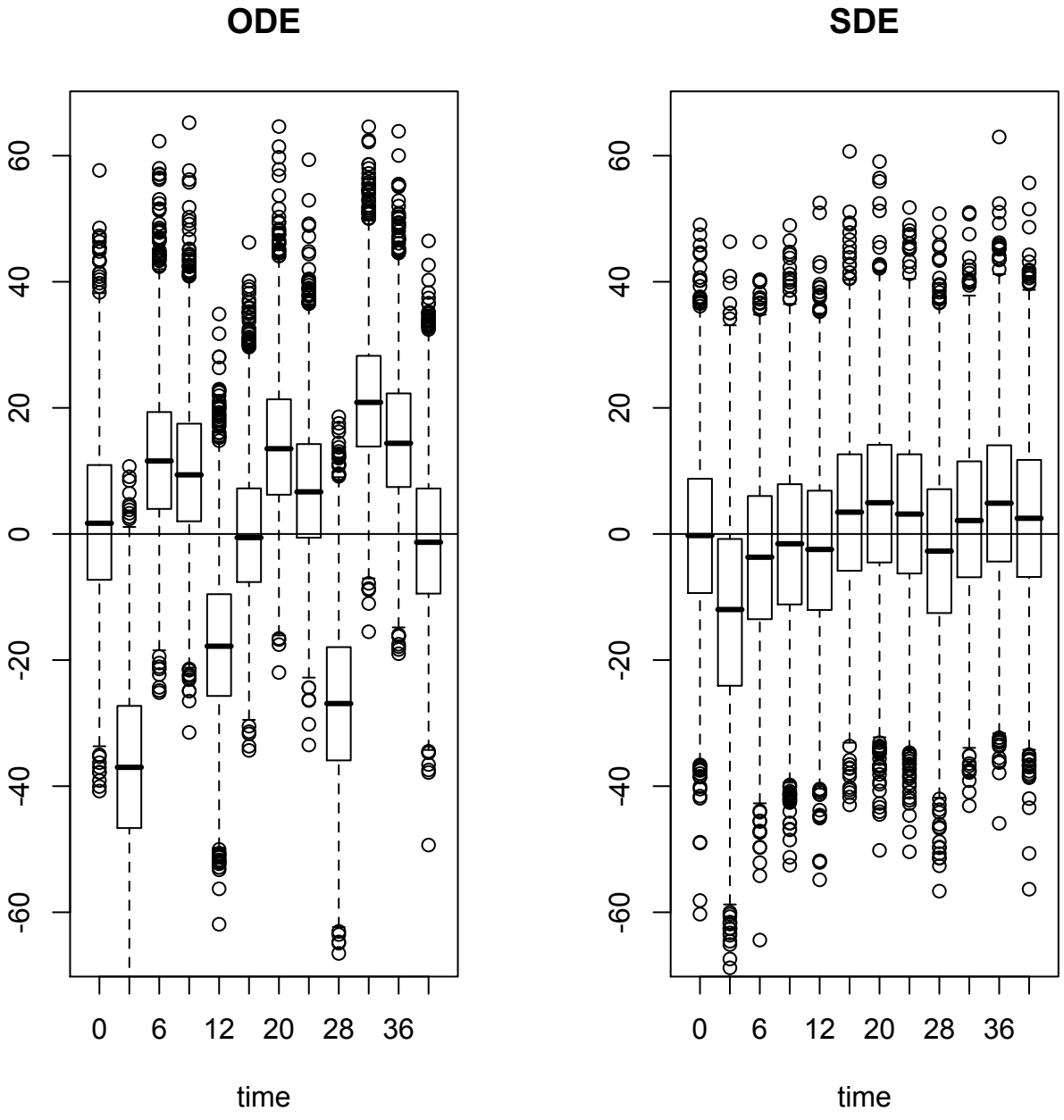

Figure 2. Posterior predictive distributions for the ODE and SDE models on chicken growth data. Posterior predictive p-values at times $(0, 4, 6, 8, 12, 16, 20, 24, 28, 32, 36, 40)$ for the ODE (resp. SDE) mixed models are equal to $(0.55, 0.00, 0.86, 0.80, 0.08, 0.48, 0.90, 0.73, 0.02, 0.99, 0.91, 0.46)$ (resp $(0.49, 0.23, 0.40, 0.45, 0.43, 0.60, 0.64, 0.60, 0.43, 0.57, 0.64, 0.56)$).

