BIC selection procedures in mixed effects models

Maud Delattre, Marc Lavielle, Marie-Anne Poursat

BIC selection procedures in mixed effects models

Maud Delattre, Marc Lavielle, Marie-Anne Poursat

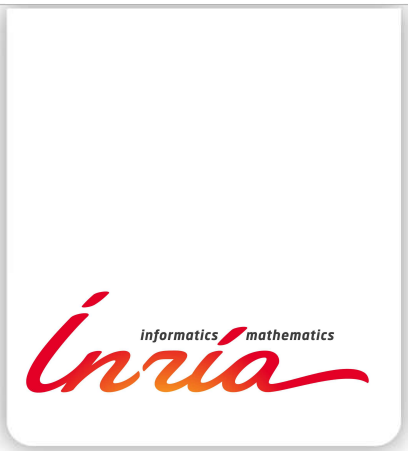
RESEARCH REPORT

N° 7948

Mai 2012

Project-Team Popix

ISSN 0249-6399 ISRN INRIA/RR--7948--FR+ENG



BIC selection procedures in mixed effects models

Maud Delattre*, Marc Lavielle*, Marie-Anne Poursat*

Project-Team Popix

Research Report n° 7948 — Mai 2012 — 17 pages

Abstract: We consider the problem of variable selection in general nonlinear mixed-effets models, including mixed-effects hidden Markov models. These models are used extensively in the study of repeated measurements and longitudinal analysis. We propose a Bayesian Information Criterion (BIC) that is appropriate for nonstandard situations where both the number of subjects N and the number of measurements per subject n tend to infinity. In this case, the consistency rates of the maximum likelihood estimators (MLE) of the parameters depend on the level of variability designed in the model. We show that the MLE of the population parameters related to subject-specific parameters are $\sqrt{N}$-consistent whereas the MLE of the parameters related to fixed parameters are $\sqrt{Nn}$-consistent. We derive a BIC criterion with a penalty based on two terms proportional to $\log N$ and $\log Nn$. Finite-sample properties of the proposed selection procedure are investigated by simulation studies.

Key-words: Consistency rate, Nonlinear mixed model, Hidden Markov mixed-effects model, Variable selection.

* Laboratoire de Mathématiques, Université Paris-Sud, France & Popix, Inria Saclay Ile-de-France

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

Procédures de sélection de variables de type BIC dans les modèles à effets mixtes

Résumé : Nous nous intéressons au problème de la sélection de variables dans des modèles non-linéaires mixtes généraux, incluant les modèles de Markov cachés à effets mixtes. Ces modèles sont très utilisés pour analyser des données répétées ou des données longitudinales. Nous proposons un critère BIC (Bayesian Information Criterion) adapté à la situation non-standard de double-asymptotique où le nombre de sujets N et le nombre d'observations par sujet n tendent vers l'infini. Dans cette situation, les vitesses de convergence des estimateurs du maximum de vraisemblance (EMV) des paramètres dépendent des niveaux de variabilité exprimés dans le modèle. Nous montrons que les EMV des paramètres de population liés aux paramètres spécifiques à chaque sujet sont $\sqrt{N}$-convergents tandis que les EMV des paramètres liés aux paramètres sans composante aléatoire sont $\sqrt{Nn}$-convergents. Nous en déduisons un critère BIC dont la pénalité est formée de deux termes en $\log N$ et $\log Nn$. Nous illustrons le comportement de la méthode de sélection de variables proposée par une étude de simulations.

Mots-clés : Modèle de Markov caché à effets mixtes, Modèle non-linéaire mixte, Sélection de variables, Vitesses de convergence.

1 Introduction

Nonlinear mixed-effects models are used in population studies where repeated measurements are observed from several independent subjects ([3]). Population studies occur in various fields such as pharmacokinetics or public health, for example to study disease evolution and to determine the effect of treatment or physiological covariates ([1], [8]). There is an extensive literature on parameter estimation in mixed models. However, the variable selection problem has been much less studied in these models. It is a standard practice to select the most relevant predictors using a Bayesian Information Criterion (BIC; [12]). The procedure consists in maximizing the observed log-likelihood penalized by the product of the dimension of the model and the logarithm of the sample size. Yet, the effective sample size is unclear in typical situations of mixed models. Therefore, the practice is to penalize the BIC either by the logarithm of the number of subjects or the logarithm of the total number of observations, without any guiding rule to choose between these two penalties. The purpose of this paper is to give a theoretical answer to the problem of choosing which penalty term is convenient in the practice to select the significant covariates.

To fix the notations, assume there are $i = 1, \ldots, N$ subjects and $j = 1, \ldots, n_i$ repeated observations nested within subject i. The ith observation consists in the vector of n_i observations $\mathrm{y}_i = (y_{i1}, \ldots, y_{in_i})$, where y_{ij}, $j = 1, \ldots, n_i$ denotes the jth measure observed at time point t_{ij}. There are two specifications in a mixed model. First, the probability distribution of the y_i's is assumed to belong to a common parametric model $p(\mathrm{y}_i|\phi_i)$ specified by a vector of individual parameters ϕ_i of length K. Second, subject effects are added into the parametric model by considering ϕ_i as a random vector. In this work, $\phi_i = (\phi_{i1}, \ldots, \phi_{iK})^T$ is defined as :

$$\phi_i = \beta X_i + \eta_i, \quad \eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega), \quad i = 1, \ldots, N. \tag{1}$$

β is a $K \times p$ matrix of fixed-effects, X_i is the $p \times 1$ vector of known covariates for subject i. The vector of random effects $\eta_i = (\eta_{i1}, \ldots, \eta_{iK})^T$ represents the between subjects variability that is not captured by the covariates. These are treated as random effects because the sampled subjects are thought to represent a population of subjects. They are assumed Gaussian and independent across subjects. The variance matrix Ω indicates the degree of heterogeneity of subjects. We denote by $\theta = (\beta, \Omega)$ the set of parameters of the global model that are to be estimated from the observations $\mathrm{y}_1, \ldots, \mathrm{y}_N$.

