Genet. Sel. Evol. 38 (2006) 583–600

DOI: 10.1051/gse:2006023

Original article

Genetic analysis of growth curves using the SAEM algorithm

Florence JAFFRÉZIC[a*], Cristian MEZA[b], Marc LAVIELLE[b], Jean-Louis FOULLEY[a]

[a] Quantitative and Applied Genetics, INRA 78352 Jouy-en-Josas Cedex, France
[b] Laboratoire de Mathématiques, Université Paris Sud, 91400 Orsay, France

(Received 2 February 2006; accepted 10 August 2006)

Abstract – The analysis of nonlinear function-valued characters is very important in genetic studies, especially for growth traits of agricultural and laboratory species. Inference in nonlinear mixed effects models is, however, quite complex and is usually based on likelihood approximations or Bayesian methods. The aim of this paper was to present an efficient stochastic EM procedure, namely the SAEM algorithm, which is much faster to converge than the classical Monte Carlo EM algorithm and Bayesian estimation procedures, does not require specification of prior distributions and is quite robust to the choice of starting values. The key idea is to recycle the simulated values from one iteration to the next in the EM algorithm, which considerably accelerates the convergence. A simulation study is presented which confirms the advantages of this estimation procedure in the case of a genetic analysis. The SAEM algorithm was applied to real data sets on growth measurements in beef cattle and in chickens. The proposed estimation procedure, as the classical Monte Carlo EM algorithm, provides significance tests on the parameters and likelihood based model comparison criteria to compare the nonlinear models with other longitudinal methods.

genetic analysis / growth curves / longitudinal data / stochastic approximation EM algorithm

1. INTRODUCTION

Many traits of interest in genetic studies are function-valued characters, *i.e.* they change in a continuous manner over time or some other independent continuous variable. Focus will be in this study on nonlinear functions applied to growth traits. They are of interest for many agricultural and laboratory species such as rabbits [2], chickens [24], pigs [11], cattle [13], mice [1] and trees [20].

Various methodologies have been proposed to analyze such longitudinal data, including random coefficient models [7], which model individual

* Corresponding author: florence.Jaffrezic@jouy.inra.fr

deviations with polynomial functions of time, and structured antedependence models [12, 25], which consider that the observation at time t is a function of previous observations. These models are in the linear mixed model framework and can be implemented in traditional mixed model softwares.

A different approach for function-valued characters, especially growth traits, is to use a parametric nonlinear function of time, with a few interpretable parameters, that are decomposed into a genetic and an environmental component. For instance, the Gompertz curve has proven suitable for modeling growth curves in rabbits [2] and chickens [24]. It has three parameters that have an interesting biological interpretation in terms of adult body weight and maturation rate. This modeling is similar in spirit to the random regression approach, but it overcomes the drawbacks encountered with the use of polynomimal functions. This nonlinear modeling of growth curves has also been used in QTL detection by Ma *et al.* [20].

Estimation procedures for these nonlinear mixed effects models are, however, much more complex, and require the use of stochastic estimation procedures. Some authors have used the Gibbs sampling for Bayesian estimations [2]. These Bayesian methods do, however, have a few drawbacks such as the choice of prior distributions, the computing time, the check of convergence and inference on the estimated parameters (significance tests, etc.).

McCulloch [21], however, has proposed using a hybrid algorithm combining a Markov chain Monte Carlo EM algorithm – MCEM [28] and a Markov chain Monte Carlo (MCMC) integration and maximization of the likelihood – MCMLE [9]. Indeed, the MCEM algorithm converges quickly to the neighborhood of the parameter estimates, but shows a great deal of variability within this neighborhood. In addition, it requires a considerable increase in the number of MCMC draws and the number of EM iterations to make the procedure accurate [3]. On the contrary, the MCMLE algorithm provides accurate estimates as well as all the elements required for parameter testing and model comparisons. It is, however, very computationally expensive and requires a reference point in the parameter space close to the actual MLE [26].

The aim of this paper was to present an extension of the stochastic approximation EM algorithm (SAEM) proposed in the statistical literature [15] and to apply it to the genetic analysis of growth curves. This methodology combines the strength of the two aforementioned algorithms. As with the MCEM algorithm, it is quite robust to starting values, but has much faster convergence to the maximum likelihood estimates, thanks to a smoothing parameter. It also provides the likelihood value and confidence intervals for all the estimated

parameters, and therefore permits the use of classical significance tests and likelihood based model comparison criteria.

A simulation study will be presented to check the properties of this algorithm in genetic studies, and an application to growth data analysis in beef cattle and in chickens will be presented.

2. MATERIALS AND METHODS

2.1. Presentation of the nonlinear genetic model

The general form of the model can be written as:

$$y_{ij} = f(t_{ij}, \boldsymbol{\phi_i}) + g(t_{ij}, \boldsymbol{\phi_i})\epsilon_{ij} \tag{1}$$

for individual i ($1 \leq i \leq N$) and measurement j ($1 \leq j \leq n_i$). In this equation, functions f and g are nonlinear functions of t_{ij}, a known continuous variable, usually time, and of an unknown random vector $\boldsymbol{\phi_i}$ of dimension $(d \times 1)$. Variable ϵ_{ij} is a residual term and is assumed to be normally distributed with mean zero and variance σ^2. In the case where f is the Brody function, for instance, and g is equal to 1, the model reduces to:

$$y_{ij} = A_i - B_i \, \mathrm{e}^{-K_i t_{ij}} + \epsilon_{ij} \tag{2}$$

where t_{ij} is the time of measurement and the individual vector of parameters is $\boldsymbol{\phi_i} = (A_i, B_i, K_i)$, which are biologically interpretable.

