Electronic Journal of Statistics
Vol. 8 (2014) 456–475
ISSN: 1935-7524
DOI: 10.1214/14-EJS890

A note on BIC in mixed-effects models

Maud Delattre

AgroParisTech, UMR 518 MIA, F-75005 Paris, France
INRA, UMR 518 MIA, F-75005 Paris, France
e-mail: maud.delattre@agroparistech.fr

Marc Lavielle

Inria Saclay – Île-de-France
Université Paris-Sud, Laboratoire de mathématiques UMR 8628, Orsay F-91405
e-mail: marc.lavielle@inria.fr

and

Marie-Anne Poursat

Université Paris-Sud, Laboratoire de mathématiques UMR 8628, Orsay F-91405
e-mail: marie-anne.poursat@math.u-psud.fr

Abstract: The Bayesian Information Criterion (BIC) is widely used for variable selection in mixed effects models. However, its expression is unclear in typical situations of mixed effects models, where simple definition of the sample size is not meaningful. We derive an appropriate BIC expression that is consistent with the random effect structure of the mixed effects model. We illustrate the behavior of the proposed criterion through a simulation experiment and a case study and we recommend its use as an alternative to various existing BIC versions that are implemented in available software.

MSC 2010 subject classifications: Primary 62J02, 62J12.
Keywords and phrases: Bayesian Information Criterion, BIC, mixed effects model, variable selection.

Received December 2012.

Contents

1. Introduction

Mixed effects models are typically used in population studies where repeated measurements are observed from several independent subjects (see [3, 22] for instance). Population studies are relevant in various fields such as pharmacokinetics or public health, for example to study a disease progression and to evaluate the effect of a treatment or the impact of physiological covariates [3, 4, 24]. The particular data structure in population approaches raises difficulties with the maximum likelihood (ML) approach. Since the likelihood cannot be expressed in a closed form, model fitting is a complicated computational issue. There is an extensive literature devoted to the development of powerful algorithms based on likelihood linearization [22] or on stochastic versions of the EM algorithm [7, 11]. Available software packages (Monolix, NLMIXED in SAS, saemix and nlme in R, nlmefitsa in Matlab, among others) allow to fit a wide range of generalized linear and nonlinear mixed models. For model selection purpose, the Bayesian Information Criterion (BIC) provides a consistent and easy-to-use method [27]. Its generic form is the negative maximized log-likelihood penalized by a term that depends on the number of estimated parameters and the sample size. Surprisingly, the BIC expression differs from one software to another. The reason is that the effective sample size and the effective number of parameters are not clearly defined in this context. The aim of the paper is to clarify how the BIC should be defined in general mixed effects models.

1.1. Model formulation

To fix the notations, let N be the number of subjects and let $y_i = (y_{i1}, \dots, y_{i,n_i})'$ be the $n_i \times 1$ vector of observations for subject i, where $y_{ij},\ j = 1, \dots, n_i$ denotes the jth measure observed under time condition t_{ij} and possible additional conditions u_{ij} (such as drug doses for instance). For simplicity, we assume in this paper that the number of repeated measurements n_i is the same for each subject: $n_i \equiv n$, $i = 1, \dots, N$, but extension of our results to different n_is is straightforward. We write $x_{ij} = (t_{ij}, u_{ij})$ the values of the so-called regression variables or design variables for subject i and we denote $x_i = (x_{i1}, \dots, x_{i,n})$. Population studies try to explain the variability of the observed profiles (x_i, y_i), $i = 1, \dots, N$, and to determine if some of the variation is associated with subject characteristics c_i that do not change with time (such as weight or age for instance).

Assuming that the N subjects are mutually independent, a mixed effects model may be presented hierarchically as a two-stage model (see for instance [3, 13, 22]):

- *Stage 1 (observations):* for each individual i, the probability distribution of the observations y_i depends on a set of regression variables x_i and a $d \times 1$ vector of individual parameters ψ_i

$$y_i | \psi_i \sim p(\cdot | \psi_i; x_i); \tag{1.1}$$

- *Stage 2 (individual parameters):* the inter-individual variability is modeled by considering ψ_i as a random vector and by introducing in the model covariates c_i
$$\psi_i \sim p(\cdot|\theta; c_i), \tag{1.2}$$
where θ is a vector of population parameters.

A linear mixed effects model assumes a linear relationship between y_i and ψ_i at stage 1
$$y_i = F_i\psi_i + \varepsilon_i, \tag{1.3}$$
where $F_i = F(x_i)$ is a $n \times d$ design matrix and where ε_i is a $n \times 1$ vector of normally distributed residual errors: $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \Sigma)$. A linear model for ψ_i is also assumed at stage 2
$$\psi_i = C_i\beta + D_i b_i, \tag{1.4}$$
where β is a $m \times 1$ vector of fixed-effects, b_i a $\ell \times 1$ vector of Gaussian random effects: $b_i \sim_{i.i.d.} \mathcal{N}(0, \Gamma)$. Here, $C_i = C(c_i)$ and $D_i = D(c_i)$ are $d \times m$ and $d \times \ell$ matrices.

By combining (1.3) and (1.4), we obtain the general representation of a linear mixed effects model [13]:
$$y_i = A_i\beta + B_i b_i + \varepsilon_i,$$
where $A_i = F_iC_i$ and $B_i = F_iD_i$ are $n \times m$ and $n \times \ell$ design matrices. We give an example of such linear mixed effects model in the simulation study Section 3.

Formulation (1.1)–(1.2) includes very general models such as nonlinear mixed effects models for continuous data
$$y_{ij} = f(x_{ij}; \psi_i) + \sigma(x_{ij}; \psi_i)\varepsilon_{ij},$$
where f describes the structural model and σ models the variability of the residual errors ε_{ij} [3, 22]. Section 4 provides a typical example of a pharmacokinetic study for which the nonlinear mixed effects model is an appropriate framework. Model (1.1)–(1.2) can also consider generalized linear mixed models for categorical, count or survival data [17, 26]. In this context, the first stage essentially consists in defining the conditional mean $\mathbb{E}(y_{ij}|\psi_i; x_{ij})$ as a function $\mu(x_{ij}, \psi_i)$ which depends on the fixed and random effects through a link function and a linear predictor:
$$g(\mathbb{E}[y_i|\psi_i; x_i]) = A_i\beta + B_i b_i.$$

Our main result presented in Section 2 is obtained under the hypothesis that the individual parameters ψ_i are normally distributed. For a sake of simplicity in the notations, we will denote $\eta_i = D_ib_i$ in the sequel. Then, (1.4) reduces to
$$\psi_i = C_i\beta + \eta_i, \tag{1.5}$$
where $\eta_i \sim_{i.i.d.} \mathcal{N}(0, \Omega)$ and where $\Omega = (\Omega_{k,k'})_{1\le k,k'\le d}$ is a $d \times d$ (possibly degenerate) variance-covariance matrix. The vector of population parameters θ includes β and the parameters in Ω.

Equation (1.5) can consider models for which certain individual parameters are random or purely fixed. Degenerate matrices Ω may have the following block-diagonal structure:

$$\Omega = \begin{pmatrix} 0 & 0 \\ 0 & \Omega_R \end{pmatrix}, \tag{1.6}$$

where Ω_R is a $d_R \times d_R$ positive-definite variance-covariance matrix, with $d_R \leq d$.