The idea of BIC is to penalize the log-likelihood by a term proportional to the logarithm of the sample size, which is N in the classic case. In mixed models, it can also be the total number of observations $N_{\mathrm{tot}} = \sum_{i=1}^{N} n_i$. For example, consider the simple linear model $\mathrm{y}_{ij} = \phi_i + \xi_{ij}$, $\phi_i = \beta X_i + \eta_i$, $i = 1, \ldots, N$, $j = 1 \ldots, n_i$, where ξ_{ij} is a Gaussian residual error independent of the random effect η_i. The effective sample size to estimate the fixed effects $(\beta_{k1}, \ldots, \beta_{kp})$ is N if $\Omega_{kk} > 0$ whereas it is N_{tot} if $\Omega_{kk} = 0$ i.e. ϕ_{ik} is not random anymore but fixed. This motivates the asymptotic study of BIC when both the number of subjects N and the numbers of measurements per subject n_i tend to infinity.

As the penalty term in BIC relies on the asymptotic approximation of the distribution of the maximum likelihood estimators (MLE) of the model parameters, we first consider the consistency rates of the MLEs in the double-asymptotic situation where $N \to \infty$ and $n_i \to \infty$, $i = 1, \ldots, N$. In Section 2, we extend the results of [10] to general nonlinear mixed models such as mixed hidden Markov models where the repeated observations nested within each subject are dependent. In the double-asymptotic framework, the consistency rates of the maximum likelihood estimators (MLE) of the parameters depend on the level of variability designed in the model. We show that the MLE of the population parameters defining the subject-specific parameters are $\sqrt{N}$-

consistent whereas the MLE of the parameters that are identical for all subjects are $\sqrt{N_{\text{tot}}}$-consistent. As a consequence, we obtain in Section 3 an appropriate BIC with a penalty based on two terms proportional to $\log N$ and $\log N_{\text{tot}}$. We illustrate the performance of this criterion with a simulation study. We conclude with a discussion in Section 4.

2 Asymptotic convergence of the MLE

For the sake of simplicity, we assume without loss of generality that the number of repeated observations n_i is the same across subjects : $n_i \equiv n$. The total number of observations is then $N_{\text{tot}} = \sum_i n_i = Nn$. We investigate the asymptotic behavior of the components of θ when $N \to \infty$ and $n \to \infty$.

Although $\sqrt{N}$-consistency of the MLE in standard parametric models are well-known results, some special features of the mixed-effects models may yield different convergence rates in the double-asymptotic framework.

We first consider two examples : the first one is a simple linear model that illustrates the special data structure of interest and the second one is a mixed hidden Markov model that motivated our study.

Example 1: Linear Mixed Model.

$$\begin{aligned} y_{ij} &= \phi_i + \xi_{ij}, \\ \phi_i &= \mu + \eta_i, \end{aligned} \tag{2}$$

where $\eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \omega^2)$, $\xi_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, N$, $j = 1, \ldots, n$.
Here, the MLE of μ is

$$\begin{aligned} \widehat{\mu}_{\text{MLE}} &= \frac{1}{Nn} \sum_{i=1}^{N} \sum_{j=1}^{n} y_{ij}, \\ &= \mu + \frac{1}{N} \sum_{i=1}^{N} \eta_i + \frac{1}{Nn} \sum_{i=1}^{N} \sum_{j=1}^{n} \xi_{ij}, \end{aligned}$$

and

$$\text{Var}(\widehat{\mu}_{\text{MLE}}) = \frac{\omega^2}{N} + \frac{\sigma^2}{Nn}.$$

Thus, if $\omega > 0$, ϕ_i is an individual random parameter and the MLE of μ is $\sqrt{N}$-consistent, while if $\omega = 0$, $\phi_i = \mu$ is not a random parameter anymore and the MLE of μ is $\sqrt{Nn}$-consistent.

Example 2: Mixed Hidden Markov Model.

We suppose that y_i is the realization of a S-state mixed hidden Markov model (MHMM) with Poisson emissions. For each $i = 1, \ldots, N$, the expression of $p(\text{y}_i|\phi_i)$ is given by

$$p(\text{y}_i|\phi_i) = p(y_{i1}) \prod_{j=2}^{n} p(y_{ij}|y_{i1}, \ldots, y_{i,j-1}, \phi_i),$$

and involves a sum over all possible sequences of n states in $\{1, \ldots, S\}$:

$$p(\text{y}_i|\phi_i) = \sum_{z_{i1}, \ldots, z_{in} \in \{1, \ldots, S\}^n} p(z_{i1}|\phi_i) \prod_{j=2}^{n} p(z_{ij}|z_{i,j-1}, \phi_i) \prod_{j=1}^{n} p(y_{ij}|z_{ij}, \phi_i). \tag{3}$$

We assume the initial state distribution to be uniform and known for all $i = 1, \ldots, N$. For $S = 2$, the specification of each individual hidden Markov model is reduced to the two Poisson parameters: λ_{1i} in state 1 and λ_{2i} in state 2, and the transition probabilities from state 1 to state 1 ($p_{11,i}$) and from state 2 to state 1 ($p_{21,i}$). This can be written as

$$\begin{aligned} \text{logit } p_{11,i} = \phi_{1i} \quad &, \quad \text{logit } p_{21,i} = \phi_{2i}, \\ \log \lambda_{1i} = \phi_{3i} \quad &, \quad \log \lambda_{2i} = \phi_{4i}, \end{aligned}$$

for all $i = 1, \ldots, N$, with $\phi_i = (\phi_{i1}, \ldots, \phi_{i4})^T$. See [8] for more details about this model and its use for epilepsy seizure count modelling.