In the case of a genetic analysis, for an animal model, vector $\boldsymbol{\phi_i}$ for individual i is decomposed as follows:

$$\boldsymbol{\phi_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i u} + \boldsymbol{p_i} \tag{3}$$

where $\boldsymbol{X_i\beta}$ are the fixed effects influencing the curve parameters (A_i, B_i, K_i), $\boldsymbol{Z_i u}$ are the genetic effects and $\boldsymbol{p_i}$ are the permanent environmental effects. Matrices $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are known incidence matrices. It is assumed that $\boldsymbol{u}$ is normally distributed: $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{A} \otimes \boldsymbol{G})$, where matrix $\boldsymbol{G}$ is of dimension $d \times d$ (*i.e.*, (3×3) in the case of the Brody function) and represents the genetic covariance matrix between the curve parameters (A_i, B_i, K_i), and matrix $\boldsymbol{A}$ is the known genetic relationship matrix. The environmental vector $\boldsymbol{p_i}$ is also assumed normally distributed, with mean zero and covariance matrix $\boldsymbol{P}$, of dimension $d \times d$, which represents the environmental covariance matrix between the curve parameters. Let $\boldsymbol{\theta}$ be the vector of parameters to be estimated: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{P}, \sigma^2)$.

In the EM framework, a possible and convenient choice for the missing data is $\boldsymbol{z} = (\boldsymbol{\phi}, \boldsymbol{u})$. The likelihood of the complete data $p(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{u})$ can therefore be decomposed as follows: $p(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{u})p(\boldsymbol{\phi}|\boldsymbol{u})p(\boldsymbol{u})$.

2.2. The SAEM algorithm for genetic studies

The Stochastic EM algorithm was first introduced by Celeux and Diebolt [4], a Stochastic Approximation version was then proposed by Delyon *et al.* [6] and improved by Kuhn and Lavielle [14, 15].

The general idea of the algorithm is to replace the Expectation phase of the EM algorithm, *i.e.* the calculation of the conditional expectation of the likelihood of the complete data, by a stochastic approximation, since this expectation cannot be analytically calculated in the case of nonlinear mixed effects models.

At iteration $[k]$, let $Q(\boldsymbol{\theta})^{[k]}$ be the expectation function of the complete likelihood conditional on the observations $\boldsymbol{y}$ and the vector of parameters $\boldsymbol{\theta}$ estimated at iteration $[k-1]$.

$$Q(\boldsymbol{\theta})^{[k]} = E[\text{Log } p(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{u}; \boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta}^{[k-1]}]. \tag{4}$$

The key idea is to recycle variates generated from the previous iterations of the EM algorithm [17]. Therefore, instead of approximating $Q(\boldsymbol{\theta})^{[k]}$ by the arithmetic mean of L evaluations of the complete likelihood, *i.e.* $(1/L)\sum_{\ell=1}^{L} \text{Log } p(\boldsymbol{y}, \boldsymbol{\phi}^{[k,\ell]}, \boldsymbol{u}^{[k,\ell]}; \boldsymbol{\theta})$ calculated from L random draws of $\boldsymbol{\phi}$ and $\boldsymbol{u}$, as for a classical Monte Carlo EM algorithm, it is replaced by the following stochastic approximation:

$$Q(\boldsymbol{\theta})^{[k]} = Q(\boldsymbol{\theta})^{[k-1]} + \gamma_k \left(\frac{1}{L} \sum_{\ell=1}^{L} \text{Log } p(\boldsymbol{y}, \boldsymbol{\phi}^{[k,\ell]}, \boldsymbol{u}^{[k,\ell]}; \boldsymbol{\theta}) - Q(\boldsymbol{\theta})^{[k-1]} \right) \tag{5}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{u}$ are simulated according to the conditional distribution $p(.|\boldsymbol{y}, \boldsymbol{\theta}^{[k-1]})$, either directly or using a Metropolis-Hastings algorithm [14]. Kuhn and Lavielle [15] also showed that the convergence of the algorithm can be considerably improved by coupling it with an MCMC procedure, *i.e.* by simulating M Monte Carlo chains for $\boldsymbol{\phi}$ and $\boldsymbol{u}$, and averaging the observed likelihood values over the M chains. Thanks to the "recycling" process presented in the equation above and in contrast to the classical Monte Carlo EM (MCEM) algorithm, the number of chains (M) and of random draws within each chain (L) do not have to be very large. Five chains and 10 to 20 random draws within the chains are often sufficient in practice. These numbers are very small in contrast to the 200 to 5000 random deviates that are recommended by McCulloch [21] to ensure convergence of the MCEM algorithm. However, practical experience has shown that choosing $L = 1$ is often not sufficient to obtain a good accuracy of the parameter estimations.

Parameter γ_k is a crucial parameter in this estimating procedure. It performs a smoothing of the calculated likelihood values from one iteration to the other and therefore considerably accelerates convergence compared to other MCMC estimation procedures. In practice, this smoothing parameter is defined as follows. During the first K iterations, γ_k=1, *i.e.* there is no smoothing performed and the algorithm is equivalent to an MCEM algorithm [28]. McCulloch [21] showed that this algorithm converged very rapidly towards a neighborhood of the ML estimates but then continued showing a great deal of variation. Therefore, from iteration $(K + 1)$ the smoothing starts in order to stabilize the estimates and converges more rapidly towards the actual ML estimates [15]. Parameter γ_k is a sequence of stepsizes within the interval [0,1]. It is recommended [15] to take $\gamma_k = (k - K)^{-1}$ for $k \geq (K + 1)$. The choice of the iteration number K can depend on the number of simulations performed at each iteration. To ensure the algorithm has already converged into a neighborhood of the MLE before the smoothing starts, it is recommended to use this algorithm with several different starting values. A detailed description of the parameter estimation is given in the Appendix.

