Estimating Fisher Information Matrix in Latent Variable Models based on the Score Function

Maud Delattre, Estelle Kuhn

Estimating Fisher Information Matrix in Latent Variable Models based on the Score Function

Delattre Maud[1], Estelle Kuhn[2]

[1] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, France.
[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France.

Abstract

The Fisher information matrix (FIM) is a key quantity in statistics as it is required for example for evaluating asymptotic precisions of parameter estimates, for computing test statistics or asymptotic distributions in statistical testing, for evaluating post model selection inference results or optimality criteria in experimental designs. However its exact computation is often not trivial. In particular in many latent variable models, it is intricated due to the presence of unobserved variables. Therefore the observed FIM is usually considered in this context to estimate the FIM. Several methods have been proposed to approximate the observed FIM when it can not be evaluated analytically. Among the most frequently used approaches are Monte-Carlo methods or iterative algorithms derived from the missing information principle. All these methods require to compute second derivatives of the complete data log-likelihood which leads to some disadvantages from a computational point of view. In this paper, we present a new approach to estimate the FIM in latent variable model. The advantage of our method is that only the first derivatives of the log-likelihood is needed, contrary to other approaches based on the observed FIM. Indeed we consider the empirical estimate of the covariance matrix of the score. We prove that this estimate of the Fisher information matrix is unbiased, consistent and asymptotically Gaussian. Moreover we highlight that none of both estimates is better than the other in terms of asymptotic covariance matrix. When the proposed estimate can not be directly analytically evaluated, we present a stochastic approximation estimation algorithm to compute it. This algorithm provides this estimate of the FIM as a by-product of the parameter estimates. We emphasize that the proposed algorithm only requires to compute the first derivatives of the complete data log-likelihood with respect to the parameters. We prove that the estimation algorithm is consistent and asymptotically Gaussian when the number of iterations goes to infinity. We evaluate the finite sample size properties of the proposed estimate and of the observed FIM through simulation studies in linear mixed effects models and mixture models. We also investigate the convergence properties of the estimation algorithm in non linear mixed effects models. We compare the performances of the proposed algorithm to those of other existing methods.

1 Introduction

The Fisher information matrix (FIM) is a key quantity in statistics as it is required for example for evaluating asymptotic precisions of parameter estimates, for computing Wald test statistics or asymptotic distributions in statistical testing (van der Vaart (2000)). It also appears in post model selection inference or in optimality criteria for experimental designs or as a particular Riemannian metric. However its exact computation is often not trivial. This is in particular the case in many latent variables models, also called incomplete data models, due to the presence of the unobserved variables. Though these models are increasingly used in many fields of application. They especially allow a better consideration of the different variability sources and when appropriate, a more precise characterization of the known mechanisms at the origin of the data generation. Let us quote some examples in pharmacology (Delattre et al. (2012)), in ecophysiology (Technow et al. (2015)), in genomic (Picard et al. (2007)) or in ecology (Gloaguen et al. (2014)). When the FIM can not be exactly computed, people focus on an estimate of the FIM and consider usually the observed FIM. When it can not be directly computed, several methods have been proposed to approximate it. Among the most frequently used approaches are Monte-Carlo methods or iterative algorithms derived from the missing information principle (Orchard and Woodbury (1972)). Indeed according to this principle, the observed Fisher information matrix can be expressed as the difference between two matrices corresponding to the complete information and the missing information due to the unobserved variables. It enables the development of alternative methods to compute the observed FIM: the Louis's method (Louis (1982)), combined with a Monte Carlo method or a stochastic approximation algorithm by Delyon et al. (1999), the Oakes method (Oakes (1999)) or the supplemented Expectation Maximization algorithm (Meng and Rubin (1991)). However as the observed FIM involves the second derivatives of the observed log-likelihood, all these methods require to compute second derivatives of the complete data log-likelihood which leads to some disadvantages from a computational point of view. More recently, Meng and Spall (2017) proposed an accelerated algorithm based on numerical first order derivatives of the conditional expectation of the log-likelihood.

In this paper, we present a new approach to evaluate the FIM in latent variables model. The advantage of our method is that only the first derivatives of the complete log-likelihood is needed. Indeed we consider the empirical estimate of the covariance matrix of the score. When the proposed estimate can not be directly analytically evaluated, we propose a stochastic approximation estimation algorithm to compute it, which provides this estimate of the FIM as a by-product of model parameter estimates.

The paper is organized as follows. In Section 2, we detail both moment estimates of the Fisher information matrix and establish their asymptotic properties. In Section 3, we give practical tools for the computation of the proposed estimate of the Fisher information matrix in incomplete data models. In particular, we introduce a new stochastic approximation procedure based on the first derivatives of the complete log-likelihood only and state its asymptotic properties. In Section 4, we illustrate the finite sample

size properties of both estimators and the convergence properties of the computation algorithm through simulations. The paper ends with conclusion and discussion.

2 Moment estimates of the Fisher information matrix

Let us consider a random variable Y and denote by g the density of Y. Assume that the log-likelihood function $\log g$ is parametric depending on some parameter θ taking values in Θ, differentiable on Θ and that $\|\partial_\theta \log g(y;\theta)(\partial_\theta \log g(y;\theta))^t\|$ is integrable. Then, by definition, the Fisher information matrix is given for all $\theta \in \Theta$ by:

$$I(\theta) = E_\theta \left[\partial_\theta \log g(Y;\theta)(\partial_\theta \log g(Y;\theta))^t\right]. \tag{1}$$

Moreover if the log-likelihood function $\log g$ is twice differentiable on Θ, the following relation also holds for all $\theta \in \Theta$:

$$I(\theta) = -E_\theta \left[\partial^2_\theta \log g(Y;\theta)\right]. \tag{2}$$

When none of these expressions can be analytically evaluated, people are interested in computing an estimate of the Fisher information matrix.