2.1 The observed likelihood

In a mixed-effects model where the unobserved ϕ_i's are random variables, the probability distribution function (pdf) for subject i, $p(\mathbf{y}_i; \theta)$, $i = 1, \ldots, N$, is obtained by integrating the conditional distribution function of the data vector $\mathbf{y}_i$ with respect to ϕ_i's distribution:

$$p(\mathbf{y}_i; \theta) = \int p(\mathbf{y}_i, \phi_i; \theta) d\phi_i, \tag{4}$$

$$= \int p(\mathbf{y}_i | \phi_i) p(\phi_i; \theta) d\phi_i. \tag{5}$$

Except for linear mixed-effects models, the integral over the ϕ_i's do not have any explicit expression :

By independence of the N subjects, the joint pdf $p(\mathbf{y}; \theta)$ is the product of the N individual pdf's $p(\mathbf{y}_i; \theta)$ and the MLE of θ is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} \sum_{i=1}^{N} \log p(\mathbf{y}_i; \theta).$$

To establish the theoretical properties of the MLE $\hat{\theta}$ of parameter θ when both the number of subjects N and the number of observations per subject n tend to infinity, we investigate the convergence rate of the components of the observed Fisher information matrix. Due to the special structure of the mixed effects models, the structure of the Fisher information matrix naturally divides the individual parameters ϕ_i into two groups : the individual parameters ϕ_{ik} which are not random (for which $\Omega_{kk} = 0$), and the individual parameters that randomly vary among subjects, corresponding to a random effect η_{ik} with non negative variance. We denote by $\phi_{F,i}$ the components of ϕ_i that are not random and $\phi_{R,i}$ the components of ϕ_i that include a random component.

If the set of fixed parameters $\phi_{F,i}$ is not empty, then the decomposition of the pdf proposed in (5) does not hold any more. Indeed, to $\phi_i = (\phi_{F,i}, \phi_{R,i})$ corresponds a natural partition of the model parameters $\theta = (\theta_F, \theta_R)$, and (5) should be replaced by

$$p(\mathbf{y}_i; \theta) = \int p(\mathbf{y}_i | \phi_{R,i}, \phi_{F,i}) p(\phi_{R,i}; \theta_R) d\phi_{R,i} \tag{6}$$

$$= \int p(\mathbf{y}_i | \phi_{R,i}, \theta_F) p(\phi_{R,i}; \theta_R) d\phi_{R,i}. \tag{7}$$

As the likelihood can not be expressed in a closed form, the study of the theoretical properties of the MLE in a mixed-effects model is not straightforward. There exists few results about the

properties of the MLE in mixed-effects models. Some well known references to this topic are papers from Nie [11, 9, 10]. In these articles, the author suggests very specific tools for studying the MLE in mixed-effects models. In particular, the demonstrations proposed in [9, 10] are based on the individual complete likelihoods $p(\mathbf{y}_i, \phi_i; \theta)$, which generally have closed form expression, rather than the marginal likelihood of the observations. Thus, according to (7), the study of the asymptotic properties of the MLE is based on the following decomposition of the individual complete log-likelihoods:

$$l_i(\mathbf{y}_i, \phi_i; \theta) = l_{i1}(\mathbf{y}_i, \phi_{R,i}, \theta_F) + l_{i2}(\phi_{R,i}, \theta_R),$$

where $l_i(\mathbf{y}_i, \phi_i; \theta) = -\log p(\mathbf{y}_i, \phi_i; \theta)$, $l_{i1}(\mathbf{y}_i, \phi_{R,i}, \theta_F) = -\log p(\mathbf{y}_i|\phi_{R,i}, \theta_F)$, $l_{i2}(\phi_{R,i}, \theta_R) = -\log p(\phi_{R,i}; \theta_R)$.

Remark: The individual complete log-likelihood is decomposed into two terms which don't have the same order, since $l_{i1}(\mathbf{y}_i, \phi_{R,i}, \theta_F)$ is the conditional log-likelihood of the vector of observations $\mathbf{y}_i$ of size n, while $l_{i2}(\phi_{R,i}, \theta_R)$ is the log-likelihood of the random vector $\phi_{R,i}$ which size is fixed as given by the model.

2.2 Assumptions

For deriving the asymptotic distribution of the MLE of parameter θ, we need four classical assumptions (H1)-(H4) ensuring the regularity of the model and both continuity and invertibility of the Fisher Information Matrix in a neighborhood of the true parameter value $\theta^\star$. These assumptions can be found in [10] and are given in the Appendix.

When the number of observations per subject tends to infinity, an additional assumption (H5) is required, which ensures the regularity of $p(\mathbf{y}_i|\phi_i)$ for $i = 1, \ldots, N$:

(H5)

(i) $$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left\{\left[{\phi_{R,i}}^T\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i, \phi_{R,i}, \theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T} + \Omega_R^{-1}\right)^{-1}\phi_{R,i}\right]_{|\phi_i=\hat{\phi}_i}\right\} = o(n),$$

(ii) $$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left\{\left[\frac{\partial^2 l_{i1}(\mathbf{y}_i, \phi_{R,i}, \theta_F)}{\partial\theta_F\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i, \phi_{R,i}, \theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T} + \Omega_R^{-1}\right)^{-1}\phi_{R,i}\right]_{|\phi_i=\hat{\phi}_i}\right\} = \mathcal{O}(n),$$

where Ω_R is the variance matrix of $\phi_{R,i}$, and $\hat{\phi}_i = \text{argmax}_{\phi_i}\ l_i(\mathbf{y}_i, \phi_i; \theta)$.

Condition (H5) is specific to the individual models and is necessary to evaluate the respective order of the components of the inverse Fisher information matrix. In Example 1, theses conditions are easy to check. Nie [10] showed that they can be verified as well in mixed-effects regression models where the repeated observations nested within each subject are independent. In more general models such as Example 2, where the repeated observations are driven by a Markov structure dependence, we showed that assumption (H5) can be relaxed to classical ergodicity conditions:

(H5') For all $j = 1, \ldots, n$, as n tends to infinity,

$$\left|\frac{\partial^2}{\partial\phi_i\partial\phi_i^T}\left(\log p(y_{ij}|y_{i,j-1}, \ldots, y_{i1}, \phi_i) - \log p(y_{ij}|y_{i,j-1}, \ldots, \phi_i)\right)_{|_{\phi_i=\phi_i^\star}}\right|$$

converges in probability to 0. $\phi_i^\star$ is the true individual parameter for subject i.

2.3 A Central Limit Theorem

We now give the asymptotic distribution of the MLE in mixed-effects models when $N, n \to +\infty$.