An advantage of the stochastic EM approach is that it remains in the classical maximum likelihood framework. It therefore allows the calculation of the likelihood value of the model using Importance Sampling and the calculation of the SE of the parameters using Louis' missing information principle [19] as presented by Lavielle [16]. This enables significance tests on the parameters (fixed effects and variance-covariance components) and also enables model comparisons using classical criteria such as likelihood ratio tests, AIC or BIC criteria.

A Matlab program is available for genetic analyses using the SAEM algorithm from the second author (Cristian.Meza@math.u-psud.fr).

3. EXAMPLES

3.1. Growth curve analysis in beef cattle

Data analyzed in this study came from an INRA experimental Charolais herd [23]. The data set comprised body weight records for 560 cows, born over an 11 year period (from 1988 to 1998), from 60 sires and 369 dams. Data were collected monthly from 1998 to 2003, but only 10 measurements from each animal were included being at around 0, 112, 224, 364, 540, 720, 900, 1260, 1620 and 1980 days. Although the same ages were considered for each animal, they were unequally spaced and some records were missing.

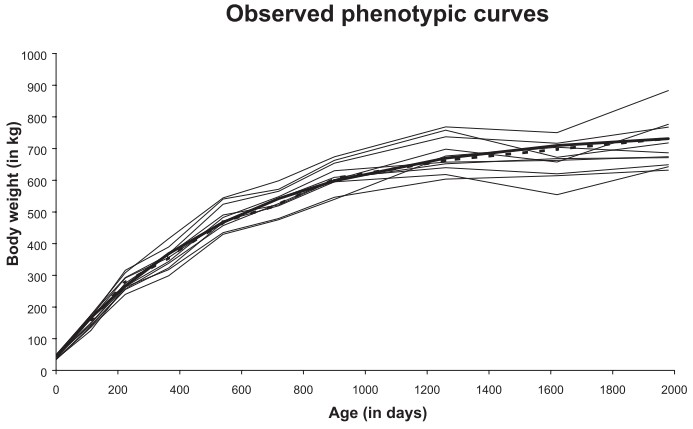

Figure 1. Example of 10 observed phenotypic growth curves for beef cattle. The bold and plain line represents the estimated mean curve obtained with the Brody function and the bold and dotted line is the observed mean curve calculated on the whole data set.

A Brody function was used to analyze these data and a sire model was considered. The model can be written as:

$$y_{ij} = A_i - B_i e^{-K_i t_j} + \epsilon_{ij} \tag{6}$$

where y_{ij} is the body weight measurement for individual i at time t_j (t_j corresponds to the ages of measurement divided by 100 000). The two individual parameters of this nonlinear function: A_i and K_i have an interesting biological interpretation. In fact, A_i represents the adult body weight for individual i and K_i is its maturation rate. A reparameterization was used for the B_i parameter of the Brody function such that $B_i = A_i - W0_i$, where $W0_i$ is the observed birth weight. The residual term ϵ_{ij} was assumed normally distributed with mean zero and constant variance σ^2. Parameters A_i and K_i were also assumed normally distributed and were decomposed using a sire model, as a special case of the animal model presented in the methodology section above (Eq. (3)).

3.1.1. Analysis with the SAEM algorithm

As shown in Figure 1, the Brody function is very appropriate to model the growth curves in beef cattle. Estimates obtained for each of the parameters are given in Table I. As expected, the genetic correlation between parameters A and K was quite high (−0.80). It still is, however, different from 1 which

Table I. Estimated genetic sire variances and correlation (VarG, CorrG) for the curve parameters A and K and permanent environmental variances and correlation (VarE, CorrE) for A and K with the SAEM algorithm for the beef cattle growth data using a Brody function. (In brackets are the SE of the parameters).

Fixed effects			
μ_A	761 (4.08)	μ_K	165 (1.32)
Variance components			
VarG_A	1270 (411)	VarE_A	4190 (309)
VarG_K	54.8 (21.9)	VarE_K	518 (43.2)
CorrG_{AK}	−0.80	CorrE_{AK}	−0.71
Residual variance		687 (14.4)	

Table II. Estimated genetic and environmental parameters with the SAEM algorithm for 400 simulated data sets with a sire model and the Brody function (θ_0 represents the starting values).

	μ_A	μ_K	VarG_A	VarG_K	CorrG_{AK}	VarE_A	VarE_K	CorrE_{AK}	σ^2
Simulated	760	165	1300	60	−0.80	4200	520	−0.72	690
θ_0	800	200	15000	6000	0.0	15000	6000	0.0	12869
Mean	760.2	164.9	1256.6	62.3	−0.80	4214.5	512.1	−0.72	690.4
Variance	31.9	2.26	103410	621.5	0.0121	106040	2275.1	0.0008	248.2
RMSE%	0.74	0.91	25.0	41.7	13.7	7.8	9.3	3.9	2.3

gives the possibility for a genetic selection for high growth rate while keeping a reasonable adult body weight, which is the goal of beef cattle breeders.