2.1 Definitions of the estimators

Considering the two expressions given in equations (1) and (2), we can derive two moment estimators for the Fisher information matrix based on a n-sample $(y_1, \ldots, y_n)$ of observations. Indeed both estimators, denoted by $I_{n,sco}(\theta)$ and $I_{n,obs}(\theta)$, are defined as empirical estimates of the Fisher information matrix based on the expressions involving respectively the score function and the Hessian as follows:

$$\begin{aligned} I_{n,sco}(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \partial_\theta \log g(y_i;\theta)(\partial_\theta \log g(y_i;\theta))^t \\ I_{n,obs}(\theta) &= -\frac{1}{n}\sum_{i=1}^{n} \partial^2_\theta \log g(y_i;\theta) \end{aligned}$$

Note that the estimator $I_{n,obs}(\theta)$ is usually called the observed Fisher information matrix. We emphasize that the estimator $I_{n,sco}(\theta)$ is defined as soon as the Fisher information matrix is whereas the estimator $I_{n,obs}(\theta)$ requires additional regularity assumptions to be defined. Moreover evaluating the estimator $I_{n,sco}(\theta)$ requires only to calculate the first derivatives of the log-likelihood whereas evaluating the estimator $I_{n,obs}(\theta)$ requires also to calculate the second ones.

2.2 Properties of the estimators

Both estimators are moment estimates and therefore unbiased. We now establish the asymptotic properties of both estimators.

Proposition 2.1 *Assume that $Y_1, \ldots, Y_n$ are independent identically distributed random variables from some parametric probability density function g depending on some parameter θ in an open subset Θ of $\mathbb{R}^p$. Assume also that $\log g$ is differentiable in θ on Θ and that for all $\theta \in \Theta$, $\partial_\theta \log g(y;\theta)(\partial_\theta \log g(y;\theta))^t$ is integrable. Then, for all $\theta \in \Theta$, the estimator $I_{n,sco}(\theta)$ is defined, consistent for $I(\theta)$ and asymptotically normal.*

Moreover, assuming additionaly that $\log g$ is twice differentiable in θ on Θ, the estimator $I_{n,obs}(\theta)$ is defined, consistent for $I(\theta)$ and asymptotically normal.

Proof The results follow by applying the law of large numbers and the central limit theorem. □

Remark 2.2 *Regarding the variance, none of both estimators is better than the other one. This can be highlighted through the following examples. First consider a Gaussian sample with unknown expectation and fixed variance. Then, the variance of the estimator $I_{n,obs}(\theta)$ is zero whereas the variance of the estimator $I_{n,sco}(\theta)$ is positive. Second consider a centered Gaussian sample with unknown variance. Then, the variance of $I_{n,sco}(\theta)$ is smaller than the variance of $I_{n,obs}(\theta)$. Therefore, none of both estimators is more suitable than the other in general.*

Since the above result does not apply if the variables $Y_1, \ldots, Y_n$ are not identically distributed, for example if they depend on some individual covariates which is often the case, we state the following result under the assumption of independent non identically distributed random variables.

Proposition 2.3 *Assume that $Y_1, \ldots, Y_n$ are independent non identically distributed random variables each having a parametric probability density function g_i depending on some common parameter θ in an open subset Θ of $\mathbb{R}^p$. Assume also that for all i the function $\log g_i$ is differentiable in θ on Θ and that for all $\theta \in \Theta$, $\partial_\theta \log g_i(y;\theta)(\partial_\theta \log g_i(y;\theta))^t$ is integrable. Moreover assume that for all θ in Θ, $lim \frac{1}{n} \sum_{i=1}^n E_\theta(\partial_\theta \log g_i(y;\theta)(\partial_\theta \log g_i(y;\theta))^t)$ exists and denotes it by $\nu(\theta)$. Then, for all $\theta \in \Theta$, the estimator $I_{n,sco}(\theta)$ is defined, converges almost surely toward $\nu(\theta)$ and is asymptotically normal. Moreover, assuming additionaly that $\log g_i$ is twice differentiable in θ on Θ, the estimator $I_{n,obs}(\theta)$ is defined, converges almost surely toward $\nu(\theta)$ and is asymptotically normal.*

Proof We prove the consistency by applying the law of large numbers for non identically distributed variables. We establish the normality result by using characteristic functions. By recentering the terms $E_\theta(\partial_\theta \log g_i(y;\theta)(\partial_\theta \log g_i(y;\theta))^t)$, we can assume that $\nu(\theta)$ equals zero. Let us denote by ϕ_Z the characteristic function for a random

variable Z. We have for all real t in a neighborhood of zero that:

$$\begin{aligned}\|\phi_{I_{n,sco}(\theta)/\sqrt{n}}(t)-1\| &= \|\prod_{i=1}^{n}(\phi_{\partial_\theta \log g_i(y;\theta)(\partial_\theta \log g_i(y;\theta))^t}(t/n)-1)\| \\ &\leq \prod_{i=1}^{n}\|\phi_{\partial_\theta \log g_i(y;\theta)(\partial_\theta \log g_i(y;\theta))^t}(t/n)-1\|\end{aligned}$$

Computing a limited expansion in t around zero, we get the result.

Noting that for all $1 \leq i \leq n$, $E_\theta(\partial_\theta \log g_i(y;\theta)(\partial_\theta \log g_i(y;\theta))^t) = -E_\theta(\partial^2_\theta \log g_i(y;\theta))$, we get the corresponding results for the estimator $I_{n,obs}(\theta)$. □

Remark 2.4 *The additional assumptions required when considering non identically distributed random variables are in the same spirit as those usually used in the literature. Let us quote for example Nie (2006), Silvapulle and Sen (2011), Baey et al. (2019).*

3 Computing the estimator $I_{n,sco}(\theta)$ in latent variable model

Let us consider independent random variables $Y_1, \ldots, Y_n$. Assume in the sequel that there exist independent random variables $Z_1, \ldots, Z_n$ such that for each $1 \leq i \leq n$, the random vector (Y_i, Z_i) admits a parametric probability density function denoted by f parametrized by $\theta \in \Theta$. We present in this section dedicated tools to compute the estimator $I_{n,sco}(\theta)$ in latent variable model when it can not be evaluated analytically.

3.1 Analytical expressions in latent variable models

In latent variable models, the estimator $I_{n,sco}(\theta)$ can be expressed using the conditional expectation as stated in the following proposition.