Lemma 1 (Asymptotic independence of $\hat{\theta}_R$ and $\hat{\theta}_F$)**.** *Under (H1)-(H5), the MLEs for parameters θ_R and θ_F are asymptotically independent as N and n simultaneously tend to infinity, with $\frac{n}{N} \to +\infty$.*

Theorem 1 (Asymptotic distribution of the MLE when $N, n \to +\infty$)**.** *Assume (H1)-(H5). As N and n simultaneously tend to infinity, with $\frac{n}{N} \to +\infty$,*

$$\begin{aligned}\sqrt{N}(\hat{\theta}_R - \theta_R^\star) &\xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_1(\theta^\star)),\\ \sqrt{Nn}(\hat{\theta}_F - \theta_F^\star) &\xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_2(\theta^\star)),\end{aligned}$$

where

$$\Gamma_1^{-1}(\theta^\star) = \lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\phi_{R,i};\theta_R)}{\partial\theta_R \partial{\theta_R}^T}_{|_{\phi_i=\hat{\phi}_i}}\right],$$

and

$$\begin{aligned}\Gamma_2^{-1}(\theta^\star) = \lim_{N,n\to+\infty} \frac{1}{Nn}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\Bigg[&\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F \partial{\theta_F}^T}\\ &-\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F \partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i} \partial{\phi_{R,i}}^T} + \Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i} \partial{\theta_F}^T}_{|_{\phi_i=\hat{\phi}_i}}\Bigg],\end{aligned}$$

where $\theta^\star$ is the true parameter value.

Proof of Lemma 1 and Theorem 1. The starting point for getting the asymptotic distribution of the MLE in any parametric model is a Taylor series expansion of the log-likelihood around the true parameter value $\theta^\star$. From this, assuming that the model fulfills classical regularity conditions results in the convergence in distribution of the normalized score function, the convergence in probability of the normalized Hessian function of the log-likelihood, and the negligibility of the rest term of the Taylor series expansion. Invertibility of the Fisher information matrix is also required. When the number of subjects tends to infinity and the number of observations per subject is finite, provided that assumptions (H1)-(H4) are fulfilled, we have:

$$\sqrt{N}\begin{pmatrix}\hat{\theta}_R - \theta_R^\star\\ \hat{\theta}_F - \theta_F^\star\end{pmatrix} \underset{\mathrm{N}\to+\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta^\star)\right), \tag{8}$$

where

$$\mathcal{I}(\theta^\star) = -\lim_{N\to+\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\begin{pmatrix}\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R \partial{\theta_R}'} & \dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R \partial{\theta_F}'}\\ \dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F \partial{\theta_R}'} & \dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F \partial{\theta_F}'}\end{pmatrix} = \begin{pmatrix}\mathcal{I}_{11} & \mathcal{I}_{12}\\ \mathcal{I}_{21} & \mathcal{I}_{22}\end{pmatrix}.$$

See [2] for details. The distribution stated in (8) is degenerated as n tends to infinity, requiring adjustment of the rates of convergence of $\hat{\theta}_R$ and $\hat{\theta}_F$ as $n \to +\infty$. We show that some block components of $\mathcal{I}^{-1}(\theta^\star)$ tend to 0 as N, n tend to infinity, and adequately normalize the different

components of the Fisher information matrix by evaluating the orders of the block components of $\mathcal{I}(\theta^\star)$ as suggested in [10]. Each component of the Fisher information matrix involves second derivatives of the marginal individual likelihoods with respect to θ_R and θ_F, but due to the expression of the likelihood as an integral over the ϕ_i's, evaluation of these derivatives is not straightforward and requires Laplace approximation of each second partial derivative of $p(\mathbf{y}_i;\theta)$:

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta\partial\theta^T} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta\partial\phi_i^T}\left(\frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\phi_i\partial\phi_i^T}\right)^{-1}\frac{\partial^2 l_i(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta^T\partial\phi_i} + \mathcal{O}(n)\right]|_{\phi_i=\hat{\phi}_i}.$$

Using assumption (H5) and Laplace approximation of the partial derivatives of the individual log-likelihoods, we get the results of Lemma 1 and Theorem 1. More details are given in the Appendix.

□

3 Model selection

We will use the convergence results obtained in the previous section to derive the BIC penalty in a population approach context for selecting the covariate model.

3.1 BIC for mixed-effects models

Deriving the BIC penalty requires to take a finite collection of mixed-effects models, $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m)$, and to consider these models and their respective parameters as random variables. In model $\mathcal{M}_k$, the data is supposed to have distribution $p(\cdot;\theta_k)$ characterized by a set of population parameter θ_k.

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ denote the observations sample, composed of n independent sequences of longitudinal data: $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})$, where the n observations for a given subject are not necessarily independent. The aim is to select model $\mathcal{M}_k$ among collection $\mathcal{M}$ presenting highest a posteriori distribution $p(\mathcal{M}_k|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_k)p(\mathcal{M}_k)$. Assuming uniform a priori distribution for the models of $\mathcal{M}$, the problem amounts to maximizing $p(\mathbf{y}|\cdot)$ among the models of the collection. For any given $k = 1, \ldots, m$,

$$p(\mathbf{y}|\mathcal{M}_k) = \int_{\Theta_k} p(\mathbf{y}|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k,$$

is evaluated using a Laplace approximation, according to [7]:

$$\begin{aligned}\log p(\mathbf{y}|\mathcal{M}_k) \quad = \quad & \left(\sum_{i=1}^{N}\log p(\mathbf{y}_i|\theta_k^\star)\right) + \log p(\theta_k^\star|\mathcal{M}_k) + \frac{\dim(\theta_k)}{2}\log(2\pi) \\ & -\frac{\dim(\theta_k)}{2}\log(N) - \frac{1}{2}\log\det\left(-\frac{1}{N}\frac{\partial^2 \log p(\mathbf{y}|\mathcal{M}_k)}{\partial\theta\partial\theta^T}|_{\theta_k=\theta_k^\star}\right) \\ & +\mathcal{O}(N).\end{aligned}$$

The Hessian matrix $-\frac{1}{N}\frac{\partial^2 \log p(\mathbf{y}|\mathcal{M}_k)}{\partial\theta\partial\theta^T}|_{\theta_k=\theta_k^\star}$ is approximated with $\mathcal{I}(\hat{\theta}_k)$, but when both N and n tend to infinity, $\log\det\mathcal{I}(\hat{\theta}_k)$ can not be evaluated as a constant as in [7].