In order to check the accuracy of the SAEM estimates, we simulated 400 data sets with these parameter values. Table II provides the mean, variance and relative mean square error (RMSE) for each of these parameters over the 400 data sets. Estimations for all the simulated data sets were performed with 700 iterations, with the smoothing parameter starting after 400 iterations, 5 chains and 8 simulations per chain at each iteration (which corresponds to a total of 28 000 MC samples).

Analysis of these 400 data sets was performed using different starting values. The SAEM algorithm was found to be robust to the choice of starting values for the variance parameters. However, starting values for the fixed effects should be quite close to the real parameter values. Good initial values for the fixed effects can easily be obtained with the NLIN procedure of SAS®, for example. The algorithm was found to converge better when initial values for the variance components were larger than the expected ones.

A comparison with other nonlinear estimation procedures on these simulated data sets is difficult due to the computing time required by Bayesian analyses and the difficulty for approximated methods such as FOCE – First Order Conditional Estimation [18] to analyze any sampled data set arising from a simulation study. In addition, most softwares based on the Gaussian quadrature such as SAS® NLMIXED do not allow a random structure as complex as this one. Concerning the computing time, the phenotypic analysis of the real data set was performed with both the SAEM algorithm and the Gibbs sampling using the winBUGS program [27]. The SAEM algorithm converged and provided accurate parameter estimations in less than 4 min (for 700 iterations, 5 chains and 8 simulations per chain), whereas the Gibbs sampling required at least 50 000 iterations, which took about 30 min to run.

3.1.2. Model comparisons

A previous study showed that the structured antedependence (SAD) models performed well to analyze this growth pattern compared to the classical random regression (RR) models [13]. The aim is now to compare these models and the proposed nonlinear approach. Model comparison was based on the likelihood values and the BIC criterion, which was calculated using the following formula: BIC = -2 LogL + n_c Log(N) where -2 LogL is minus twice the log-likelihood value, n_c is the number of covariance parameters in the model and N is the total number of observations. Notice that N in the previous formula has to be replaced by (N-p) (where p is the number of fixed effects, also equal to rank (**X**)) in the case of REML estimation.

In order to compare the different methodologies, the same mean curve was used as fixed effects, *i.e.* the Brody curve presented above ($f(t) = a - b\exp(-kt)$). For the SAD and RR models, since the nonlinear parameter k could not be estimated with ASREML [10], the value obtained with the SAEM algorithm was used. The aim was to compare the flexibility of the three approaches to model the covariance structure. To do so, the variances and correlations were calculated at each of the 10 ages with the three methods (SAD, RR, Brody). Since no analytical form is available for a nonlinear model to calculate the variance and correlation functions, they were calculated by simulations. In order to have a 'reference' model, this analysis was performed in the phenotypic case, and the three estimated covariance structures were compared to a completely unstructured model.

To make sure the likelihood values were comparable, the 10 by 10 phenotypic covariance matrix was calculated with the parameters obtained with each

Table III. Likelihood values and BIC criterion for the phenotypic analysis (the smaller are the values the better is the model). 'US' is the completely unstructured model with a 10 by 10 estimated covariance matrix; 'SAD2 quad-const' corresponds to a second order structured antedependence model with a quadratic first order antedependence parameter and constant second order; 'RR cubic' corresponds to a random regression model based on a polynomial of order 3. Nb Par Cov is the number of parameters in the covariance structure. To make the model comparisons easier a constant ($c = -40\,000$) was added to all the likelihood values.

Model	Nb Par Cov	−2 LogL	BIC
US	55	901.6	1374.8
SAD2 quad-const	7	1592.2	1652.4
RR cubic	11	2732.2	2826.8
BRODY	4	3382.4	3416.8

of the models and fixed in ASREML (for US, SAD, RR and Brody) to obtain the likelihood values.

Table III gives the likelihood values and BIC criterion for the different models. The unstructured model (US) was found here to have the smallest BIC value and is considered as the 'reference' model. It was found that although the nonlinear shape of the curve is very appropriate to model the phenotypic growth phenomenon, it is less flexible than the structured antedependence and even the cubic random regression model to fit the covariance structure. In fact, as shown in Figure 2, the Brody model did not fit the correlation pattern very well; the estimated correlations were underestimated at early ages and slightly overestimated at late ages. Similarly, the phenotypic variance shown in Figure 3 was overestimated at early ages and underestimated at late ages. On the contrary, although the likelihood value and BIC criterion were higher for the cubic random regression model than for the Brody function, Figure 2 shows that the use of the nonlinear Brody function avoided the main drawbacks of the random regression models based on polynomial functions, which are the border effects.

The Brody model also requires the estimation of only very few parameters and allows the direct prediction of individual genetic values for the adult body weight and the maturation rate, which is quite difficult to define with other longitudinal models.