The basic model defined in (1.5) for the individual parameters can be extended to more general models. First, we can assume that there exists a monotonic transformation h such that $\phi_i = h(\psi_i) = C_i\beta + \eta_i$. For instance, the log-transformation is frequently used for non negative parameters and the logit transformation for parameters such as proportions which are known to take their values between 0 and 1. Equation (1.5) can also be extended to nonlinear Gaussian models such as $\phi_i = h(\psi_i) = \mu(\beta, c_i) + \eta_i$, where μ is a possibly nonlinear function of c_i and β. Indeed, the proof of our results use of the parametrization that involves the ϕ_i's instead of the ψ_i's, and does not make any assumption of linearity for the function μ.

1.2. Covariate selection

In this paper, we focus on the covariate selection problem, *i.e.* the selection of the non zero elements of β. In equation (1.5), our model formulation emphasizes that this variable selection problem consists in choosing the most relevant components in the c_i's for describing the between-subjects variability of the individual parameters of the model.

To clarify the variable selection problem tackled in the present work, we consider a basic example based on the following linear mixed-effects model

$$y_{ij} = \psi_{i0} + \psi_{i1}\, t_{ij} + \varepsilon_{ij},\ i = 1, \ldots, N,\ j = 1, \ldots, n,$$

specified by the individual parameters $\psi_i = \left(\psi_{i0}, \psi_{i1}\right)'$ and the regression variables $x_{ij} = t_{ij}$, $i = 1, \ldots, N$, $j = 1, \ldots, n$, with a Gaussian measurement noise $\varepsilon_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. We assume that the slope may vary with one covariate c_i, so that $\psi_i = C_i\beta + \eta_i$ with

$$C_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & c_i \end{pmatrix};\ \beta = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \alpha_1 \end{pmatrix};\ \eta_i \sim_{i.i.d.} \mathcal{N}\left(0, \Omega\right),\ \Omega = \begin{pmatrix} \omega_0^2 & 0 \\ 0 & \omega_1^2 \end{pmatrix}.$$

The decision on whether or not to include the covariate c_i in the model can be based on the BIC of the fits with and without it. Our goal is to compute the appropriate BIC; to answer this question, we assume that we know which of the parameters are purely fixed or random, *i.e.* which diagonal element of Ω is null or not. The problem of selecting the variance-covariance structure, *i.e.* the non zero elements of Ω is beyond the scope of the paper.

The BIC is defined as $-2\mathcal{LL} + \text{pen}$ where $\mathcal{LL}$ is the maximized log-likelihood and pen is a term equal to the product of the number of estimated parameters and the logarithm of the sample size; this term is referred to as the *BIC penalty*. In a pure fixed effects model where all the observations $y_{ij},\ i = 1, \dots, N,\ j = 1, \dots, n_i$ are independent, the effective sample size is the total number of observations $n_{\text{tot}} = \sum_{i=1}^{N} n_i$. In a pure random effects model where all the components of ψ_i are random, the estimation of the fixed effects β is based on N independent random vectors $(y_1, \dots, y_N)$. In such situation, the effective sample size is the number of subjects N. In a mixed effects model which combines fixed and random components of ψ_i, the definition of the effective sample size is not straightforward. Yet, in the mixed effects model literature, the BIC penalty usually involves $\log n_{\text{tot}}$. From a practical point of view, the $\log n_{\text{tot}}$ penalty is implemented in the R package nlme [23] and in the SPSS procedure MIXED [28] while the $\log N$ penalty is used in Monolix [18], saemix [2] or in the SAS proc NLMIXED [25].

From a theoretical point of view, the BIC penalty relies on asymptotic approximations. Nie [20] showed that convergence rates of maximum likelihood estimators can differ from parameter to parameter according to the level of variability designed in the mixed model. Based on the work of [20], we consider the double-asymptotic framework where both the number of subjects N and the numbers of measurements per subject n_i, $i = 1, \dots, N$ tend to infinity. We use an appropriate decomposition of the complete log-likelihood combined with the Laplace approximation to derive the asymptotic BIC approximation. We obtain a BIC penalty based on two terms proportional to $\log N$ and $\log n_{\text{tot}}$ that adapts to the mixed effects structure of the model.

BIC-type procedures were studied by Jiang and Rao [9] in linear mixed effects models. Jiang et al. [10] pointed out the difficulties encountered with these criteria in non conventional situations including mixed models and introduced new strategies called fence methods. Recently, several authors clarified how the number of parameters should be chosen to define a conditional AIC in mixed models [8, 15, 16, 29]. Bondell et al. [1], Lai et al. [12] and Schelldorfer and Bühlmann [26] studied the problem of fixed and random effects selection with a lasso-type penalized likelihood approach in different mixed effects models. In the three procedures, a BIC step was implemented to choose the regularization parameters. The three papers used different BICs with different choices of effective sample sizes and effective degrees of freedom. Our work gives an answer on how to define the BIC accounting for the covariance structure in the model.

The rest of the article is organized as follows. We describe in Section 2 the asymptotic framework of our study and give the theoretical penalty term to select covariates using BIC in mixed models. Sections 3 and 4 are devoted to a simulation experiment and a case study. We illustrate the performance of the proposed BIC and compare it with standard BIC criteria that are implemented in available softwares. The article concludes with a discussion in Section 5.

2. BIC for mixed-effects models

Denote by $\mathbf{y} = (y_1, \dots, y_N)$ the N vectors of $n_i \equiv n$ observations. Then, the total number of observations is $n_{\text{tot}} = \sum_i n_i = Nn$. From here on, the c_i's and the x_i's will be omitted in the notations of the probability distributions. The conditional distribution of y_i will be denoted $p(\cdot|\psi_i)$ rather than $p(\cdot|\psi_i; x_i)$ and the distribution of ψ_i will be denoted $p(\cdot|\theta)$ rather than $p(\cdot|\theta; c_i)$. The distributions' dependency on rather the c_i's or the x_{ij}'s is nevertheless still accounted for in the following theoretical developments.

The BIC statistic arises in the Bayesian approach for model selection. Bayesian choice is based on the posterior probability of a given model m: $p(m|\mathbf{y}) \propto p(m)p(\mathbf{y}|m)$. Assuming that the prior over models m is uniform, we need some way of approximating $p(\mathbf{y}|m)$. The Laplace approximation [14] of the integral

$$p(\mathbf{y}|m) = \int p(\mathbf{y}|\theta, m)p(\theta|m)d\theta,$$

where $p(\theta|m)$ denotes the prior for the parameters θ of each model m, gives

$$\log p(\mathbf{y}|m) \approx \log p(\mathbf{y}|\widehat{\theta}, m) - \frac{1}{2}\log\det\left(H_{\widehat{\theta}}\right). \tag{2.1}$$

Here $\widehat{\theta}$ denotes the maximum likelihood estimator (MLE) of the model parameters and $H_{\widehat{\theta}} = -\frac{\partial^2 \log p(\mathbf{y}|m)}{\partial\theta\partial\theta'}|_{\theta=\widehat{\theta}}$ is the negative Hessian matrix computed at the MLE and also referred to as the observed information matrix. We use $\approx$ to mean "is approximately equal to", corresponding to asymptotic equivalence as sample size goes to infinity. In fixed-effects models, under basic regularity assumptions, the Hessian H_θ behaves as N times a full-rank matrix $\mathcal{I}_\theta$, namely the information matrix of the model. Then,

$$\frac{1}{2}\log\det\left(H_{\widehat{\theta}}\right) \approx \frac{q}{2}\log N + \frac{1}{2}\log\det\left(\mathcal{I}_{\widehat{\theta}}\right),$$

where $q = \dim(\theta)$ is the number of parameters in the model, *i.e.* the number of elements of θ. The second term does not grow with N and by dropping it and substituting into (2.1) we get the classical BIC approximation

$$\log p(\mathbf{y}|m) \approx \log p(\mathbf{y}|\widehat{\theta}, m) - \frac{q}{2}\log N.$$