We assume here a linear model for the individual parameters as described in (1). We therefore decompose the vector of random effect η_i into $(\eta_{F,i}, \eta_{R,i})$ where $\eta_{F,i} = 0$ is associated to the fixed individual parameters and $\eta_{R,i}$ to the random ones:

$$\phi_{F,i} = \beta_F X_i \tag{9}$$
$$\phi_{R,i} = \beta_R X_i + \eta_{R,i}, \quad \eta_{R,i} \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega_R), \quad i = 1, \ldots, N, \tag{10}$$

where Ω_R is a positive-definite variance covariance matrix. We consider here that all the models of collection $\mathcal{M}$ have the same covariance structure in the sense that the covariance matrix Ω of the individual parameters ϕ_i's has the same structure in the m models. In other words, the decomposition $\phi_i = (\phi_{F,i}, \phi_{R,i})$ is the same for all the models We focus on the covariate selection problem, *i.e.* the selection of the non zero elements of β_F and β_R, but not on the selection of the non zero elements of Ω. Here, θ is decomposed into (θ_R, θ_F), where $\theta_R = (\beta_R, \Omega_R)$ and θ_F is β_F, eventually augmented with model-specific fixed parameters such as the error variance σ^2 as in Example 1.

Using asymptotic independence of $\hat{\theta}_R$ and $\hat{\theta}_F$ (Lemma 1), and the convergence rates of $\hat{\theta}_R$ and $\hat{\theta}_F$ as $N, n \to +\infty$ (Theorem 1), we get

$$-\log\det\mathcal{I}(\hat{\theta}_k) \approx -\log\det(\Gamma_1(\hat{\theta}_k)) - \log\det(n\Gamma_2(\hat{\theta}_k)),$$

where, using the same notations as in previous section, $\Gamma_1(\hat{\theta}_k)$ and $n\Gamma_2(\hat{\theta}_k)$ represent the diagonal block components of $\mathcal{I}(\hat{\theta}_k)^{-1}$. According to Theorem 1, $\Gamma_1(\hat{\theta}_k)$ and $\Gamma_2(\hat{\theta}_k)$ are evaluated as constants when both $N, n \to +\infty$. Thus, correctly normalizing each term of Laplace approximation expressed above and neglecting all terms remaining bounded as $N, n \to +\infty$, we get

$$\log p(\mathbf{y}|\mathcal{M}_k) \approx \log p(\mathbf{y}|\hat{\theta}_k) - \frac{\dim(\theta_{R,k})}{2}\log N - \frac{\dim(\theta_{F,k})}{2}\log Nn$$

as $N, n \to +\infty$ and $\dfrac{n}{N} \to +\infty$.

Theorem 2. *The BIC procedure consists in selecting the model that minimizes*

$$BIC(\mathcal{M}_k) = -2\log p(\mathbf{y}|\hat{\theta}_k, \mathcal{M}_k) + \dim(\theta_{R,k})\log N + \dim(\theta_{F,k})\log(Nn),$$

among the models $\mathcal{M}_k$ of collection $\mathcal{M}$.

Remark: Replace Nn by $N_{tot} = \sum_{i=1}^{N} n_i$ in BIC expression when the number of observations per subject differs from one subject to an other.

3.2 Simulation study

We now investigate the contributions of the hybrid BIC penalty in variable selection problems occuring in mixed-effects models. More precisely, we confront the new criteria $BIC_{N,Nn}$ to BIC penalized with either the logarithm of the number of subjects (BIC_N) or the logarithm of the total number of observations (BIC_{Nn}), in a variable selection problem via a short simulation study.

3.2.1 Model

We use the following linear mixed-effects model for the simulations, which generalizes the simple model of Example 1. For $i = 1, \ldots, N$ and $j = 1, \ldots, n$,

$$\begin{aligned} y_{ij} &= \mu + \phi_{i1} t_{ij} + \phi_{i2} t_{ij}^2 + \xi_{ij} \quad , \quad \xi_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \\ \phi_{i1} &= a_0 + a_1 x_i + \eta_i \quad , \quad \eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \omega^2), \\ \phi_{i2} &= b_0 + b_1 x_i, \end{aligned} \tag{11}$$

where x_i is a covariate for subject i and t_{ij} the observation time of y_{ij}. Here, $\phi_{R,i} = \phi_{i1}$, $\phi_{F,i} = \phi_{i2}$, $X_i = (1, x_i)$ and $\Omega_R = \omega^2$.

In this simple model, the distribution of the observations is explicit: $y_{ij} \sim \mathcal{N}(m_{ij}, \gamma_{ij}^2)$, where $m_{ij} = \mu + (a_0 + a_1 x_i) t_{ij} + (b_0 + b_1 x_i) t_{ij}^2$ and $\gamma_{ij} = t_{ij}^2 \omega^2 + \sigma^2$. Thus, the likelihood can be easily derived for any vector of population parameter θ. Four sub-models, M_{11}, M_{10}, M_{01}, M_{00}, can be derived from model (11), depending on wether the coefficients a_1 and b_1 are null or not. The expression of the BIC hybrid penalty for each sub-model is given in Table 1. Note that in this specific model, the variable selection problem arises at two levels of the model: on the random parameter ϕ_{i1} on the one hand, on the non random parameter ϕ_{i2} on the other hand.