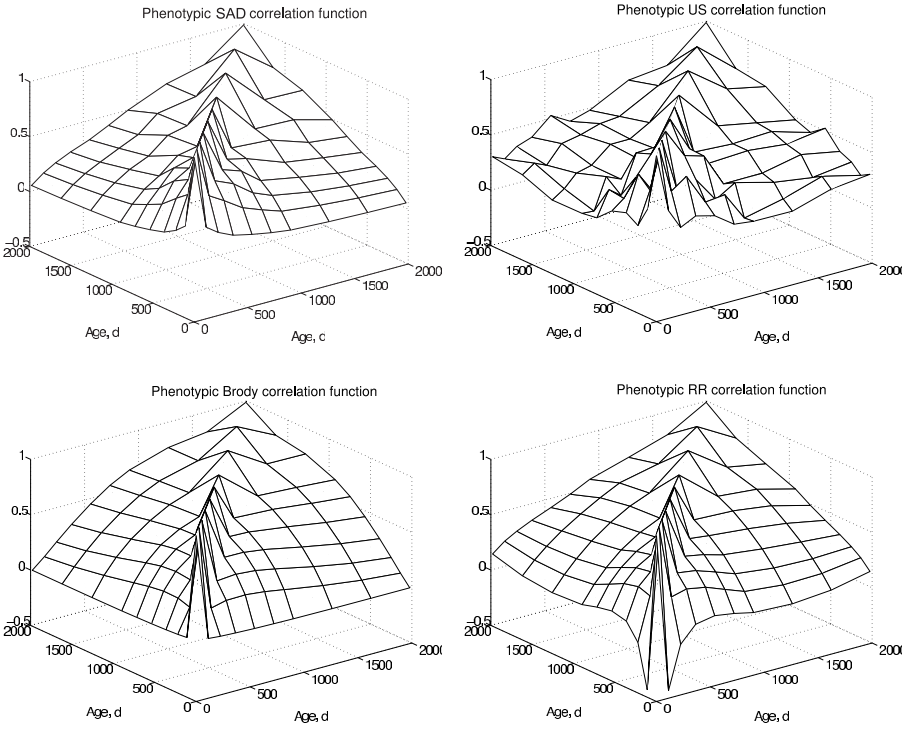

Figure 2. Estimated phenotypic correlation functions obtained with the unstructured (US), SAD, RR and Brody models presented in Table III.

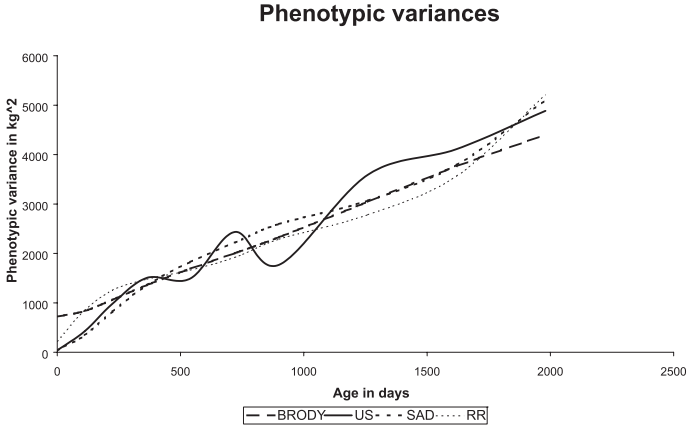

Figure 3. Estimated phenotypic variance functions obtained with the unstructured (US), SAD, RR and Brody models presented in Table III.

3.2. Growth curve analysis in the chicken

This data set corresponds to the last generation of selection from the experiment presented by Mignon-Grasteau *et al.* [24]. Data originated from a selection experiment on the form of the growth curve initiated by F. Ricard in 1960 on meat-type chickens. Line X+− was selected for high juvenile body weight at 8 weeks and low adult body weight at 36 weeks. In contrast, line X−+ was selected for low juvenile body weight and high adult body weight. In line X++, chickens were selected for high body weights at both ages and, in the opposite line, X− −, they were selected for low body weights at both ages. Line X00 was an unselected control line. The data set comprised in total 265 chickens, from 71 sires, and about 12 measurements for each animal at ages 0, 4, 6, 8, 12, 16, 20, 24, 28, 32, 36 and 40 weeks. Only chickens with more than 5 measurements were included in the analyses.

The Gompertz function was used and the model can be written as follows:

$$y_{ij} = A_i \exp(-B_i \exp(-K_i t_j)) + \epsilon_{ij} \tag{7}$$

where A_i is the asymptotic body weight of chicken i, *i.e.* the weight at an infinite age. Parameter B_i is equal to $\ln(A_i/W_{i0})$ where W_{i0} is the estimated hatching weight of chicken i. Parameter K_i corresponds to the maturation rate, *i.e.* the rate at which the animal approaches its asymptotic weight. In this equation, the times t_j correspond to the ages of measurement listed above divided by 100. A sire model was used for each of these three parameters, and the different lines were fitted as fixed effects. As before, the three parameters of the curve were assumed normally distributed and correlated. The residuals ϵ_{ij} were also assumed normally distributed with mean zero and constant variance σ^2.

As mentioned in the methodology section, the SAEM approach allows to perform significance tests on the parameters. Using a likelihood ratio test, it was found that the environmental covariance between parameters A and B of the Gompertz function was not significant. It was therefore set to zero.

On the contrary, it was found that the line effects were all significantly different for the three parameters of the curve. As expected and as shown in Table IV, the mean effect for parameter A, *i.e.* the asymptotic body weight, was found to be the highest for lines X++ and X−+, and the lowest for lines X−− and X+−. On the contrary, the maturation rate (parameter K) was found to be the lowest for line X−+ and the highest for line X+−.

Table V provides the estimated genetic and environmental variance and correlation parameters. Since they were calculated only on the last generation of selection, they were found to be different from the results obtained by Mignon-Grasteau *et al.* [24].

Table IV. Estimated fixed effects with the SAEM algorithm for the chicken growth data using a sire model and the Gompertz function. (In brackets are the SE of the parameters).