In mixed effects models, this is often not true, depending on whether $n \to \infty$. The order of accuracy of the Laplace approximation depends both on the number of subjects N and the number of observations per subject n; the second term $\log\det(\mathcal{I}_{\widehat{\theta}})$ cannot be evaluated anymore as a constant. To evaluate the respective contributions of N and n in (2.1), we investigate the orders of magnitude of the components of $H_{\widehat{\theta}}$ with respect to both N and n. It requires to expand the likelihood $p(\mathbf{y}|m)$ around the random effects rather than around the average effect over all subjects. To see this, we write $\log p(\mathbf{y}|m) = \sum_{i=1}^{N}\log p(y_i|\theta)$ and consider the marginal likelihood for subject i, $p(y_i|\theta)$, that is obtained by integrating the conditional distribution function of the data vector y_i with respect to ψ_i's distribution:

$$p(y_i|\theta) = \int p(y_i|\psi_i)p(\psi_i|\theta)d\psi_i. \tag{2.2}$$

Except for particular models, the integral in (2.2) does not have a tractable expression. Along the lines of [19, 20], the individual complete likelihood $p(y_i, \psi_i|\theta)$ is decomposed into two terms according to the covariance structure in the model. Equation (1.6) naturally divides the ψ_i's into two components: the individual parameters ψ_{ik} which are not random (for which $\Omega_{kk} = 0$), and the individual parameters that randomly vary among subjects, corresponding to a non-degenerate random effect η_{ik} with non negative variance, $k = 1, \ldots, d$. We denote by $\psi_{F,i}$ the components of ψ_i that are not random and $\psi_{R,i}$ the components of ψ_i that include a random component, leading to following notations:

$$\psi_i = \begin{pmatrix} \psi_{F,i} \\ \psi_{R,i} \end{pmatrix}, \; C_i = \begin{pmatrix} C_{F,i} & 0 \\ 0 & C_{R,i} \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_F \\ \beta_R \end{pmatrix}, \; \eta_i = \begin{pmatrix} \eta_{F,i} \\ \eta_{R,i} \end{pmatrix}, \tag{2.3}$$

in such a way equations (1.5) and (1.6) are still valid. Thus the population parameter θ is decomposed into (θ_R, θ_F), where $\theta_R = (\beta_R, \Omega_R)$ and $\theta_F = \beta_F$. Note that although the components of β and θ are now indexed by either "F" or "R", the whole population parameters are fixed parameters in the model. Indexes "F" and "R" only refer to the fixed and random components of ψ_i. Some precise illustration of such a decomposition is provided in Section 3 and summarized in Table 1.

Thus, (2.2) can be replaced by

$$\begin{aligned} p(y_i|\theta) &= \int p(y_i|\psi_{R,i}, \psi_{F,i}) p(\psi_{R,i}|\theta_R) d\psi_{R,i}, \\ &= \int p(y_i|\psi_{R,i}, \theta_F) p(\psi_{R,i}|\theta_R) d\psi_{R,i}. \end{aligned} \tag{2.4}$$

The negative Hessian matrix H_θ can be written as

$$H_\theta = -\sum_{i=1}^{N} \begin{pmatrix} \dfrac{\partial^2 \log p(y_i|\theta)}{\partial\theta_R \partial{\theta_R}'} & \dfrac{\partial^2 \log p(y_i|\theta)}{\partial\theta_R \partial{\theta_F}'} \\ \dfrac{\partial^2 \log p(y_i|\theta)}{\partial\theta_F \partial{\theta_R}'} & \dfrac{\partial^2 \log p(y_i|\theta)}{\partial\theta_F \partial{\theta_F}'} \end{pmatrix}. \tag{2.5}$$

Under suitable regularity conditions, by using (2.4) and Laplace approximations of the partial derivatives of the individual log-likelihoods, we obtain

$$\log\det\left(H_{\widehat{\theta}}\right) \approx \log\det(N\mathcal{I}_1) + \log\det(Nn\mathcal{I}_2),$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ represent full rank matrices of constant order of magnitude. By dropping constant terms and substituting into (2.1) we get the appropriate BIC approximation

$$\log p(\mathbf{y}|m) \approx \log p(\mathbf{y}|\hat{\theta}, m) - \frac{\dim(\theta_R)}{2}\log N - \frac{\dim(\theta_F)}{2}\log(Nn).$$

Details of the proof are given in the Appendix.

Both N and n need to be sufficiently large if this approximation is to work. It is worth noting that the leading term of the Laplace approximation of the individual log-likelihood derivatives is at most $\mathcal{O}(1)$, the reminder having order of $\mathcal{O}(1/n)$. Hence, the order of accuracy of the leading term of the Laplace approximation to the sample log-likelihood (2.1) is $\mathcal{O}(N/n)$, which is negligible provided n grows faster than N. Hence the BIC approximation is obtained under the assumption that the number of observations per subject grows at a faster rate than the number of subjects. If n grows at a rate slower than N, the Laplace leading term no longer converges to the log-likelihood, although the MLEs still attain consistency [30].

Under the assumptions given in the Appendix, we obtain the following result.

Main result. Assume that the data $\mathbf{y}$ are modeled by (1.1)–(1.2), and (1.5). The BIC procedure consists in selecting the model that minimizes

$$BIC_h = -2 \log p(\mathbf{y}|\hat{\theta}) + \dim(\theta_R) \log N + \dim(\theta_F) \log n_{\text{tot}}. \tag{2.6}$$

Remarks:

1. The new BIC criterion penalizes the size of θ_R with the logarithm of the number of subjects and the size of θ_F with the logarithm of the total number of observations. In a pure fixed-effects model, θ_R is empty and $\theta = \theta_F$. Then, the proposed criterion is the BIC with a $\log n_{\text{tot}}$ penalty:

   $$BIC_{n_{\text{tot}}} = -2 \log p(\mathbf{y}|\hat{\theta}) + \dim(\theta) \log n_{\text{tot}}.$$

   In the other extreme situation where all the individual parameters are random, all the population parameters are components of θ_R and the proposed criterion is the BIC with a $\log N$ penalty:

   $$BIC_N = -2 \log p(\mathbf{y}|\hat{\theta}) + \dim(\theta) \log N.$$

   Thus, the BIC_h proposed in equation (2.6) appears to be an hybrid BIC version that automatically adapts to the random-effects structure of a mixed model.
2. In the Akaike Information criteria (AIC) adjusted to mixed models [16, 29], the log-likelihood is penalized by a term that depends on the number of degrees of freedom ρ of the model. ρ is either the number of fixed parameters (marginal AIC) or the effective number of parameters (conditional AIC, [8, 15]). The choice between the two formulae depends on the focus of the model selection, *i.e.* prediction about the fixed effects or prediction about the random effects themselves. Our work on the BIC does not address the same issue. We give the optimal BIC to perform covariate selection, *i.e.* to find the non-zero elements of the fixed-effects β, whatever the random structure of the mixed model.
3. Deriving the BIC_h's value involves the computation of the population parameters' estimators $\hat{\theta}$ and the computation of the observed log-likelihood. These calculations can be intricate in nonlinear models but several algorithms are available [11] and are implemented in standard packages. The penalty term is straightforward to compute once the decomposition of $\theta = (\theta_R, \theta_F)$ is specified for the model under study. An illustrative example is given in the next section.