Proposition 3.1 *Assume that for all $\theta \in \Theta$ the function $\log g(\cdot;\theta)$ is integrable, that for all y the function $\log g(y;\cdot)$ is differentiable on Θ and that there exists an integrable function h_1 such that for all $\theta \in \Theta$, $\| \partial_\theta \log g(y;\theta)\| \leq h_1(y)$. Then for all $\theta \in \Theta$ and all $n \in \mathbb{N}^*$:*

$$I_{n,sco}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{Z_i|Y_i;\theta}(\partial_\theta \log f(Y_i, Z_i;\theta))\mathrm{E}_{Z_i|Y_i;\theta}(\partial_\theta \log f(Y_i, Z_i;\theta))^t$$

where $\mathrm{E}_{Z|Y;\theta}$ denotes the expectation under the law of Z conditionally to Y.

Proof Applying the Leibniz integral rule, we get that for all $\theta \in \Theta$:

$$\partial_\theta \log g(Y;\theta) = \mathrm{E}_{Z|Y;\theta}(\partial_\theta \log f(Y, Z;\theta))$$

This equality allows to express explicitely the first derivatives of the logarithm of the marginal density of Y as the expectation of the first derivatives of the logarithm of the

complete likelihood with respect to the conditional distribution of the latent variables. This statement is indeed in the same spirit as the well-known Louis formulae for the observed Fisher information matrix estimate. □

Remark 3.2 *In some specific cases the conditional expectations involved in the previous proposition admit exact analytical expressions for example in mixture models which are developed in Section 4 in some simulation studies.*

3.2 Computing $I_{n,sco}(\theta)$ using stochastic approximation algorithm

When exact computation of the estimator $I_{n,sco}(\theta)$ is not possible, we propose to evaluate its value by using a stochastic algorithm which provides the estimate $I_{n,sco}(\theta)$ as a by-product of the parameter estimates of θ.

3.2.1 Description of the algorithm in curved exponential family model

We consider an extension of the stochastic approximation Expectation Maximization algorithm proposed by Delyon et al. (1999) which allows to compute the maximum likelihood estimate in general latent variables model. We assume that the complete log-likelihood belongs to the curved exponential family for stating the theoretical results. However our algorithm can be easily extended to general latent variables models (see Section 3.2.3). As our estimate involves individual conditional expectations, we have to consider an extended form of sufficient statistics for the model. Therefore we introduce the following notations and assumptions.

The individual complete data likelihood function is given for all $1 \leq i \leq n$ by:

$$f_i(z_i;\theta) = \exp\left(-\psi_i(\theta) + \langle S_i(z_i), \phi_i(\theta)\rangle\right),$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product, S_i is a function on $\mathbb{R}^{d_i}$ taking its values in a subset $\mathcal{S}_i$ of $\mathbb{R}^{m_i}$.

Let us denote for all $1 \leq i \leq n$ by L_i the function defined on $\mathcal{S}_i \times \Theta$ by $L_i(s_i;\theta) \triangleq -\psi_i(\theta) + \langle s_i, \phi_i(\theta)\rangle$ and by $L : \mathcal{S} \times \Theta \to \mathbb{R}$ the function defined as $L(s,\theta) = \sum_i L_i(s_i;\theta)$ with $\mathcal{S} = \prod_i \mathcal{S}_i$ and $s = (s_1, \ldots, s_n)$. For sake of simplicity, we omitted here all dependency on the observations $(y_i)_{1\leq i\leq n}$ since the considered stochasticity relies on the latent variables.

Finally let us denote by $(\gamma_k)_{k\geq 1}$ a sequence of positive step sizes.

Moreover we assume that there exists a function $\widehat{\theta} :\ \mathcal{S} \to \Theta$, such that $\forall s \in \mathcal{S},\ \ \forall \theta \in \Theta,\ \ L(s;\widehat{\theta}(s)) \geq L(s;\theta)$.

- **Initialization step:** Initialize arbitrarily for all $1 \leq i \leq n$ s_i^0 and θ_0.
- **Repeat until convergence the three following steps:**
  - **Simulation step:** for $1 \leq i \leq n$ simulate a realization Z_i^k from the conditional distribution denoted by p_i using the current parameter value θ_{k-1}.

- **Stochastic approximation step:** compute the quantities for all $1 \leq i \leq n$

$$s_i^k = (1-\gamma_k)s_i^{k-1} + \gamma_k S_i(Z_i^k)$$

where (γ_k) is a sequence of positive step sizes satisfaying $\sum \gamma_k = \infty$ and $\sum \gamma_k^2 < \infty$.

- **Maximisation step:** update of the parameter estimator according to:

$$\theta_k = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \left(-\psi_i(\theta) + \left\langle s_i^k, \phi_i(\theta) \right\rangle \right) = \hat{\theta}(s^k)$$

- **When convergence is reached, say at iteration K of the algorithm, evaluate the FIM estimator according to:**

$$I_{n,sco}^K = \frac{1}{n} \sum_{i=1}^{n} \hat{\Delta}_i \left(\hat{\theta}\left(s^K\right) \right) \hat{\Delta}_i \left(\hat{\theta}\left(s^K\right) \right)^t$$

where $\hat{\Delta}_i(\hat{\theta}(s)) = -\partial\psi_i(\hat{\theta}(s)) + \left\langle s_i, \partial\phi_i(\hat{\theta}(s)) \right\rangle$ for all s.

Remark 3.3 *In the cases where the latent variables can not be simulated from the conditional distribution, one can apply the extension coupling the stochastic algorithm with a Monte Carlo Markov Chain procedure as presented in Kuhn and Lavielle (2004). All the following results can be extended to this case.*

3.2.2 Theoretical convergence properties

In addition to the exponential family assumption for each individual likelihood, we also make the same type of regularity assumptions as those presented in Delyon et al. (1999) at each individual level. These assumptions are detailed in the appendix section.

Theorem 3.4 *Assume that assumptions* $(M1')$ *and* $(M2')$, $(M3)$ *to* $(M5)$ *and* $(SAEM1)$ *to* $(SAEM4)$ *are fulfilled. Assume also that with probability 1* $\operatorname{clos}(\{s_k\}_{k\geq 1})$ *is a compact subset of* $\mathcal{S}$. *Let us define* $\mathcal{L} = \{\theta \in \Theta, \partial_\theta l(y;\theta) = 0\}$ *the set of stationary points of the observed log-likelihood* l. *Then, for all* $\theta_0 \in \Theta$, *for fixed* $n \in \mathbb{N}^*$, *we get:* $\lim_k d(\theta_k, \mathcal{L}) = 0$ *and* $\lim_k d(I_{n,sco}^k, \mathcal{I}) = 0$ *where* $\mathcal{I} = \{I(\theta), \theta \in \mathcal{L}\}$.