	Line X−+	Line X+−	Line X++	Line X− −	Line X00
μ_A	3070 (49.4)	1960 (47.3)	3110 (44.0)	1750 (41.0)	2350 (30.2)
μ_B	4.73 (0.0971)	3.36 (0.203)	4.36 (0.13)	4.39 (0.0502)	3.72 (0.0573)
μ_K	12.7 (0.454)	16.7 (0.811)	16.5 (0.586)	15.4 (0.332)	14.8 (0.29)

Table V. Estimated genetic and environmental variances and correlations obtained with the SAEM algorithm for the chicken growth data using a sire model and the Gompertz function. On the diagonal are the variances and off-diagonal are the correlations. (In brackets are the SE of the parameters).

Genetic components			
A	6220 (6960)	−0.12	−0.36
B		0.0428 (0.00866)	0.96
K			1.28 (0.128)
Environmental components			
A	49000 (7450)	0	−0.35
B		0.0194 (0.015)	0.88
K			2.25 (1.58)
Residual variance			8970 (296.0)

The likelihood values were used to compare the Gompertz curve with two other nonlinear curves: the Logistic function and the Brody function, in a phenotypic analysis. The Brody function was defined as in equation (2) and the Logistic function was:

$$y_{ij} = \frac{A_i}{1 + B_i \exp(-K_i t_{ij})} + \epsilon_{ij}. \tag{8}$$

The three nonlinear curves had the same number of parameters, and the likelihood (−2 LogL) values obtained were 696 for the Gompertz function, 1112 for the Logistic function and 3776 for the Brody function (a constant $c = 46\,000$ was added to the three likelihood values to make them more easily comparable). As expected, it was found that the Gompertz function was more appropriate to model this growth phenomenon. It is useful, however, to have a likelihood criterion for nonlinear model comparisons when a less well known character is analyzed. Any nonlinear function can be defined in the available SAEM program.

Phenotypic analyses of these data with the Gompertz function were also performed with winBUGS [27], for a Bayesian Gibbs Sampling analysis. Many convergence problems were encountered, especially for fitting different line effects for the B parameter, and the algorithm showed a great sensitivity to the choice of the prior distributions. On the contrary, the SAEM algorithm proved to be more robust to the choice of starting values and showed a much faster convergence.

4. DISCUSSION

The Stochastic Approximation EM (SAEM) algorithm presented in this paper is conceptually very simple and has several advantages compared to a classical Monte Carlo EM algorithm [28]. Firstly, thanks to the "recycling" of the simulated values from one iteration to the next, it considerably reduces the number of Monte Carlo simulations required. Secondly, the smoothing parameter considerably accelerates convergence to the MLE. A comparison of the SAEM algorithm with approximated estimation procedures such as First Order Conditional Estimation (FOCE), Laplacian methods or the Gaussian quadrature [5] was performed by Kuhn and Lavielle [15]. The SAEM algorithm was found to perform better than the other methods in terms of robustness with regards to the choice of the starting values, especially for the variance components, and accuracy of the estimates. It is also much faster to converge than classical Bayesian methods using the Gibbs sampling. These properties of the SAEM algorithm were confirmed here in the simulation study. The SAEM algorithm is implemented in a specialized software for the phenotypic analysis of nonlinear mixed effects models called "Monolix", which can be freely downloaded from the following address: `http://www.math.u-psud.fr/~lavielle/monolix/logiciels`. A Matlab program for the sire model extension is available from the second author.

Another advantage of the stochastic EM algorithm is that it remains within the maximum likelihood framework, and therefore allows to use classical model comparison criteria such as AIC or BIC. It is possible, in particular, to compare nonlinear mixed models to other longitudinal models such as random regression or structured antedependence models. In this study, for example, it was found that the structured antedependence models [13, 25] were able to better fit the covariance structure than the nonlinear Brody function. This shows that it might be necessary to define more flexible nonlinear functions for growth curves, which would still have interpretable parameters in terms of

adult body weight and maturation rates, but would have additional parameters to better capture the variance and correlation patterns of the data. For example, functions defined by differential equations might be more appropriate. Indeed, extension of the SAEM algorithm for differential equation models is under investigation for phenotypic analyses. It was also found that, although mathematically equivalent, different parameterizations of the growth curve models (Brody, Gompertz, Richards) may improve convergence.

The aim of this paper was to present this novel and efficient estimation procedure, namely the SAEM algorithm. It was applied here for the genetic analysis of nonlinear longitudinal characters such as growth traits. This algorithm is, however, very general and can also be extended for estimation in the context of mixture models, for the classification of genes with regards to their expression profile dynamics, for example. Or, it can be used for inference in generalized linear mixed models (GLMM), for the analysis of categorical traits such as fertility, or the joint analysis of discrete and continuous variables for the genetic analysis of disease resistance characters. Another extension of the SAEM algorithm could also be for QTL detection for nonlinear traits, such as growth trajectories [20], or for QTL detection of discrete traits such as disease resistance characters.

It was found that the speed of convergence of the SAEM algorithm can be improved by the use of a PX modification [8,22]. This proved to be particularly efficient during the first iterations, when the parameters were highly correlated, as is the case for growth curve models. A REML extension of the SAEM algorithm is under development in the phenotypic case and proved to improve the accuracy of the variance parameter estimates in similar proportions as in linear mixed models.

ACKNOWLEDGEMENTS

The beef cattle data were provided by the INRA experimental center located near Bourges, France. We are most grateful to Eric Venot, Gilles Renand for interesting discussions on the beef cattle data analysis, and to Sandrine Mignon-Grasteau for kindly providing the chicken growth data. Thank you to two anonymous referees for useful comments.

REFERENCES

[1] Atchley W.R., Zhu J., Developmental quantitative genetics, conditional epigenetic variability and growth in mice, Genetics 147 (1997) 765–776.