3. Simulation study

The objective of this section is to compare numerically the performances of the proposed "hybrid" BIC, called BIC_h, with those of the two most widely used BIC versions: $BIC_{n_{tot}}$ and BIC_N, which systematically penalize any model using $\log(n_{tot})$ and $\log(N)$ respectively.

We will consider a basic variable selection problem based on the following linear mixed-effects model:

$$y_{ij} = \psi_{i0} + \psi_{i1}\, t_{ij} + \psi_{i2}\, t_{ij}^2 + \varepsilon_{ij},\ i = 1, \dots, N,\ j = 1, \dots, n, \tag{3.1}$$

specified by the individual parameters $\psi_i = \big(\psi_{i0}, \psi_{i1}, \psi_{i2}\big)'$, $i = 1, \dots, N$, with a Gaussian measurement noise $\varepsilon_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Here we have $\psi_i = C_i\beta + \eta_i$, with

$$C_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & c_i & 0 \\ 0 & 0 & 1 & 0 & c_i \end{pmatrix} ; \quad \beta = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} .$$

Furthermore, $\eta_i \sim_{i.i.d.} \mathcal{N}(0, \Omega)$ where Ω is a diagonal matrix with diagonal elements $(\omega_0^2, \omega_1^2, \omega_2^2)$.

Here, covariate model selection reduces to selecting the non zero elements of (α_1, α_2). There are therefore 4 possible covariate models to compare using different information criteria:

$$\begin{aligned} M_1 &: \ \alpha_1 = 0,\ \alpha_2 = 0, \\ M_2 &: \ \alpha_1 \neq 0,\ \alpha_2 = 0, \\ M_3 &: \ \alpha_1 = 0,\ \alpha_2 \neq 0, \\ M_4 &: \ \alpha_1 \neq 0,\ \alpha_2 \neq 0. \end{aligned}$$

Unlike BIC_N and $BIC_{n_{tot}}$, BIC_h's penalty depends on both the number of random individual parameters and the number of covariates in the model. Table 1 displays the elements of θ_R and θ_F as well as the three different penalization terms used by the three different BIC, for the four covariate models $(M_k, 1 \leq k \leq 4)$ and the four following variance models:

$$\begin{aligned} O_1 &: \ \omega_1^2 = 0,\ \omega_2^2 = 0, \\ O_2 &: \ \omega_1^2 \neq 0,\ \omega_2^2 = 0, \\ O_3 &: \ \omega_1^2 = 0,\ \omega_2^2 \neq 0, \\ O_4 &: \ \omega_1^2 \neq 0,\ \omega_2^2 \neq 0. \end{aligned}$$

The aim of this numerical experiment is to investigate the behavior of the three different versions of BIC in different situations, *i.e.* using different covariate models, different variance models and different designs.

We have then simulated data under the $4 \times 4 \times 4 = 64$ possible situations by combining the covariate models $(M_k, 1 \leq k \leq 4)$, the variance models $(O_\ell, 1 \leq \ell \leq 4)$ and four different designs obtained with different numbers of subjects

TABLE 1. *Elements of θ_R and θ_F and penalization terms used by BIC_N, BIC_{ntot} and BIC_h for different variance and covariate models*

variance model	covariate model	β_R	Ω_R	θ_F	BIC_N	BIC_{ntot}	BIC_h
O_1	M_1	μ_0	ω_0	σ^2, μ_1, μ_2	$5 \log N$	$5 \log n_{\text{tot}}$	$2 \log N + 3 \log n_{\text{tot}}$
	M_2	μ_0	ω_0	$\sigma^2, \mu_1, \mu_2, \alpha_1$	$6 \log N$	$6 \log n_{\text{tot}}$	$2 \log N + 4 \log n_{\text{tot}}$
	M_4	μ_0	ω_0	$\sigma^2, \mu_1, \mu_2, \alpha_1, \alpha_2$	$7 \log N$	$7 \log n_{\text{tot}}$	$2 \log N + 5 \log n_{\text{tot}}$
O_2	M_1	μ_0, μ_1	ω_0, ω_1	σ^2, μ_2	$6 \log N$	$6 \log n_{\text{tot}}$	$4 \log N + 2 \log n_{\text{tot}}$
	M_2	μ_0, μ_1, α_1	ω_0, ω_1	σ^2, μ_2	$7 \log N$	$7 \log n_{\text{tot}}$	$5 \log N + 2 \log n_{\text{tot}}$
	M_3	μ_0, μ_1	ω_0, ω_1	$\sigma^2, \mu_2, \alpha_2$	$7 \log N$	$7 \log n_{\text{tot}}$	$4 \log N + 3 \log n_{\text{tot}}$
	M_4	μ_0, μ_1, α_1	ω_0, ω_1	$\sigma^2, \mu_2, \alpha_2$	$8 \log N$	$8 \log n_{\text{tot}}$	$5 \log N + 3 \log n_{\text{tot}}$
O_4	M_1	μ_0, μ_1, μ_2	$\omega_0, \omega_1, \omega_2$	σ^2	$7 \log N$	$7 \log n_{\text{tot}}$	$6 \log N + \log n_{\text{tot}}$
	M_2	$\mu_0, \mu_1, \mu_2, \alpha_1$	$\omega_0, \omega_1, \omega_2$	σ^2	$8 \log N$	$8 \log n_{\text{tot}}$	$7 \log N + \log n_{\text{tot}}$
	M_4	$\mu_0, \mu_1, \mu_2, \alpha_1, \alpha_2$	$\omega_0, \omega_1, \omega_2$	σ^2	$9 \log N$	$9 \log n_{\text{tot}}$	$8 \log N + \log n_{\text{tot}}$