Proof Let us denote by $S(Z) = (S_1(Z_1), \ldots, S_n(Z_n))$ the sufficient statistics of the model we consider in our approach. Note as recall in (Delyon et al., 1999), these are not unique. Let us also define $H(Z,s) = S(Z) - s$ and $h(s) = \mathrm{E}_{Z|Y;\theta}(S(Z)) - s$. Our assumptions $(M1')$ and $(M2')$ imply that assumptions $(M1)$ and $(M2)$ of Theorem 5 of Delyon et al. (1999) are fulfilled. Indeed our assumptions focus on expressions and regularity properties of the individual likelihood functions and the corresponding sufficient statistics for each indice $i \in \{1, \ldots, n\}$. Then by linearity of the log-likelihood

function and of the stochastic approximation and applying Theorem 5 of Delyon et al. (1999), we get that $\lim_k d(\theta_k, \mathcal{L}) = 0$. Moreover we get that for $1 \leq i \leq n$, each sequence (s_i^k) converges almost surely toward $\mathrm{E}_{Z_i|Y_i;\theta}(S_i(Z_i))$. Since assumption $(M2')$ ensures that for all $1 \leq i \leq n$ the functions ψ_i and ϕ_i are twice continuously differentiable and assumption $(M5)$ ensures that the function $\hat{\theta}$ is continuously differentiable, the function Φ_n defined by $\Phi_n(s^k) = \frac{1}{n}\sum_{i=1}^n \hat{\Delta}_i(\hat{\theta}(s^k))\hat{\Delta}_i(\hat{\theta}(s^k))$ is continuous. Therefore we get that $\lim_k d(I_{n,sco}^k, \mathcal{I}) = 0$. □

We now establish the asymptotic normality of the estimate $\bar{I}_{n,sco}^k$ defined as $\bar{I}_{n,sco}^k = \Phi_n(\bar{s}^k)$ with $\bar{s}^k = \sum_{l=1}^k s^l/k$ using the results stated by Delyon et al. (1999). Let us denote by $Vect(A)$ the vector composed of the elements of the triangular superior part of matrix A ordered by columns.

Theorem 3.5 *Assume that assumptions $(M1')$ and $(M2')$, $(M3)$ to $(M5)$, $(SAEM1')$, $(SAEM2)$, $(SAEM3)$, $(SAEM4')$ and $(LOC1)$ to $(LOC3)$ are fulfilled. Then, there exists a regular stable stationary point $\theta^* \in \Theta$ such that $\lim_k \theta_k = \theta^*$ a.s. Moreover the sequence $(\sqrt{k}(Vect(\bar{I}_{n,sco}^k) - Vect(\bar{I}_{n,sco}(\theta^*))))\mathbb{1}_{\lim \|\theta_k - \theta^*\| = 0}$ converges in distribution toward a centered Gaussian random vector when k goes to infinity. The asymptotic covariance matrix is characterised.*

Proof The proof follows the lines of this of Theorem 7 of Delyon et al. (1999). Assumptions $(LOC1)$ to $(LOC3)$ are those of Delyon et al. (1999) and ensures the existence of a regular stable stationary point s^* for h and therefore of $\theta^* = \hat{\theta}(s^*)$ for the observed log-likelihood l. Then applying Theorem 4 of Delyon et al. (1999), we get that:

$$\sqrt{k}(\bar{s}^k - s^*)\mathbb{1}_{\lim \|s^k - s^*\| = 0} \xrightarrow{\mathcal{L}} \mathcal{N}(0, J(s^*)^{-1}\Gamma(s^*)J(s^*)^{-1})\mathbb{1}_{\lim \|s_k - s^*\| = 0}$$

where the function Γ defined in assumption $(SAEM4')$ and J is the Jacobian matrix of the function h. Applying the Delta method, we get that:

$$\sqrt{k}(Vect(\Phi_n(\bar{s}^k)) - Vect(\Phi_n(s^*)))\mathbb{1}_{\lim \|s^k - s^*\| = 0} \xrightarrow{\mathcal{L}} W\mathbb{1}_{\lim \|s^k - s^*\| = 0}$$

where $W \sim \mathcal{N}(0, \partial Vect(\Phi_n(s^*))J(s^*)^{-1}\Gamma(s^*)J(s^*)^{-1}\partial Vect(\Phi_n(s^*))^t)$ which leads to the result. □

Note that as usually in stochastic approximation results, the rate $\sqrt{k}$ is achieved when considering an average estimator (see Theorem 7 of Delyon et al. (1999) e.g).

3.2.3 Description of the algorithm for general latent variables models

In general settings, the SAEM algorithm can yet be applied to approximate numerically the maximum likelihood estimate of the model parameter. Nevertheless there are no more theoretical guarantees of convergence for the algorithm. However we propose an extended version of our algorithm which allows to get an estimate of the Fisher information matrix as a by-product of the estimation algorithm.

- **Initialization step:** Initialize arbitrarily Δ_i^0 for all $1 \leq i \leq n$, Q_0 and θ_0.
- **Repeat until convergence the three following steps:**
  - **Simulation step:** for $1 \leq i \leq n$ simulate a realization Z_i^k from the conditional distribution p_i using the current parameter θ_{k-1}.
  - **Stochastic approximation step:** compute the quantities for all $1 \leq i \leq n$

$$\begin{aligned} Q_k(\theta) &= (1-\gamma_k)Q_{k-1}(\theta) + \gamma_k \sum_{i=1}^{n} \log f(y_i, Z_i^k; \theta) \\ \Delta_i^k &= (1-\gamma_k)\Delta_i^{k-1} + \gamma_k \partial_\theta \log f(y_i, Z_i^k; \theta_{k-1}) \end{aligned}$$

  - **Maximisation step:** update of the parameter estimator according to:

$$\theta_k = \underset{\theta}{\operatorname{argmax}}\, Q_k(\theta).$$

- **When convergence is reached, say at iteration K of the algorithm, evaluate the FIM estimator according to**:

$$I_{n,sco}^K = \frac{1}{n} \sum_{i=1}^{n} \Delta_i^K (\Delta_i^K)^t.$$

We illustrate through simulations in a nonlinear mixed effects model the performance of this algorithm in Section 4.2.