[2] Blasco A., Piles M., Varona L., A Bayesian analysis of the effect of selection for growth rate on growth curves in rabbits, Genet. Sel. Evol. 35 (2003) 21–41.

[3] Booth J.G., Hobert J.P., Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, J. R. Statist. Soc. B. 61 (1999) 265–285.

[4] Celeux G., Diebolt J., The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, Comp. Statist. Quat. 2 (1985) 73–82.

[5] Davidian M., Giltinan D., Nonlinear models for repeated measurement data: an overview and update, J. Agric. Biol. Env. Satist. 4 (2003) 387–419.

[6] Delyon B., Lavielle M., Moulines E., Convergence of a stochastic approximation version of the EM algorithm, Ann. Statist. 27 (1999) 94–128.

[7] Diggle P.J., Liang K.Y., Zeger S.L., Analysis of longitudinal data, Oxford University Press, Oxford, 1994.

[8] Foulley J.L., van Dyk D.A., The PX EM algorithm for fast fitting of Henderson's mixed model, Genet. Sel. Evol. 32 (2000) 143–163.

[9] Geyer C.J., On the convergence of Monte Carlo maximum likelihood calculations, J. R. Statist. Soc. B. 56 (1994) 261–274.

[10] Gilmour A.R., Thompson R., Cullis B.R., Welham S.J., ASREML Manual, New South Wales Department of Agriculture, Orange, Australia, 2002.

[11] Huisman A.E., Veerkamp R.F., van Arendonk J.A.M., Genetic parameters for various random regression models to describe weight data of pigs, J. Anim. Sci. 80 (2002) 575–582.

[12] Jaffrézic F., Thompson R., Hill W.G., Structured antedependence models for genetic analysis of multivariate repeated measures in quantitative traits, Genet. Res. 82 (2003) 55–65.

[13] Jaffrézic F., Venot E., Laloë D., Vinet A., Renand G., Use of structured antedependence models for the genetic analysis of growth curves, J. Anim. Sci. 82 (2004) 3465–3473.

[14] Kuhn E., Lavielle M., Coupling a stochastic approximation version of EM with a MCMC procedure, ESAIM Prob. Statist. 8 (2004) 115–131.

[15] Kuhn E., Lavielle M., Maximum likelihood estimation in nonlinear mixed effects models, Comput. Statist. Data Anal. 49 (2005) 1020–1038.

[16] Lavielle M., Monolix User Guide Manual, 2005, http://www.math.u-psud.fr/~lavielle/monolix/logiciels.

[17] Levine R.A., Casella G., Implementations of the Monte Carle EM algorithm, J. Comp. Graph. Statist. 10 (2001) 1–18.

[18] Lindstrom M.J., Bates D.M., Nonlinear mixed-effects models for repeated measures data, Biometrics 46 (1990) 673–687.

[19] Louis T.A., Finding the observed information matrix when using the EM algorithm, J. R. Statist. Soc. B. 44 (1982) 226–233.

[20] Ma C.-X., Casella G., Wu R.L., Functional mapping of quantitative trait loci underlying the character process: a theoretical framework, Genetics 161 (2002) 1751–1762.
[21] McCulloch C.E., Maximum likelihood algorithms for generalized linear mixed models, J. Am. Statist. Assoc. 92 (1997) 162–170.
[22] Meng X.L., van Dyk D.A., Fast EM-type implementations for mixed effects models, J. R. Statist. Soc. B. 60 (1998) 559–578.
[23] Mialon M.M., Renand G., Krauss D., Ménissier F., Variability of the postpartum recovery of sexual activity of Charolais cows, Livest. Prod. Sci. 69 (2001) 217–226.
[24] Mignon-Grasteau S., Piles M., Varona L., de Rochambeau H., Poivey J.P., Blasco A., Beaumont C., Genetic analysis of growth curve parameters for male and female chickens resulting from selection on shape of growth curve, J. Anim. Sci. 78 (2000) 2515–2524.
[25] Nunez-Anton V., Zimmerman D.L., Modeling non-stationary longitudinal data, Biometrics 56 (2000) 699–705.
[26] Pletcher S.D., Jaffrézic F., Generalized character process models: estimating the genetic basis of traits that cannot be observed and that change with age or environmental conditions, Biometrics 58 (2002) 157–162.
[27] Spiegelhalter D.J., Thomas A., Best N.G., WinBUGS Version 1.4 User Manual, Cambridge: Medical Research Council Biostatistics Unit, http://www.mrc-bsu.cam.ac.uk/bugs, 2004.
[28] Wei G.C.G., Tanner M.A., A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, J. Am. Statist. Assoc. 85 (1990) 699–704.