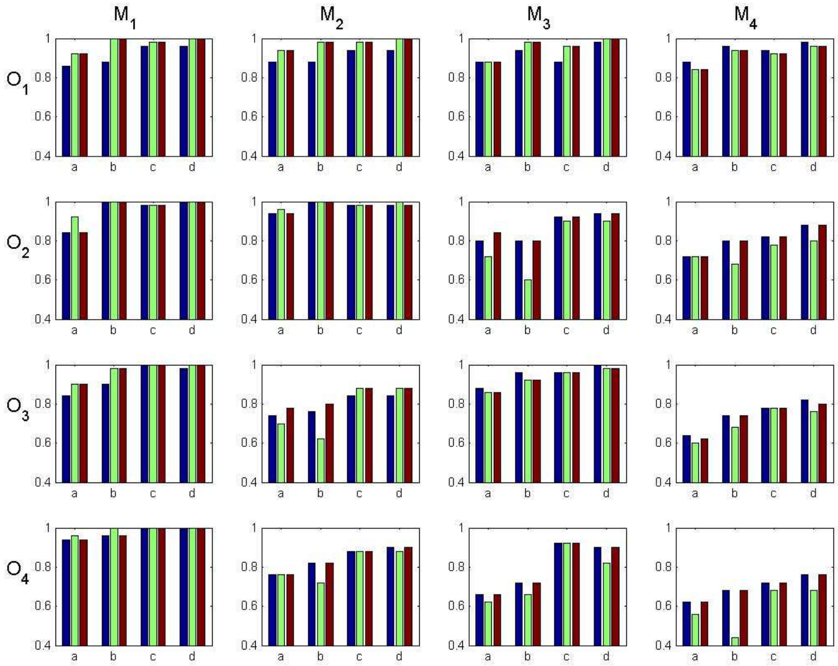

Fig 1. *Frequency of correct covariate selection for the three BIC versions: BIC_N (blue), $BIC_{n_{tot}}$ (green) and BIC_h (brown) under different covariate models $M_1(\alpha_1 = 0, \alpha_2 = 0)$, $M_2(\alpha_1 \neq 0, \alpha_2 = 0)$, $M_3(\alpha_1 = 0, \alpha_2 \neq 0)$, $M_4(\alpha_1 \neq 0, \alpha_2 \neq 0)$, different variance models $O_1(\omega_1^2 = 0, \omega_2^2 = 0)$, $O_2(\omega_1^2 \neq 0, \omega_2^2 = 0)$, $O_3(\omega_1^2 = 0, \omega_2^2 \neq 0)$, $O_4(\omega_1^2 \neq 0, \omega_2^2 \neq 0)$ and different designs $a(N = 20, n = 5)$, $b(N = 20, n = 100)$, $c(N = 100, n = 5)$, $d(N = 100, n = 100)$.*

N and different numbers of observations per subject n: $(N = 20, n = 5)$, $(N = 20, n = 100)$, $(N = 100, n = 5)$, $(N = 100, n = 100)$.

For each of these 64 models, 50 datasets were simulated as follows: the n observation time points $t_{i1}, \dots, t_{in}$ were equally spaced in $[0, 10]$ and the residual error variance was fixed to $\sigma^2 = 1$. In order to consider different situations, the covariates (c_i) and the variances $(\omega_m^2, 0 \leq m \leq 2)$ where randomly drawn for each replicate: $c_i \sim \mathcal{N}(0, 1)$, $\mu_0, \alpha_1, \alpha_2 \sim \mathcal{N}(0.01, 1)$, $\mu_1 \sim \mathcal{N}(0.005, 1)$, $\mu_2 \sim \mathcal{N}(0.0025, 1)$ and $\omega_m^2 \sim \mathcal{U}_{[0.01, 1.01]}$, $m = 0, 1, 2$.

For each simulated dataset, the EM algorithm was used for estimating the population parameters and the observed likelihood was computed as a Gaussian likelihood, according to (3.1). We could then derive the three versions of BIC using the penalization terms displayed in Table 1 and thus obtain three selected covariate models. The selected models were then compared to the original covariate model used for generating the data.

The results of the Monte-Carlo simulation study are displayed in Figure 1. The results obtained with the 64 models are displayed as follows: each column is associated to a covariate model, each row to a variance model. Then, any of the 16 subplots displays the results obtained with the four designs for a given covariate and variance model.

We can remark first that the performances of BIC_N and $BIC_{n_{tot}}$ significantly differ according to the covariance structure of the model. Figure 1 especially shows that $BIC_{n_{tot}}$ is more likely to select the right covariate model than BIC_N when there are no random components (model O_1). In this situation BIC_h behaves exactly as $BIC_{n_{tot}}$ since the penalizations are the same, up to a constant term (see Table 1). On the contrary, BIC_N better behaves than $BIC_{n_{tot}}$ in a model with a large number of random parameters (model O_4). Such result was expected according to Remark 1 in Section 2. In this extreme situation, we can remark that BIC_h now behaves exactly as BIC_N (see Table 1). In models with an intermediate covariance structure (models O_2 and O_3), the comparison between BIC_N and $BIC_{n_{tot}}$ seems to depend both on the design and the covariate model. In some situations, $BIC_{n_{tot}}$ underestimates the number of covariates and in other situations, BIC_N overestimates this number while BIC_h uses the most adequate penalization whatever the model and the design.

In summary, BIC_h is globally the best selection criterion in the present covariate selection problem, since it behaves in various situations at least as well as both standard criteria BIC_N and $BIC_{n_{tot}}$. On the other hand, the performances of BIC_N and $BIC_{n_{tot}}$ highly depend on the number of random parameters in the model. Thus, by automatically adapting its penalization according to the covariance structure of the model, BIC_h is a useful compromise between BIC_N and $BIC_{n_{tot}}$ for covariate selection in a population approach.

4. Application to the warfarin data

We will use a classical clinical pharmacology study [21] for illustrating the proposed method with a real data example. This data is available with the Monolix software and all the results presented in this section can easily be reproduced.

In this well known study, $N = 32$ healthy volunteers received a 1.5 mg/kg single oral dose of warfarin, an anticoagulant normally used in the prevention of thrombosis. The warfarin plasma concentration C was then measured at different times for these patients and a total number of $n_{\text{tot}} = 251$ measurements was obtained.

We consider a one compartment model for this data:

$$C(t, D, ka, V, Cl) = \frac{D\,ka}{V\,ka - Cl}\left(e^{-(Cl/V)\,t} - e^{-ka\,t}\right), \tag{4.1}$$

where D is the initial dose of drug, ka the absorption rate constant, V the volume of distribution and Cl the clearance of the drug. We then model the observations $(y_{ij}, 1 \leq i \leq n_i)$ of patient i using a proportional error model:

$$y_{ij} = C(x_{ij}, \psi_i) + C(x_{ij}, \psi_i)\varepsilon_{ij}, \tag{4.2}$$

where $x_{ij} = (t_{ij}, D_i)$ are the regression variables, $\psi_i = (ka_i, V_i, Cl_i)$ are the PK (pharmacokinetic) parameters for patient i and the residual errors ε_{ij} are

Table 2

BIC_N, BIC_{ntot} and BIC_h for the four covariate models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$

Model	dimension of θ_R	dimension of θ_F	$-2\,LL$	BIC_N	BIC_{ntot}	BIC_h
$\mathcal{M}_1$	5	2	928.1	952.4	966.8	956.5
$\mathcal{M}_2$	6	2	923.3	951.0	967.5	955.1
$\mathcal{M}_3$	5	3	923.0	950.7	967.2	956.9
$\mathcal{M}_4$	6	3	918.1	949.3	967.8	955.5

i.i.d. centered Gaussian random variables with variance σ^2. Due to positivity constraints, we assume that (ka_i, V_i, Cl_i) are log-normal random variables, *i.e.* $\log\psi_i$ is Gaussian. We also assume a diagonal covariance matrix Ω. Available covariate for patient i is its weight w_i. Here w_i has been centered by the empirical mean computed on the 32 patients.