4 Simulation study

In this section, we investigate both the properties of the estimators $I_{n,sco}(\theta)$ and $I_{n,obs}(\theta)$ when the sample size n grows and the properties of the proposed algorithm when the number of iterations grows.

4.1 Asymptotic properties of the estimators $I_{n,sco}(\theta)$ and $I_{n,obs}(\theta)$

4.1.1 Simulation setting

First we consider the following linear mixed effects model $y_{ij} = \beta + z_i + \varepsilon_{ij}$, where $y_{ij} \in \mathbb{R}$ denotes the j^{th} observation of individual i, $1 \leq i \leq n$, $1 \leq j \leq J$, $z_i \in \mathbb{R}$ the unobserved random effect of individual i and $\varepsilon_{ij} \in \mathbb{R}$ the residual term. The random effects (z_i) are assumed independent and identically distributed such that $z_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \eta^2)$, the residuals (ε_{ij}) are assumed independent and identically distributed such that $\varepsilon_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ and the sequences (z_i) and (ε_{ij}) are assumed mutually independent. Here, the model parameters are given by $\theta = (\beta, \eta^2, \sigma^2)$. We set $\beta = 3$, $\eta^2 = 2$, $\sigma^2 = 5$ and $J = 12$.

Second we consider the following Poisson mixture model where the distribution of each observation y_i $(1 \leq i \leq n)$ depends on a state variable z_i which is latent leading

to $y_i|z_i = k \sim \mathcal{P}(\lambda_k)$ with $P(z_i = k) = \alpha_k$ and $\sum_{k=1}^{K} \alpha_k = 1$. The model parameters are $\theta = (\lambda_1, \ldots, \lambda_K, \alpha_1, \ldots, \alpha_{K-1})$. For the simulation study, we consider a mixture of $K = 3$ components, and the following values for the parameters $\lambda_1 = 2$, $\lambda_2 = 5$, $\lambda_3 = 9$, $\alpha_1 = 0.3$ and $\alpha_2 = 0.5$.

4.1.2 Results

For each model, we generate $M = 500$ datasets for different sample sizes $n \in \{20, 100, 500\}$. We do not aim at estimating the model parameters. We assume them to be known, and in the following we denote by $\theta^\star$ the true parameter value. For each value of n and for each $1 \leq m \leq M$, we derive $I^{(m)}_{n,sco}(\theta^\star)$ and $I^{(m)}_{n,obs}(\theta^\star)$. We compute the empirical bias and the root mean squared deviation of each component (ℓ, ℓ') of the estimated matrix as:

$$\frac{1}{M}\sum_{m=1}^{M} I^{(m)}_{n,sco,\ell,\ell'}(\theta^\star) - I_{\ell,\ell'}(\theta^\star) \quad \text{and} \quad \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(I^{(m)}_{n,sco,\ell,\ell'}(\theta^\star) - I_{\ell,\ell'}(\theta^\star)\right)^2}.$$

In the previous quantities, $I(\theta^\star)$ is explicit in the linear mixed effects model and approximated by a Monte-Carlo estimation based on a large sample size in the Poisson mixture model. The results are presented in Tables 1 and 2 for the linear mixed effects model and in Tables 3 and 4 for the mixture model. We observe that whatever the model and whatever the components of $I_{n,sco}(\theta^\star)$ and $I_{n,obs}(\theta^\star)$, the bias is very small even for small values of n. Note that in this particular model the second derivatives with respect to parameter β is deterministic, which explains why the bias and the dispersion of the estimations $I_{n,obs}(\theta^\star)$ are zero for every value of n. The bias and the standard deviation decrease as n increases overall, which illustrates the consistency of both M-estimators. We also represent in Figures 1 and 2 the distributions of the normalized estimations $\sqrt{n}\left(I^{(m)}_{n,sco}(\theta^\star) - I(\theta^\star)\right)$ and $\sqrt{n}\left(I^{(m)}_{n,obs}(\theta^\star) - I(\theta^\star)\right)$ for $n = 500$ for some components of the matrices. The empirical distributions have the shape of Gaussian distributions and illustrate the asymptotic normality of the two estimators. The numerical results highlight that neither $I_{n,sco}(\theta^\star)$ nor $I_{n,obs}(\theta^\star)$ is systematically better than the other one in terms of bias and asymptotic covariance matrix.

4.2 Properties of the stochastic approximation algorithm

4.2.1 Curved exponential family model

We consider the following nonlinear mixed effects model which is widely used in pharmacokinetics for describing the evolution of drug concentration over time.

$$y_{ij} = \frac{d_i ka_i}{V_i ka_i - Cl_i}\left[e^{-\frac{Cl_i}{V_i}t_{ij}} - e^{-ka_i t_{ij}}\right] + \varepsilon_{ij}, \tag{3}$$

where ka_i, Cl_i and V_i are individual random parameters such that $\log ka_i = \log(ka) + z_{i,1}$, $\log Cl_i = \log(Cl) + z_{i,2}$, $\log V_i = \log(V) + z_{i,3}$. For all $1 \leq i \leq n$, $1 \leq j \leq J$, y_{ij} denotes the measure of drug concentration on individual i at time t_{ij}, d_i the dose of drug

Table 1: Linear mixed effects model. Empirical bias and squared deviation to the Fisher Information matrix (in brackets) of $I_{n,sco}$ for different values of n.

n	$I_{n,sco}(\beta,\beta)$	$I_{n,sco}(\eta^2,\eta^2)$	$I_{n,sco}(\sigma^2,\sigma^2)$	$I_{n,sco}(\beta,\eta^2)$	$I_{n,sco}(\beta,\sigma^2)$	$I_{n,sco}(\eta^2,\sigma^2)$
20	0.015	0.007	-0.007	0.002	-0.004	-3.10^{-4}
	(0.141)	(0.009)	(0.085)	(0.102)	(0.068)	(0.032)
100	-0.001	-2.10^{-4}	-0.001	0.001	0.002	4.10^{-4}
	(0.057)	(0.030)	(0.039)	(0.039)	(0.031)	(0.014)
500	-0.001	-7.10^{-4}	-1.10^{-4}	5.10^{-4}	-4.10^{-4}	-4.10^{-5}
	(0.026)	(0.014)	(0.017)	(0.018)	(0.013)	(0.006)