APPENDIX
PARAMETERS ESTIMATION WITH THE SAEM ALGORITHM

The complete likelihood function of the model defined by equations (1) and (3) can be written as:

$$\text{Log } p(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{u};\boldsymbol{\theta}) = \text{Log } p(\boldsymbol{y}|\boldsymbol{\phi},\boldsymbol{u};\boldsymbol{\theta}) + \text{Log } p(\boldsymbol{\phi}|\boldsymbol{u};\boldsymbol{\theta}) + \text{Log } p(\boldsymbol{u};\boldsymbol{\theta})$$

$$\begin{aligned}\text{Log } p(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{u};\boldsymbol{\theta}) = &-\frac{N_{tot}+dN+N_a}{2}\text{Log}(2\pi) - \frac{N_{tot}}{2}\text{Log}\sigma^2 \\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij}-f(t_{ij},\boldsymbol{\phi_i}))^2 - \frac{N}{2}\text{Log}(|\boldsymbol{P}|) \\ &-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\phi_i}-\boldsymbol{Z_i u}-\boldsymbol{X_i\beta})'\boldsymbol{P^{-1}}(\boldsymbol{\phi_i}-\boldsymbol{Z_i u}-\boldsymbol{X_i\beta}) \\ &-\frac{N_a}{2}\text{Log}(|\boldsymbol{\Gamma}|) - \frac{1}{2}\boldsymbol{u'\Gamma^{-1}u}\end{aligned}$$

where N is the number of individuals with observations, $N_{tot} = \sum_{i=1}^{N} n_i$ is the total number of observations, d is the dimension of vector $\boldsymbol{\phi_i}$, for all individuals i ($d = 3$ for a Brody function for example: $\boldsymbol{\phi_i} = (A_i, B_i, K_i)$), and N_a is the number of animals in the relationship matrix. Let $\boldsymbol{\Gamma} = \boldsymbol{A} \otimes \boldsymbol{G}$ be the genetic covariance matrix.

In the E step, the conditional expectation of the complete likelihood is calculated: $E(\text{Log } p(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{u}; \boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta} = \boldsymbol{\theta}^{[k]})$.

The M step consists of maximizing this conditional expectation. By deriving it with respect to each of the parameters, it follows that:

$$\begin{aligned}
\boldsymbol{\beta}^{[k]} &= \sum_{i=1}^{N} (\boldsymbol{X_i'} \boldsymbol{P^{-1}} \boldsymbol{X_i})^{-1} E_c \left[\sum_{i=1}^{N} \boldsymbol{X_i'} \boldsymbol{P^{-1}} (\boldsymbol{\phi_i}^{[k]} - \boldsymbol{Z_i} \boldsymbol{u}^{[k]}) \right] \\
&= \sum_{i=1}^{N} (\boldsymbol{X_i'} \boldsymbol{P^{-1}} \boldsymbol{X_i})^{-1} S_1^{[k]} \\
\boldsymbol{G}_{(l,m)}^{[k]} &= \frac{E_c(\boldsymbol{u_l'} \boldsymbol{A^{-1}} \boldsymbol{u_m})}{N_a} = \frac{S_{2(l,m)}^{[k]}}{N_a} \quad \text{for } l, m = 1, ..., d
\end{aligned}$$

where $\boldsymbol{u_l}$ and $\boldsymbol{u_m}$ are of dimension $(N_a \times 1)$.

$$\begin{aligned}
\boldsymbol{P}_{(l,m)}^{[k]} &= \frac{E_c[\sum_{i=1}^{N} (\boldsymbol{\phi_{i,l}}^{[k]} - \boldsymbol{Z_{i,l}} \boldsymbol{u_l}^{[k]} - \boldsymbol{X_{i,l}} \boldsymbol{\beta_l}^{[k]})' (\boldsymbol{\phi_{i,m}}^{[k]} - \boldsymbol{Z_{i,m}} \boldsymbol{u_m}^{[k]} - \boldsymbol{X_{i,m}} \boldsymbol{\beta_m}^{[k]})]}{N} \\
&= \frac{S_{3(l,m)}^{[k]}}{N} \\
\sigma^{2[k]} &= \frac{E_c(\sum_{i,j} (y_{ij} - f(t_{ij}, \boldsymbol{\phi_i}^{[k]}))^2)}{N_{tot}} = \frac{S_4^{[k]}}{N_{tot}}
\end{aligned}$$

where E_c denotes the conditional expectation $E(.|\boldsymbol{y}, \boldsymbol{\theta} = \boldsymbol{\theta}^{[k]})$.

In the SAEM algorithm, the above conditional expectations are replaced with the following stochastic approximations (for one Markov chain and at

iteration k):

$$S_1^{[k]} = S_1^{[k-1]} + \gamma_k \left[\sum_{i=1}^{N} \boldsymbol{X_i' P^{-1}} (\boldsymbol{\phi_i}^{[k]} - \boldsymbol{Z_i u}^{[k]}) - S_1^{[k-1]} \right]$$

$$S_{2(l,m)}^{[k]} = S_{2(l,m)}^{[k-1]} + \gamma_k [\boldsymbol{u_l' A^{-1} u_m} - S_{2(l,m)}^{[k-1]}]$$

$$S_{3(l,m)}^{[k]} = S_{3(l,m)}^{[k-1]} + \gamma_k \left[\sum_{i=1}^{N} (\boldsymbol{\phi_{i,l}}^{[k]} - \boldsymbol{Z_i u_l}^{[k]} - \boldsymbol{X_i \beta_l}^{[k]})' (\boldsymbol{\phi_{i,m}}^{[k]} - \boldsymbol{Z_i u_m}^{[k]} - \boldsymbol{X_i \beta_m}^{[k]}) - S_{3(l,m)}^{[k-1]} \right]$$

$$S_4^{[k]} = S_4^{[k-1]} + \gamma_k \left[\sum_{i,j} (y_{ij} - f(t_{ij}, \boldsymbol{\phi_i}^{[k]}))^2 - S_4^{[k-1]} \right]$$

where $\boldsymbol{\phi}^{[k]}$ and $\boldsymbol{u}^{[k]}$ are simulated according to the conditional distribution $p(.|\boldsymbol{y}, \boldsymbol{\theta}^{[k-1]})$ either directly or using a Metropolis-Hastings algorithm [14].