We have compared several possible covariate models for the PK parameters but we only report here four representative models. For these four models, we use the same following model for V_i

$$\log(V_i) = \log(V_{\text{pop}}) + \beta_V w_i + \eta_{V,i},$$

where $\eta_{V,i} \underset{i.i.d.}{\sim} \mathcal{N}(0, \omega_V^2)$. We also assume that there is no random effect on ka_i. We then consider different covariate models for ka_i and Cl_i:

$$\begin{aligned}
\mathcal{M}_1 &: \quad \log(ka_i) = \log(ka_{\text{pop}}); & \log(Cl_i) &= \log(Cl_{\text{pop}}) + \eta_{Cl,i},\\
\mathcal{M}_2 &: \quad \log(ka_i) = \log(ka_{\text{pop}}); & \log(Cl_i) &= \log(Cl_{\text{pop}}) + \beta_{Cl} w_i + \eta_{Cl,i},\\
\mathcal{M}_3 &: \quad \log(ka_i) = \log(ka_{\text{pop}}) + \beta_{ka} w_i; & \log(Cl_i) &= \log(Cl_{\text{pop}}) + \eta_{Cl,i},\\
\mathcal{M}_4 &: \quad \log(ka_i) = \log(ka_{\text{pop}}) + \beta_{ka} w_i; & \log(Cl_i) &= \log(Cl_{\text{pop}}) + \beta_{Cl} w_i + \eta_{Cl,i},
\end{aligned}$$

where $\eta_{Cl,i} \underset{i.i.d.}{\sim} \mathcal{N}(0, \omega_{Cl}^2)$.

The results are reported in Table 2. The numbers of non-random and random components of θ in each model are indicated in columns 2–3 and the selection criteria BIC_N, BIC_{ntot} and BIC_h are computed in columns 5–7. We can remark that the four covariate models are not ranked in the same way according to the BIC penalty. BIC_N increases from the largest model to the smallest one and is minimum for $\mathcal{M}_4$. $BIC_{n_{tot}}$ ranks the models in the reverse order and selects the simplest one $\mathcal{M}_1$. As expected, BIC_h chooses the intermediary model $\mathcal{M}_2$ which includes in the model a weight effect on Cl.

We don't pretend here that $\mathcal{M}_2$ is the "best" model and that BIC_h is able to select it. This example only illustrates the fact that different models can be selected according to the penalization which is used. It is the simulation study presented in the previous section that illustrates the good statistical properties of BIC_h.

5. Discussion

The classical definition of BIC is inappropriate for mixed effects models where the information associated with a given model structure is affected by the number of random effects. This article derives the appropriate BIC penalty for mixed effects models depending on both the number of subjects N and the numbers of observations per subject $n_i, i = 1, \ldots, N$. The selection procedure is consistent if N and $\min(n_i)$ tend to infinity. Intuitively, the $\log N$ term comes from standard asymptotic theory while the $\log n_{\text{tot}}$ term comes from the Laplace approximation of the integrated likelihood. Although validated only if $\min(n_i)$ goes to infinity faster than N, the result of equation (2.6) is expected to work well when the number of measurements is moderate, as illustrated in the simulated example. Equation (2.6) is derived under usual regularity assumptions satisfied by generalized linear and nonlinear mixed-effects models. It can be extended to more general models involving a dependence structure within each subject, such as mixed-effects hidden Markov models [4, 5].

We have performed a simulation study for comparing the behavior of the proposed BIC with two standard BIC criteria that are implemented in different softwares used for the analysis of mixed effects models. Using a simple linear mixed effects model, we have found that the proposed BIC mainly behaves as the best of the two standard BIC, whatever the random structure of the model. This is predicted by the theory since the proposed criterion automatically adapts to the number of random factors that define the model structure. Additional simulations involving more complex models such as "inter-occasion" models often used in pharmacology are required to investigate the empirical properties of the proposed criterion, which will be implemented soon in the next version of Monolix.

The proposed BIC is designed only for covariate selection when the structure of the random effects of the model is given. The selection of significant random effects cannot be treated in the same way and deriving the BIC for joint selection of covariate and covariance components remains an open problem.

Appendix: Technical details

The key for deriving the expected BIC penalty in mixed-effects models is to get an appropriate approximation of $\log\det H_{\widehat{\theta}}$ when both the number of subjects N and the number of observations per subject n tend to infinity. The study of the Hessian matrix is based on the asymptotic evaluation of the partial second derivatives of the individual log-likelihood $\log p(y_i|\theta)$. The main steps of the proof are given below.

A.1. Decomposition of the log-likelihood

Let us first give the general form of the Laplace approximation of the partial second derivatives of subject i's observed log-likelihood [20]:

$$
\begin{aligned}
-\frac{\partial^2 \log p(y_i|\theta)}{\partial\theta\partial\theta'} = {} & \frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\theta\partial\theta'} \\
& - \frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\theta\partial\psi_i'}\left(\frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\psi_i\partial\psi_i'}\right)^{-1}\frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\theta'\partial\psi_i}\Big|_{\psi_i=\widehat{\psi}_i} + \mathcal{O}(n^{-1}),
\end{aligned}
\tag{A.1}
$$

where
$$l_i(y_i,\psi_i|\theta) = -\log p(y_i,\psi_i|\theta),$$
and
$$\widehat{\psi}_i = \underset{\psi_i}{\operatorname{argmax}}\, l_i(y_i,\psi_i|\theta).$$

As a result of the partition of θ between θ_R and θ_F (2.3), $l_i(y_i,\psi_i|\theta)$ naturally divides into
$$l_i(y_i,\psi_i|\theta) = l_{i1}(y_i|\psi_{R,i},\theta_F) + l_{i2}(\psi_{R,i}|\theta_R),$$
where
$$\begin{array}{rcl} l_{i1}(y_i|\psi_{R,i},\theta_F) & = & -\log p(y_i|\psi_{R,i},\theta_F), \\ l_{i2}(\psi_{R,i}|\theta_R) & = & -\log p(\psi_{R,i}|\theta_R), \end{array}$$
according to the Bayes formula.

A.2. Regularity assumptions

Some conditions on the model are required to study the order of magnitude of $H_{\widehat{\theta}}$'s components when $N, n \to +\infty$. Let ϑ denote an open subset of the parameter space Θ and let $\theta^\star$ denote the true population parameter value. We assume that for any given n:

(H1) For all $i = 1,\ldots,N$, and for all $\theta \in \vartheta$, $p(y_i|\theta)$ admits all first, second and third derivatives with respect to θ for almost all y_i.

(H2) (i) There exists $M > 0$ such that for all $i = 1,\ldots,N$, for all $\theta \in \vartheta$ and all $k,l = 1,\ldots,q$,
$$\mathbb{E}_{y_i|\theta^\star}\left[\frac{\partial \log p(y_i|\theta)}{\partial\theta_k}_{|\theta=\theta^\star}\right]^2 < M \text{ and } \mathbb{E}_{y_i|\theta^\star}\left[\frac{\partial^2 \log p(y_i|\theta)}{\partial\theta_k\partial\theta_l}_{|\theta=\theta^\star}\right]^2 < M.$$

(ii) Moreover, there exists a sequence of functions $\{G_1(y_1),\ldots,G_N(y_N)\}$ such that for all $\theta\in\vartheta$, for all $i = 1,\ldots,N$ and for all $k,l,h = 1,\ldots,q$,
$$\left|\frac{\partial^3 \log p(y_i|\theta)}{\partial\theta_k\partial\theta_l\partial\theta_h}\right| \le G_i(y_i) \text{ and } \mathbb{E}_{y_i|\theta^\star}\left[G_i^2(y_i)\right] \le M.$$

(H3) Let $V_\theta = (H_{\theta^\star}^{-\frac{1}{2}})H_\theta(H_{\theta^\star}^{-\frac{1}{2}})'$ for all $\theta\in\vartheta$. Matrix $H_{\theta^\star}$ is positive definite and invertible, and
$$\max_{\theta\in\vartheta}||V_\theta - I_q|| \underset{N\to+\infty}{\longrightarrow} 0,$$
where I_q stands for the $q\times q$ identity matrix.