Table 2: Linear mixed effects model. Empirical bias and squared deviation to the Fisher Information matrix (in brackets) of $I_{n,obs}$ for different values of n.

n	$I_{n,obs}(\beta,\beta)$	$I_{n,obs}(\eta^2,\eta^2)$	$I_{n,obs}(\sigma^2,\sigma^2)$	$I_{n,obs}(\beta,\eta^2)$	$I_{n,obs}(\beta,\sigma^2)$	$I_{n,obs}(\eta^2,\sigma^2)$
20	0.000	0.007	-0.002	0.001	1.10^{-4}	5.10^{-4}
	(0.000)	(0.058)	(0.042)	(0.058)	(0.005)	(0.005)
100	0.000	-5.10^{-4}	2.10^{-4}	-0.002	-2.10^{-4}	4.10^{-5}
	(0.000)	(0.023)	(0.018)	(0.026)	(0.002)	(0.002)
500	0.000	-5.10^{-4}	-8.10^{-5}	4.10^{-4}	4.10^{-5}	-4.10^{-5}
	(0.000)	(0.011)	(0.009)	(0.012)	(0.001)	(0.001)

Table 3: Mixture model. Empirical bias and squared deviation to the Fisher Information matrix (in brackets) of some components of $I_{n,sco}$ for different values of n.

n	$I_{n,sco}(\lambda_2,\lambda_2)$	$I_{n,sco}(\lambda_3,\lambda_3)$	$I_{n,sco}(\alpha_1,\alpha_1)$	$I_{n,sco}(\alpha_2,\alpha_2)$	$I_{n,sco}(\lambda_2,\lambda_3)$	$I_{n,sco}(\lambda_3,\alpha_2)$
20	8.10^{-5}	3.10^{-5}	0.060	0.047	9.10^{-5}	-0.002
	(0.007)	(0.015)	(1.202)	(1.056)	(0.003)	(0.110)
100	-3.10^{-5}	-2.10^{-4}	-0.040	-0.041	-7.10^{-5}	0.003
	(0.003)	(0.007)	(0.526)	(0.469)	(0.001)	(0.046)
500	7.10^{-5}	7.10^{-5}	0.019	0.011	2.10^{-5}	-1.10^{-4}
	(0.001)	(0.003)	(0.232)	(0.205)	(0.001)	(0.021)

Table 4: Mixture model. Empirical bias and squared deviation to the Fisher Information matrix (in brackets) of some components of $I_{n,obs}$ for different values of n.

n	$I_{n,obs}(\lambda_2,\lambda_2)$	$I_{n,obs}(\lambda_3,\lambda_3)$	$I_{n,obs}(\alpha_1,\alpha_1)$	$I_{n,obs}(\alpha_2,\alpha_2)$	$I_{n,obs}(\lambda_2,\lambda_3)$	$I_{n,obs}(\lambda_3,\alpha_2)$
20	-3.10^{-4}	5.10^{-4}	0.060	0.047	9.10^{-5}	9.10^{-4}
	(0.022)	(0.009)	(1.202)	(1.055)	(0.003)	(0.034)
100	2.10^{-4}	-4.10^{-4}	-0.040	-0.041	-7.10^{-5}	-5.10^{-4}
	(0.010)	(0.004)	(0.526)	(0.469)	(0.001)	(0.016)
500	-3.10^{-4}	1.10^{-4}	0.019	0.011	2.10^{-5}	5.10^{-4}
	(0.005)	(0.002)	(0.232)	(0.205)	(6.10^{-4})	(0.007)

Figure 1: Linear mixed effects model. Kernel density estimates of the normalized values $\sqrt{n}\left(I^{(m)}_{n,sco,\ell,\ell'}(\theta^\star) - I_{\ell,\ell'}(\theta^\star)\right)$ and $\sqrt{n}\left(I^{(m)}_{n,obs,\ell,\ell'}(\theta^\star) - I_{\ell,\ell'}(\theta^\star)\right)$ of three components of the estimated Fisher information matrix computed from the $M = 500$ simulated datasets when $n = 500$.

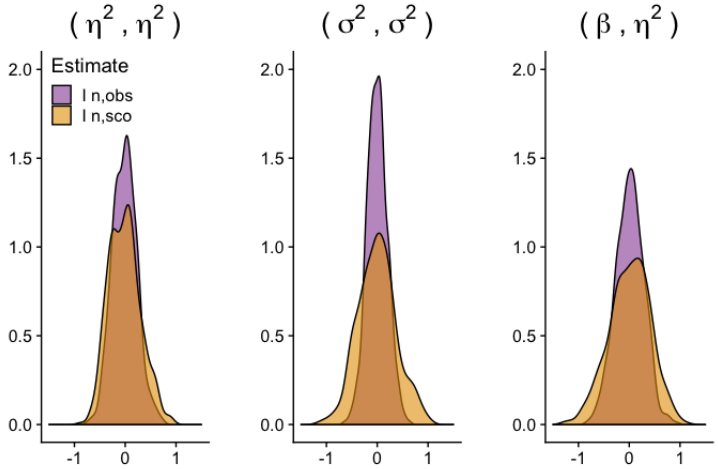
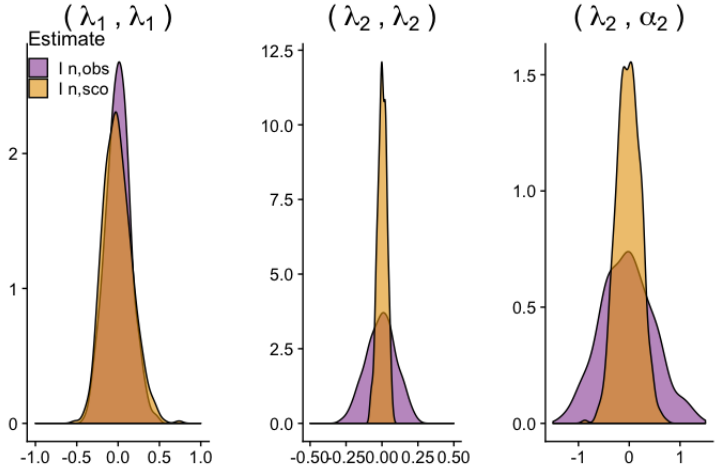