Model	Description	θ_R	θ_F	pen($BIC_{N,Nn}$)
M_{11}	($a_1 \neq 0$, $b_1 \neq 0$)	a_0, a_1, ω^2	μ, b_0, b_1, σ^2	$3 \log N + 4 \log Nn$
M_{10}	($a_1 \neq 0$, $b_1 = 0$)	a_0, a_1, ω^2	μ, b_0, σ^2	$3 \log N + 3 \log Nn$
M_{01}	($a_1 = 0$, $b_1 \neq 0$)	a_0, ω^2	μ, b_0, b_1, σ^2	$2 \log N + 4 \log Nn$
M_{00}	($a_1 = 0$, $b_1 = 0$)	a_0, ω^2	μ, b_0, σ^2	$2 \log N + 3 \log Nn$

Table 1: Definition of submodels M_{11}, M_{10}, M_{01} and M_{00}, and derivation of the corresponding penalties for hybrid BIC.

3.2.2 Design for the simulations

We will perform a Monte-Carlo experiment in order to evaluate the performance of our covariate model selection criteria.

Each of the $K = 500$ replicates of the Monte-Carlo experiment consists in simulating a new dataset $\mathcal{Y}_{N,n}^{k,k'}$ with N subjects and n observations per subject from model $M_{kk'}$, $k, k' \in \{0, 1\}^2$. The parameters θ_R and θ_F of models M_{11}, M_{10}, M_{01}, M_{00}, are estimated from the simulated observations using the EM algorithm. Using the estimated values $\hat{\theta}_R$ and $\hat{\theta}_F$, the observed likelihood is computed in each model of the collection, and the corresponding values for BIC_N, BIC_{Nn} and $BIC_{N,Nn}$ are derived. Then, minimization of each criteria leads to the selection of one of the four possible models. We can then obtain the number of times each model has been chosen according to each criteria.

Different designs (number N of subjects and number n of observations per subject) were investigated using $N = (20, 50, 100)$ and $n = (20, 100, 200, 500, 1000)$. The variance of the residual error was set to $\sigma^2 = 1$. The other model components of the model are randomly drawn at each replicate of the Monte Carlo experiment: $C_i \sim \mathcal{N}(0, 1)$, $\mu, a_1, b_1 \sim \mathcal{N}(0.01, 1)$, $b_0 \sim \mathcal{N}(0.005, 1)$, $b_1 \sim \mathcal{N}(0.0025, 1)$ and $\omega^2 \sim \mathcal{U}_{[0.01, 1.01]}$. The t_{ij}'s are equally spaced in $[0, 10]$.

3.2.3 Results

The results are displayed in Figure 1. The Monte-Carlo study mainly shows that the new criteria has very similar properties than BIC penalized with the number of subjects, which is currently the most frequently used model selection criteria in a population approach in practice. Nevertheless, when the variable selection problem arises at random effect level $\phi_{R,i}$ of the model (models M_{10} and M_{11}), we notice most important differences between the three criteria. When there is a covariate effect on $\phi_{R,i}$ only (model M_{10}), the hybrid BIC is even slightly better than BIC_N - in the sense that $BIC_{N,Nn}$ most frequently selects the correct model than BIC_N - especially when the number of subjects is small ($N = 20$). We also show very bad performances of BIC_{Nn} in this context. When a covariate effect arises at least at random effect level of the model, it over-penalizes parameters θ_R of the model, resulting in most frequent selection of the model without covariate on $\phi_{R,i}$. On the other hand, BIC_{Nn} gives slightly better results when no covariate arises at random level of the model.

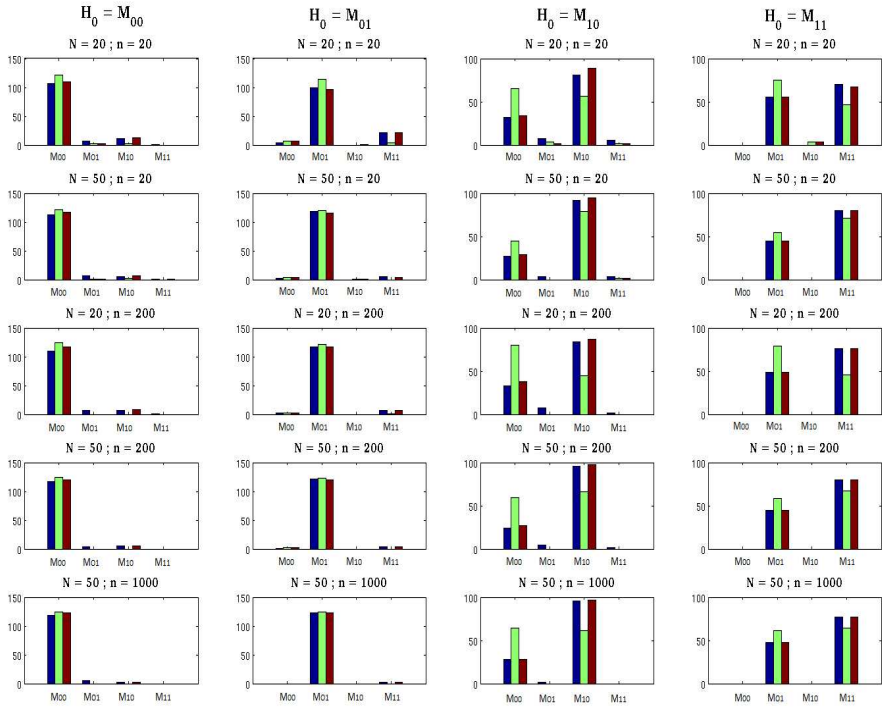


Figure 1: Results given by the three versions of BIC on 500 simulated datasets for the different sub-models of (11) (columns) and different values for (N, n) (raws). FOr each design, the histograms represent the numbers of times each model (from left to right: M_{00}, M_{01}, M_{10}, M_{11}) is selected by each version of BIC: BIC_N (blue), BIC_{Nn} (green), $BIC_{N,Nn}$ (red).

We have replicated the Monte-Carlo experiment with different values for the parameters and the t_{ij}'s. As expected, the performances of the three criteria are sensitive to the parameter scales. When greater weight is granted to the random effect term $\phi_{R,i}$ of the model, it is intuitively easier to detect a covariate in $\phi_{R,i}$ than in $\phi_{F,i}$, and vice versa. For example, when the observation time interval is $[0, 1]$, the hybrid BIC show much better properties than above when the problem

is to detect a covariate in $\phi_{R,i}$ and not in $\phi_{F,i}$ (model M_{10}), but mainly selects model M_{10} even when the true model is M_{11}. Similar results are recorded with the other criteria.