(H4) $\liminf_{N\to+\infty}\lambda_N = \lambda > 0$ where λ_N is the smallest eigenvalue of matrix $\mathbb{E}(\frac{1}{N}H_{\theta^\star})$.

(H5) (i) $$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y_i|\theta^\star}\left\{\left[\psi_{R,i}{}'\left(\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\psi_{R,i}\partial\psi_{R,i}{}'}+\Omega_R^{-1}\right)^{-1}\psi_{R,i}\right]_{|\psi_i=\hat{\psi}_i}\right\} = o(n^{-1}),$$

(ii) $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y_i|\theta^\star}\left\{\left[\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\theta_F\partial{\psi_{R,i}}'}\left(\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\psi_{R,i}\partial{\psi_{R,i}}'}+\Omega_R^{-1}\right)^{-1}\psi_{R,i}\right]_{|\psi_i=\hat{\psi}_i}\right\}$
$= \mathcal{O}(1).$

Assumptions (H1) to (H4) are classical regularity conditions encountered in most studies relative to the asymptotic maximum likelihood theory. They were adapted to mixed-effects models by [20]. Condition (H5) is necessary to evaluate the respective order of the components of the Hessian matrix and could be seen as a regularity condition on the distributions $p(y_i|\psi_i)$. (H5) is a realistic assumption in the sense that it is verified in most mixed models, even in models including a dependance structure within each subject under classical regularity conditions on $p(y_i|\psi_i)$ (see [20] and [6] for illustrative examples).

A.3. Asymptotic evaluation of $H_{\widehat{\theta}}$

In the following, we consider the Hessian matrix normalized with the number of subjects, *i.e.* $\frac{1}{N}H_{\widehat{\theta}}$, instead of $H_{\widehat{\theta}}$ itself. Denote

$$\mathcal{G} = \mathbb{E}\left[\frac{1}{N}H_{\theta^\star}\right].$$

Due to the decomposition $\theta = (\theta_R, \theta_F)$, $\mathcal{G}$ can be written as a block matrix:

$$\mathcal{G} = -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y_i|\theta^\star}\begin{pmatrix}\frac{\partial^2\log p(y_i|\theta)}{\partial\theta_R\partial\theta'_R} & \frac{\partial^2\log p(y_i|\theta)}{\partial\theta_R\partial\theta'_F}\\ \frac{\partial^2\log p(y_i|\theta)}{\partial\theta_F\partial\theta'_R} & \frac{\partial^2\log p(y_i|\theta)}{\partial\theta_F\partial\theta'_F}\end{pmatrix} = \begin{pmatrix}\mathcal{G}_{RR} & \mathcal{G}_{RF}\\ \mathcal{G}_{FR} & \mathcal{G}_{FF}\end{pmatrix}. \quad \text{(A.2)}$$

Since $\log\det\mathcal{G}^{-1} = -\log\det\mathcal{G}$, we rather study the orders of magnitude of $\mathcal{G}^{-1}$'s block components. By inverting (A.2), we have:

$$\mathcal{G}^{-1} = \begin{pmatrix}(\mathcal{G}_{RR}-\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1}\mathcal{G}_{FR})^{-1} & -(\mathcal{G}_{RR}-\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1}\mathcal{G}_{FR})\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1}\\ (-(\mathcal{G}_{RR}-\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1}\mathcal{G}_{FR})\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1})' & (\mathcal{G}_{FF}-\mathcal{G}_{FR}\mathcal{G}_{RR}^{-1}\mathcal{G}_{RF})^{-1}\end{pmatrix}. \quad \text{(A.3)}$$

The asymptotic evaluation of $\mathcal{G}^{-1}$ first requires to study the orders of magnitude of $\mathcal{G}_{RR}$, $\mathcal{G}_{RF}$ and $\mathcal{G}_{FF}$ separately. This relies on the Laplace approximations of $\frac{\partial^2\log p(y_i|\theta)}{\partial\theta_R\partial\theta'_R}$, $\frac{\partial^2\log p(y_i|\theta)}{\partial\theta_R\partial\theta'_F}$ and $\frac{\partial^2\log p(y_i|\theta)}{\partial\theta_F\partial\theta'_F}$ respectively and assumption (H5).

For example, using (A.1) to evaluate $\mathcal{G}_{RF}$, we write

$$\begin{aligned}&-\frac{\partial^2\log p(y_i|\theta)}{\partial\theta_F\partial{\theta_R}'}\\ &\quad= \frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\theta_F\partial{\theta_R}'} - \frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\theta_F\partial{\psi_{R,i}}'}\left(\frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\psi_{R,i}\partial{\psi_{R,i}}'}\right)^{-1}\frac{\partial^2 l_i(y_i,\psi_i|\theta)}{\partial\psi_{R,i}\partial{\theta_R}'}|_{\psi_i=\widehat{\psi}_i}\\ &\quad+\mathcal{O}(n^{-1}),\end{aligned}$$

$$= -\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\theta_F\partial{\psi_{R,i}}'}\left(\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\psi_{R,i}\partial{\psi_{R,i}}'}+\Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i2}(\psi_{R,i}|\theta_R)}{\partial\psi_{R,i}\partial{\theta_R}'}|_{\psi_i=\widehat{\psi}_i}$$
$$+\,\mathcal{O}(n^{-1}). \tag{A.4}$$

Then,

$$-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y_i|\theta^\star}\left[\frac{\partial^2\log p(y_i|\theta)}{\partial\theta_F\partial{\theta_R}'}\right] \tag{A.5}$$
$$= -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y_i|\theta^\star}\left[\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\theta_F\partial{\psi_{R,i}}'}\left(\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\psi_{R,i}\partial{\psi_{R,i}}'}+\Omega_R^{-1}\right)^{-1}\right.$$
$$\left.\times\frac{\partial^2 l_{i2}(\psi_{R,i}|\theta_R)}{\partial\psi_{R,i}\partial{\theta_R}'}|_{\psi_i=\widehat{\psi}_i}\right]+\mathcal{O}(n^{-1}). \tag{A.6}$$

As we assume ψ_i is Gaussian (see (1.2)), $\frac{\partial^2 l_{i2}(\psi_{R,i}|\theta_R)}{\partial\psi_{R,i}\partial{\theta_R}'}$ can be expressed as a first-order polynomial function of ψ_i. Thus,

$$\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\theta_F\partial{\psi_{R,i}}'}\left(\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\psi_{R,i}\partial{\psi_{R,i}}'}+\Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i2}(\psi_{R,i}|\theta_R)}{\partial\psi_{R,i}\partial{\theta_R}'}$$

and

$$\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\theta_F\partial{\psi_{R,i}}'}\left(\frac{\partial^2 l_{i1}(y_i|\psi_{R,i},\theta_F)}{\partial\psi_{R,i}\partial{\psi_{R,i}}'}+\Omega_R^{-1}\right)^{-1}\psi_{R,i}$$

have similar behaviours when $N,n\to+\infty$. According to assumption (H5)-(ii), we deduce that $\mathcal{G}_{RF}$ is of the order of a constant when $N,n\to+\infty$.