Figure 2: Mixture model. Kernel density estimates of the normalized values $\sqrt{n}\left(I^{(m)}_{n,sco,\ell,\ell'}(\theta^\star) - I_{\ell,\ell'}(\theta^\star)\right)$ and $\sqrt{n}\left(I^{(m)}_{n,obs,\ell,\ell'}(\theta^\star) - I_{\ell,\ell'}(\theta^\star)\right)$ of three components of the estimated Fisher information matrix computed from the $M = 500$ simulated datasets when $n = 500$.



administered to individual i, and V_i, ka_i and Cl_i respectively denote the volume of the central compartment, the drug's absorption rate constant and the drug's clearance of individual i. The terms $z_i = (z_{i,1}, z_{i,2}, z_{i,3})' \in \mathbb{R}^3$ are unobserved random effects which are assumed independent and identically distributed such that $z_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega)$, where $\Omega = \text{diag}(\omega_{ka}^2, \omega_{Cl}^2, \omega_V^2)$, the residuals (ε_{ij}) are assumed independent and identically distributed such that $\varepsilon_{ij} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ and the sequences (z_i) and (ε_{ij}) are assumed mutually independent. Here, the model parameter is given by $\theta = (ka, V, Cl, \omega_{ka}^2, \omega_V^2, \omega_{Cl}^2, \sigma^2)$. In this model, as in a large majority of nonlinear mixed effects models, the likelihood does not have any analytical expression. As a consequence, neither the Fisher Information Matrix, nor the estimators $I_{n,sco}(\theta)$, $I_{n,obs}(\theta)$ have explicit expressions. However, as the complete data log-likelihood is explicit, stochastic approximations of $I_{n,sco}(\theta)$, $I_{n,obs}(\theta)$ can be implemented. We take the following values for the parameters $V = 31$, $ka = 1.6$, $Cl = 1.8$, $\omega_V^2 = 0.40$, $\omega_{ka}^2 = 0.40$, $\omega_{Cl}^2 = 0.40$ and $\sigma^2 = 0.75$. We consider the same dose $d_i = 320$ and the same observation times (in hours): 0.25,0.5, 1, 2, 3.5, 5, 7, 9, 12, 24 for all the individuals. We simulate one dataset with $n = 100$ individuals under model (3). On this simulated dataset, we run the stochastic approximation algorithm described in section 3.2.1 for computing $I_{n,sco}(\hat{\theta})$ together with $\hat{\theta}$ and the algorithm of Delyon et al. (1999) for computing $I_{n,obs}(\hat{\theta})$ $M = 500$ times. We perform $K = 3000$ iterations in total for each algorithm by setting $\gamma_k = 0.95$ for $1 \leq k \leq 1000$ (burn in iterations) and $\gamma_k = (k - 1000)^{-3/5}$ otherwise. At any iteration, we compute the empirical relative bias and the empirical relative standard deviation of each component (ℓ, ℓ') of $I_{n,sco}$ defined respectively as:

$$\frac{1}{M}\sum_{m=1}^{M} \frac{\widehat{I_{n,sco,\ell,\ell'}^{(k,m)}} - I_{n,sco,\ell,\ell'}^{\star}}{I_{n,sco,\ell,\ell'}^{\star}} \quad \text{and} \quad \sqrt{\frac{1}{M}\sum_{m=1}^{M} \left(\frac{\widehat{I_{n,sco,\ell,\ell'}^{(k,m)}} - I_{n,sco,\ell,\ell'}^{\star}}{I_{n,sco,\ell,\ell'}^{\star}} \right)^2}$$

where $\widehat{I_{n,sco}^{(k,m)}}$ denote the estimated value of $I_{n,sco}(\hat{\theta})$ at iteration k of the m^{th} algorithm. We compute the same quantities for $I_{n,obs}$. As the true values of $I_{n,sco}^{\star} = I_{n,sco}(\theta^{\star})$ and $I_{n,obs}^{\star} = I_{n,obs}(\theta^{\star})$ are not known, they are estimated by Monte-Carlo integration. The results are displayed in Figures 3 and 4.

We observe that the bias and the standard deviations of the estimates of the components of both matrices decrease over iterations, and that for both estimates the bias is nearly zero when the convergence of the algorithm is reached. According to these simulation results, there is no evidence that one method is better than the other in terms of bias or standard deviation.

4.2.2 A general latent variable model

We use model (3) again, but we now consider that individual parameter V_i is fixed, *i.e.* $V_i \equiv V \ \forall i = 1, \ldots, n$. The model is no longer exponential in the sense of equation (3.2.1). We must therefore use the general version of the stochastic approximation algorithm from

Figure 3: Non linear mixed effects model. Representation over iterations of the mean relative biais of the diagonal components of the estimated Fisher information matrix computed from the $M = 500$ runs of the stochastic algorithm. Red line corresponds to $I_{n,sco}(\theta)$ and blue line corresponds to $I_{n,obs}(\theta)$. The burn-in iterations of the algorithm are not depicted.

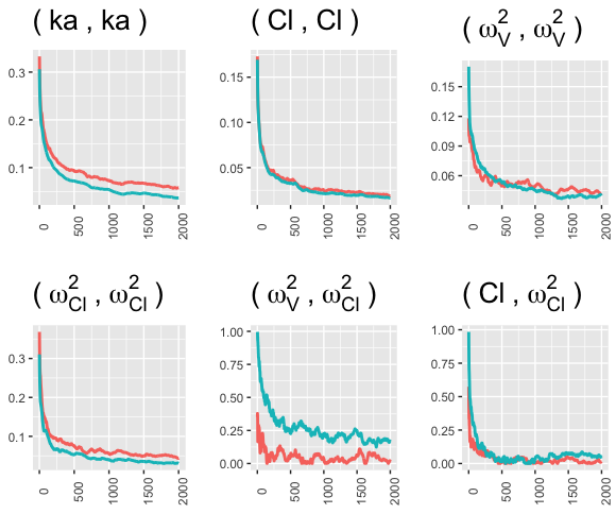
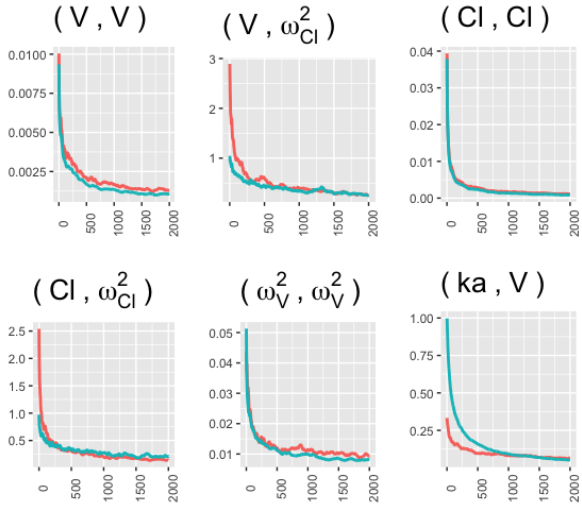