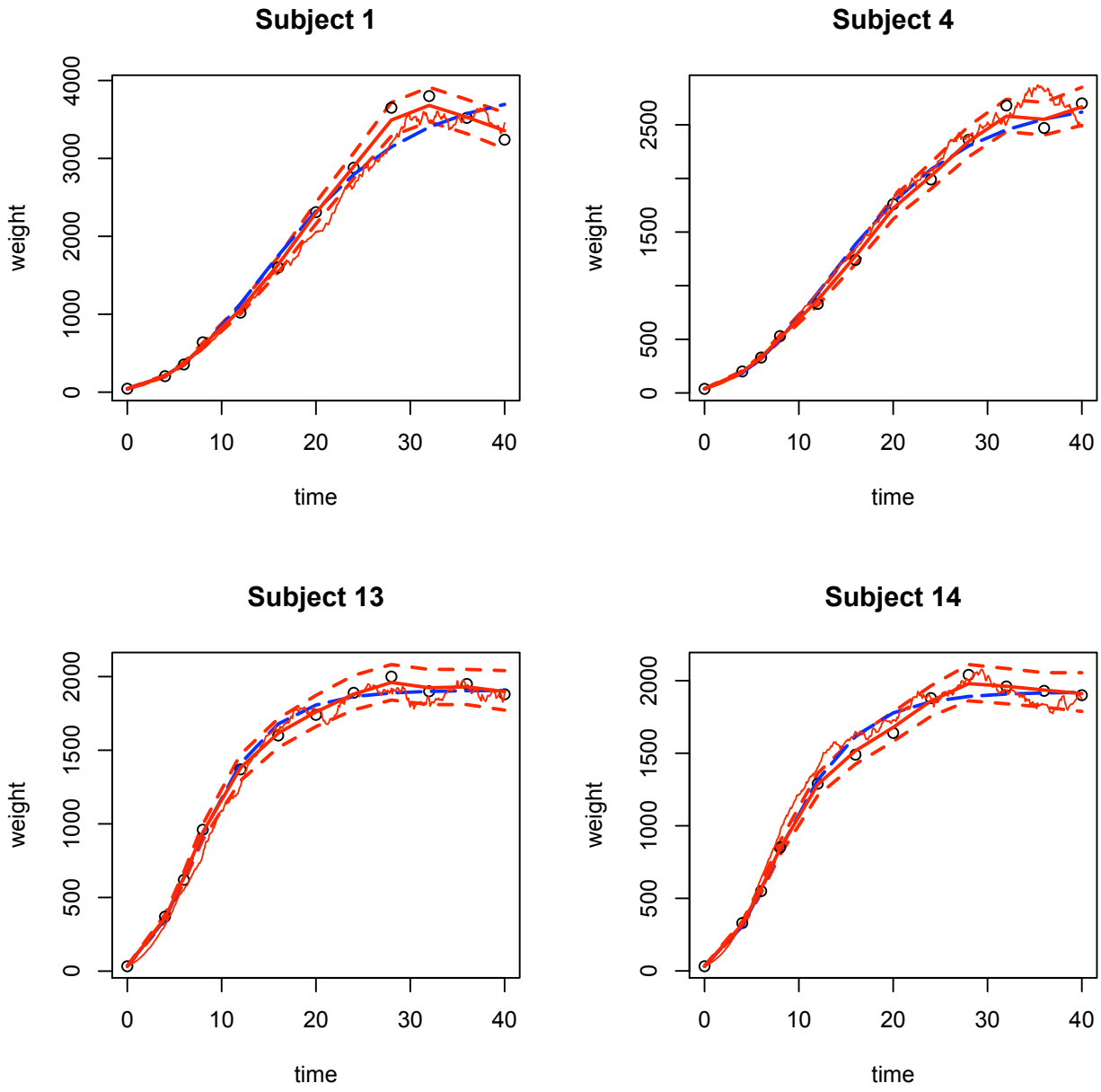

Figure 3. Observations (circles), predictions obtained with the ODE mixed model (long dashed line), mean SDE prediction (smooth solid line), 95% credibility interval obtained with the SDE mixed model (dotted line) and one SDE realization (solid line), for subjects 1, 4 13 and 14.

Table 1
Mean estimates and standard errors in brackets obtained from the ODE and the SDE mixed models on 100 datasets simulated with the ODE or the SDE mixed model.

Simulation model	true	ODE ($\gamma^2 = 0$)		SDE ($\gamma^2 = 1$)	
Estimation model	value	ODE	SDE	ODE	SDE
μ_{lnA}	8.01	8.00 (0.04)	8.03 (0.06)	7.84 (0.07)	8.02 (0.09)
μ_B	5.00	4.99 (0.08)	5.02 (0.08)	4.83 (0.09)	5.01 (0.11)
μ_{lnC}	2.64	2.64 (0.04)	2.63 (0.04)	2.69 (0.05)	2.63 (0.05)
ω_{lnA}^{-2}	100.00	122.24 (39.95)	160.84 (27.84)	8.63 (3.08)	113.58 (29.25)
ω_B^{-2}	100.00	106.70 (22.17)	103.16 (23.74)	87.38 (34.72)	103.50 (24.02)
ω_{lnC}^{-2}	100.00	126.27 (45.69)	131.03 (55.02)	125.31 (47.20)	114.53 (47.69)
γ^2		- (-)	0.19 (0.02)	- (-)	0.96 (0.25)
σ^{-2}	5.00	5.05 (0.39)	5.35 (0.43)	3.67 (0.26)	5.12 (0.40)

Table 2
Posterior distributions for the ODE and SDE models on chicken growth data: mean estimated parameters and their 95% credibility intervals (95% CI).

	ODE		SDE	
	mean	95% CI	mean	95% CI
$\log a$	7.77	[7.70; 7.84]	7.75	[7.67; 7.83]
b	4.17	[4.11; 4.23]	4.15	[4.08; 4.22]
$\log c$	2.75	[2.70; 2.81]	2.78	[2.71; 2.84]
$\Omega^{-1}_{\log a,\log a}$	117.30	[66.53; 190.90]	93.89	[59.02; 139.10]
$\Omega^{-1}_{\log a,b}$	-128.50	[-217.90; -68.32]	-88.46	[-143.10; -45.65]
$\Omega^{-1}_{\log a,\log c}$	-4.57	[-29.53; 16.81]	4.40	[-14.41; 25.06]
$\Omega^{-1}_{b;b}$	172.10	[94.21; 287.90]	146.10	[85.54; 231.9]
$\Omega^{-1}_{b,\log c}$	22.64	[-4.73; 57.69]	23.86	[-1.73; 59.07]
$\Omega^{-1}_{\log c,\log c}$	36.68	[23.04; 54.99]	38.04	[22.52; 61.61]
σ^{-2}	225.5	[197.40; 255.50]	630.22	[463.78; 797.90]
γ^2			0.09	[0.07; 0.12]