The orders of $\mathcal{G}_{RR}$ and $\mathcal{G}_{FF}$ are obtained on the same pattern. More precisely, using (H5)-(i), we show that

$$\mathcal{G}_{RR} = -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y_i|\theta^\star}\left[\frac{\partial^2 l_{i2}(\psi_{R,i}|\theta_R)}{\partial\theta_R\partial{\theta_R}'}|_{\psi_i=\widehat{\psi}_i}\right]+o(n^{-1})$$

where $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y_i|\theta^\star}[\frac{\partial^2 l_{i2}(\psi_{R,i}|\theta_R)}{\partial\theta_R\partial{\theta_R}'}|_{\psi_i=\widehat{\psi}_i}]$ is positive definite when $N,n\to+\infty$, thus $\mathcal{G}_{RR}=\mathcal{O}(1)$. Similarly, we can show that $\mathcal{G}_{FF}=\mathcal{O}(n)$, since $\mathcal{G}_{FF}$ only involves the derivatives of l_{i1} of order $\mathcal{O}(n)$.

In a second step, the asymptotic behaviors of $\mathcal{G}_{RR}$, $\mathcal{G}_{RF}$ and $\mathcal{G}_{FF}$ are combined together. We get:

$$\begin{array}{rcl}(\mathcal{G}_{RR}-\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1}\mathcal{G}_{FR})^{-1} & \approx & \mathcal{I}_1,\\ (\mathcal{G}_{FF}-\mathcal{G}_{FR}\mathcal{G}_{RR}^{-1}\mathcal{G}_{RF})^{-1} & \approx & n^{-1}\mathcal{I}_2,\\ -(\mathcal{G}_{RR}-\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1}\mathcal{G}_{FR})\mathcal{G}_{RF}\mathcal{G}_{FF}^{-1} & \approx & 0,\end{array}$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ are respectively $\dim(\theta_R)\times\dim(\theta_R)$ and a $\dim(\theta_F)\times\dim(\theta_F)$ full rank matrices of constant order of magnitude.

From (A.3), we can write

$$\begin{aligned}\log\det\mathcal{G}^{-1} &\approx \log\left(\det\mathcal{I}_1\det\left(n^{-1}\mathcal{I}_2\right)\right),\\ &= \log\det\mathcal{I}_1 - \dim(\theta_F)\log n + \log\det\mathcal{I}_2.\end{aligned}$$

Finally, to get BIC_h's penalty, $\log\det H_{\widehat{\theta}}$ is approximated by $\log\det(N\mathcal{G})$:

$$\begin{aligned}\log\det\left(N\mathcal{G}\right) &= \log\left(N^{\dim(\mathcal{G})}\det\mathcal{G}\right),\\ &= \dim(\theta)\log N - \log\det\mathcal{G}^{-1},\\ &\approx \dim(\theta_F)\log(nN) + \dim(\theta_R)\log N - \log\det\mathcal{I}_1 - \log\det\mathcal{I}_2.\end{aligned}$$

Hence the BIC approximation given in equation (2.6).

Acknowledgement

The authors thank two referees for helpful comments that improved the manuscript.

The research leading to these results has received support from Innovative Medicines Initiative Joint. Undertaking under grant agreement n. 115156, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme and EFPIA companies in kind contribution. The DDMoRe project is also financially supported by contributions from Academic and SME partners.

The third author was partially supported by the FMJH and the labex LMH (Government Program: ANR-10-CAMP-0151-02 and ANR-11-LABX-0056-LMH).

References

[1] Bondell, H., Krishna, A., and Ghosh, S. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, **66** 1069–1077. MR2758494

[2] Comets, E., Lavenu, A., and Lavielle, M. (2011). *saemix: stochastic approximation expectation maximization SAEM algorithm*. R package version 0.96.

[3] Davidian, M. and Giltinan, D. (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, **8** 387–419.

[4] Delattre, M. and Lavielle, M. (2012). Maximum likelihood estimation in discrete mixed hidden Markov models using the SAEM algorithm. *Computational Statistics & Data Analysis*, **56(6)** 2073–2085. MR2892400

[5] Delattre, M. and Lavielle, M. (2013). Coupling the SAEM algorithm and the extended Kalman filter for maximum likelihood estimation in mixed-effects diffusion models. *Statistics and its interface*, **6(4)** 519–532.

[6] Delattre, M. (2012). *Inférence statistique dans les modèles mixtes à dynamique Markovienne*. PhD thesis, Université Paris-Sud.

[7] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of stochastic approximation version of the EM algorithm. *Annals of Statistics*, **27** 94–128. MR1701103

[8] Donohue, M., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, **98** 685–700. MR2836414

[9] Jiang, J. and Sunil Rao, J. (2003). Consistent procedures for mixed linear model selection. *Sankhyã: The Indian Journal of Statistics*, **65** 23–42. MR2016775

[10] Jiang, J., Sunil Rao, J., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, **36** 1669–1692. MR2435452

[11] Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed-effects models. *Computational Statistics and Data Analysis*, **49** 1020–1038. MR2143055

[12] Lai, R., Huang, H., and Lee, T. (2012). Fixed and random effects selection in nonparametric additive mixed models. *Electronic Journal of Statistics*, **6** 810–842. MR2988430

[13] Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, **38** 963–974.

[14] Lebarbier, E. and Mary-Huard, T. (2006). Une introduction au critère BIC: Fondements théoriques et interprétation. *Journal de la Société Française de Statistique*, **147(1)** 39–57. MR2500590

[15] Lian, H. (2012). A note on conditional Akaike information for Poisson regression with random effects. *Electronic Journal of Statistics*, **6** 1–9. MR2879670

[16] Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95** 773–778. MR2443190

[17] McCulloch, C. and Searle, S. (2008). *Generalized, Linear, and Mixed Models*. 2nd ed., Wiley, New-York. MR1884506

[18] `http://www.lixoft.com/monolix` (2013). *Monolix 4.2.2 User's Guide*.

[19] Nie, L. (2006). Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, **63** 123–143. MR2242536

[20] Nie, L. (2007). Convergence rate of the MLE in generalized linear and nonlinear mixed-effects models: Theory and applications. *Journal of Statistical Planning and Inference*, **137** 1787–1804. MR2323863

[21] O'Reilly, R. A. and Aggeler, P. M. (1968). Studies on coumarin anticoagulant drugs initiation of warfarin therapy without a loading dose. *Circulation*, **38(1)** 169–177.

[22] Pinheiro, J. and Bates, D. (2009). *Mixed-Effects Models in S and S-PLUS*. 1st ed., 2nd printing, Springer-Verlag, New York Inc.

[23] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2012). *nlme: linear and lonlinear mixed effects models*. R package version 3.1-105.

[24] Samson, A., Lavielle, M., and Mentré, F. (2007). The SAEM algorithm for group comparison tests in longitudinal analysis based on non-linear mixed-effects models. *Statistics in medecine*, **26** 4860–4875. MR2405484

[25] SAS (2008). *SAS/STAT 9.2 User's Guide*, chapter 61, 4337–4435.

[26] Schelldorfer, J., Meier, L., and Bühlmann, P. (2013). GLMM-Lasso: An algorithm for high-dimensional generalized linear mixed models using l_1-penalization. *Journal of Computational and Graphical Statistics*, DOI:10.1080/10618600.2013.773239.

[27] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6** 461–464. MR0468014

[28] SPSS (2002). *Linear mixed-effects modeling in SPSS. An introduction to the MIXED procedure*. Technical Report.

[29] Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92** 351–370. MR2201364

[30] Vonesh, E. (1996). A note on the use of Laplace's approximation for non-linear mixed-effects models. *Biometrika*, **83(2)** 447–452. MR1439795