4 Discussion

The contribution of the present paper is twofold. First, we have extended the asymptotic results established by Nie [10] for nonlinear mixed-effects models to more general models involving a dependence structure within each subject, such as mixed-effects hidden Markov models. We consider the double-asymptotic framework where both the number of subjects N and the number of repeated measures n tend to infinity. We show that, provided some classical ergodicity conditions on the individual Markov chains, the rates of convergence of the MLEs depend on the levels of variability in the model : the parameters θ_R involved in the random components of the individual parameters ϕ_i, $i = 1, \ldots, N$ are $\sqrt{N}$-consistent while the parameters θ_F related to the non-random individual parameters are $\sqrt{Nn}$-consistent. We have then derived from these results a specific version of BIC for covariate selection in mixed-effects models. 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.

We have performed a simulation study for comparing the behavior of the proposed BIC with 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 hybrid BIC mainly behaves as the classical BIC penalized with the logarithm of the number of subjects but outperforms in most cases the BIC penalized with the logarithm of the total number of observations. In this illustrative example, we have also noticed some slight superiority of the new criteria in some situations, even with moderate N and n.

Additional simulations involving more complex models such as mixed-effects hidden Markov models are required to investigate the empirical properties of the proposed criterion. Furthermore, the proposed BIC is designed only for covariate selection when the structure of the random effects of the model is given. Deriving a new criteria for selecting the covariance structure of the random effects as well remains an open problem. BIC-type procedures were studied by [4] to select both fixed and random effects but they are limited to linear mixed models. Few papers addressed the problem of variable selection in nonlinear mixed models. [5] proposed a "fence" method to select predictors in generalized linear models, [6] implemented a fully Bayesian selection approach in mixed logistic models, [13] proposed a double-penalized likelihood approach for simultaneous model selection and estimation in semiparametric mixed models. In the future we hope to develop an appropriate criterion that would be able to perform both covariate and covariance selection in a population approach.

5 Appendix

5.1 Assumption of Theorem 1

We formulate conditions providing asymptotic distribution of the MLE when $N, n \to +\infty$.

Let ϑ denote an open subset of Θ. We assume that for any given n:

(H1) for all $i = 1, \ldots, N$, and for all $\theta \in \vartheta$, $p(\mathbf{y}_i; \theta)$ admits all first, second and third derivatives with respect to θ for almost all $\mathbf{y}_i$.

(H2) (i) There exists $M > 0$ such that for all $i = 1, \ldots, N$, for all $\theta \in \vartheta$ and all $k = 1, \ldots, r$,

$$\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial \log p(\mathbf{y}_i;\theta)}{\partial \theta_k}|_{\theta=\theta^\star}\right]^2 < M \text{ and } \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial \theta_k \partial \theta_l}|_{\theta=\theta^\star}\right]^2 < M.$$

(ii) Moreover, there exists a sequence of functions $\{G_1(\mathbf{y}_1), \ldots, G_N(\mathbf{y}_N)\}$ such that for all $\theta \in \vartheta$, for all $i = 1, \ldots, N$ and for all $k, l, h = 1, \ldots, r$,

$$\left|\frac{\partial^3 \log p(\mathbf{y}_i;\theta)}{\partial \theta_k \partial \theta_l \partial \theta_h}\right| \leq G_i(\mathbf{y}_i) \text{ and } \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[G_i^2(\mathbf{y}_i)\right] \leq M.$$

(H3) Let $V_N(\theta) = H_N(\theta^\star)^{-\frac{1}{2}} H_N(\theta) H_N(\theta^\star)^{-\frac{T}{2}}$, and $H_N(\theta) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta\partial\theta^T}$ for all $\theta \in \vartheta$.

Matrix $H_N(\theta^\star)$ is positive definite and invertible, and

$$\max_{\theta\in\vartheta} ||V_N(\theta) - I_r|| \underset{N\to+\infty}{\longrightarrow} 0,$$

where I_r stands for the $r \times r$ identity matrix.

(H4) $\liminf_{N\to+\infty} \lambda_N = \lambda > 0$ where λ_N is the smallest eigenvalue of matrix

$$F_N = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta\partial\theta^T}\right].$$

5.2 Evaluation of the Fisher Information Matrix when $N, n \to +\infty$

Here, we focus on the evaluation of the components of the inverse Fisher Information Matrix as $N, n \to +\infty$ and $\dfrac{n}{N} \to +\infty$. The upper and lower diagonal blocks of $\mathcal{I}^{-1}$ are respectively given by $\left(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1}$ and $\left(\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\right)^{-1}$ and the anti-diagonal blocs are given by $-(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21})\mathcal{I}_{12}\mathcal{I}_{22}^{-1}$ and its transpose.

First of all, key issue is to get the orders of magnitude of $\mathcal{I}_{11}, \mathcal{I}_{21}, \mathcal{I}_{12}$ and $\mathcal{I}_{22}$ as $n \to +\infty$. Let us detail the approach on block $\mathcal{I}_{21}$. Getting an evaluation of $\mathcal{I}_{21}$ requires Laplace approximation of $-\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F \partial\theta_R{}^T}$ for all $i = 1, \ldots, N$. It is given by:

$$-\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F \partial\theta_R{}^T} = \left[\frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta_F \partial\theta_R{}^T} - \frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\theta_F \partial\phi_{R,i}{}^T}\left(\frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\phi_{R,i}\partial\phi_{R,i}{}^T}\right)^{-1}\frac{\partial^2 \log p(\mathbf{y}_i,\phi_i;\theta)}{\partial\phi_{R,i}\partial\theta_R{}^T}\right]|_{\phi_i=\hat{\phi}_i} + \mathcal{O}(n),$$

and can be simplified into

$$-\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R \partial\theta_F{}^T} = -\left[\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F \partial\phi_{R,i}{}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial\phi_{R,i}{}^T} + \Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\phi_{R,i}\partial\theta_R{}^T}\right]|_{\phi_i=\hat{\phi}_i} + \mathcal{O}(n).$$

Linear Gaussian model for the ϕ_i's makes $\dfrac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\phi_{R,i}\partial{\theta_R}^T}$ expressed as a first-order polynomial function of ϕ_i. Using results of Lemma 1, and correctly normalizing each component in Laplace approximation of $\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R\partial{\theta_F}^T}$ with n, we establish that as $N, n \to +\infty$ such that $n/N \to +\infty$:

$$-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_R}^T}\right] = \mathcal{O}(n)$$

In other words, $\mathcal{I}_{21}$ converges to a constant matrix as $n \to +\infty$. As $\mathcal{I}_{21} = \mathcal{I}_{12}^T$, we also evaluate $\mathcal{I}_{12}$ as a constant as $n \to +\infty$. Similarly, we use result of Lemma 1 as well and Laplace approximation of $-\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_R\partial{\theta_R}^T}$, and we get

$$\lim_{N,n\to+\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\theta_R\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}+\Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\phi_{R,i}\partial{\theta_R}^T}\right]_{|_{\phi_i=\hat{\phi}_i}} = 0,$$

thus convergence of $\mathcal{I}_{11}$ to a constant positive definite matrix as $n \to +\infty$. Indeed, $\mathcal{I}_{11} = \lim_{N\to+\infty}\dfrac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\dfrac{\partial^2 l_{i2}(\phi_{R,i},\theta_R)}{\partial\theta_R\partial{\theta_R}^T}\right]$. Recall that the ϕ_i's are independent and identically distributed random variables. Thus, $\mathcal{I}_{11} = \dfrac{\partial^2 l_{12}(\phi_1^R,\theta_R)}{\partial\theta_R\partial{\theta_R}^T}$, which is a positive and definite matrix, and does not depend on the number of observations per subject.

Finally, using Laplace approximation of $\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}$, we get equivalence between $-\dfrac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\dfrac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}\right]$ and

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F\partial{\theta_F}^T} - \frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\theta_F\partial{\phi_{R,i}}^T}\left(\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial{\phi_{R,i}}^T}+\Omega_R^{-1}\right)^{-1}\frac{\partial^2 l_{i1}(\mathbf{y}_i,\phi_{R,i},\theta_F)}{\partial\phi_{R,i}\partial\theta_F}\right]_{|_{\phi_i=\hat{\phi}_i}}$$

as n tends to infinity.

Using (H5) and previous Laplace approximation of the partial second derivatives of the individual log-likelihoods, we are able to evaluate each block component of matrix $\mathcal{I}$. Providing that $\hat{\phi}_i$ is a consistent estimate of $\phi_i^\star$ for any value of θ, we show that $\left(\mathcal{I}_{11}-\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1}$ converges in probability to 0 as N tends to infinity for any value of n, giving result of Lemma 1. Similarly, we show that $\left(\mathcal{I}_{11}-\mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\right)^{-1}$ is of the order of a constant, and that the distribution of $\sqrt{N}(\hat{\theta}_F-\theta_F^\star)$ is degenerated since $\left(\mathcal{I}_{22}-\mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\right)^{-1}$ converges in probability to 0 as N and n simultaneously tend to infinity. We take again the classical frame for demonstration of convergence in distribution of the maximum likelihood estimate, focusing on θ_F only, and normalizing the log-likelihood with Nn. To get $\sqrt{Nn}$ convergence in distribution of $\hat{\theta}_F$, we

only need positive definiteness of $\lim_{N,n\to+\infty} -\frac{1}{Nn}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{y}_i|\theta^\star}\left[\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}\right]$. This results from assumption (H5) by evaluating N terms $\frac{\partial^2 \log p(\mathbf{y}_i;\theta)}{\partial\theta_F\partial{\theta_F}^T}$ with its Laplace approximation.

References

[1] Marc Lavielle Adeline Samson and France Mentré. The saem algorithm for group comparison tests in longitudinal analysis based on non-linear mixed-effects models. *Statistics in medecine*, 26:4860–4875, 2007.

[2] Ralph A. Bradley and John J. Gart. The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214, 1962.

[3] Marie Davidian and David M. Giltinan. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8:387419, 2003.

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

[5] Zhonghua Gu Jiming Jiang, J. Sunil Rao and Thuan Nguyen. Fence methods for mixed model selection. *The Annals of Statistics*, 36:1669–1692, 2008.

[6] Satkartar K. Kinney and David B. Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63:690–698, 2007.

[7] Emilie Lebarbier and Tristan Mary-Huard. Une introduction au critère bic: Fondements théoriques et interprétation. *Journal de la Société française de statistique*, 147(1):39–57, 2006.

[8] Delattre M. Inference in mixed hidden markov models and applications to medical studies. *Journal de la Société française de statistique*, 151(1):90–105, 2010.

[9] Lei Nie. Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, 63:123 – 143, 2006.

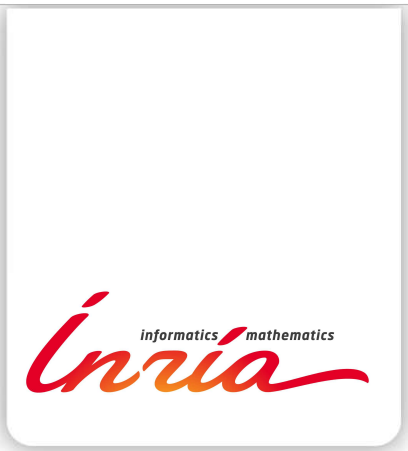
[10] Lei Nie. Convergence rate of the mle in generalized linear and nonlinear mixed-effects models: Theory and applications. *Journal of Statistical Planning and Inference*, 137:1787–1804, 2007.

[11] Lei Nie and Min Yang. Strong consitency of mle in nonlinear mixed-effects models with large cluster size. *Sankhya, The Indian Journal of Statistics*, 67:736–763, 2005.

[12] Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[13] Daowen Zhang Xiao Ni and Hao Helen Zhang. Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, 66:79–88, 2010.

Contents



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399