Figure 4: Non linear mixed effects model. Representation over iterations of the mean relative standard error of the diagonal components of the estimated Fisher information matrix computed from the $M = 500$ runs of the stochastic algorithm. Red line corresponds to $I_{n,sco}(\theta)$ and blue line corresponds to $I_{n,obs}(\theta)$. The burn-in iterations of the algorithme are not depicted.



section 3.2.3 to compute $I_{n,sco}(\hat{\theta})$. We simulate 500 datasets according to this model and we then estimate $I_{n,sco}(\hat{\theta})$ and $\hat{\theta}$ for each one. We then compute the 500 asymptotic confidence intervals of the model parameters $[\hat{\theta}_k^{(\ell)} - q_{1-\alpha/2}\,\hat{\sigma}_k^{(\ell)}, \hat{\theta}_k^{(\ell)} + q_{1-\alpha/2}\,\hat{\sigma}_k^{(\ell)}]$, $\ell = 1, \ldots, 6$ and then deduce from them empirical coverage rates. The $\hat{\sigma}_k^{(\ell)}$'s are obtained through the diagonal terms of the inversed $V_n(\hat{\theta}_k)$'s, and $q_{1-\alpha/2}$ stands for the quantile of order $1-\alpha/2$ of a standard Gaussian distribution with zero mean. We obtain for the six parameters $(ka, V, Cl, \omega_{ka}^2, \omega_{Cl}^2, \sigma^2)$ empirical covering rates of 0.946,0.928,0.962,0.944,0.950,0.942 respectively for a nominal covering rate of 0.95. Although theoretical guarantee is missing in non exponential models, the stochastic approximation algorithm proposed in section 3.2.3 converges in practice on this example for both the estimation of the model parameters and the estimation of the Fisher information matrix as shown by Figures 5 and 6.

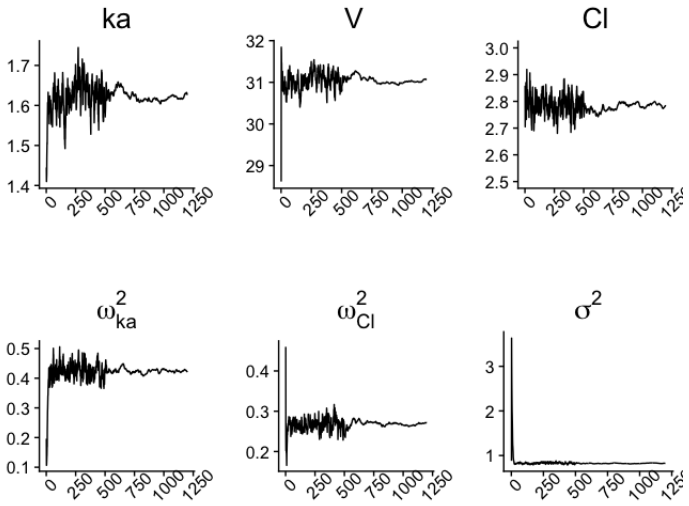


Figure 5: Convergence plot for the parameter estimates over iterations of the stochastic approximation algorithm.

4.3 Comparison with other methods

To the best of our knowledge, although there exists contributions focusing on the estimation of the Fisher information matrix in latent variable models, there is currently no method based on the first derivatives of the log-likelihood. We compare to Meng and Spall (2017) who proposed an iterative method based on numerical first order derivatives of the Q function that is computed at each E-step of the EM algorithm. The model used by Meng and Spall (2017) in their simulation study is a mixture of two Gaussian distributions with unknown expectations μ_1 and μ_2, fixed variances equal to 1 and unknown proportion π. The model parameters are denoted by $\theta = (\mu_1, \mu_2, \pi)$.

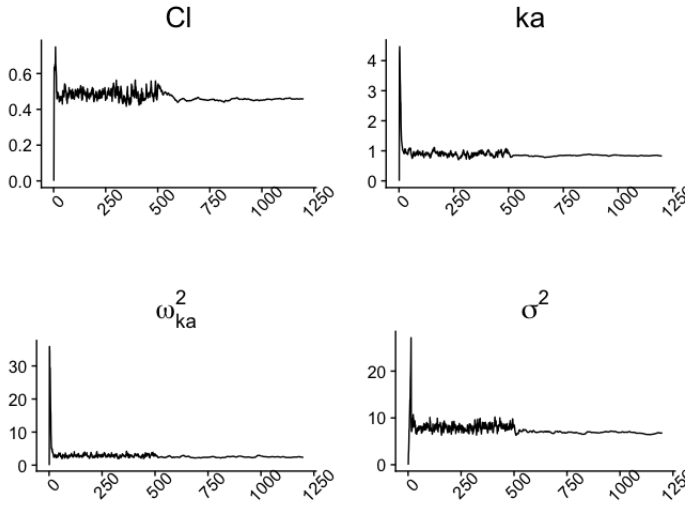


Figure 6: Convergence plot for the some diagonal components of $I_{n,sco}(\hat{\theta})$ over iterations of the stochastic approximation algorithm.

We simulate 10000 datasets according to this Gaussian mixture model, using the same setting as Meng and Spall (2017), i.e. $n = 750$, $\pi = 2/3$, $\mu_1 = 3$ and $\mu_2 = 0$. For each dataset $k = 1, \ldots, 10000$, we compute the parameter maximum likelihood estimate $\hat{\theta}_k = (\hat{\pi}_k, \widehat{\mu_1}_k, \widehat{\mu_2}_k)$ with an EM algorithm and then we derive $I_{n,sco}(\hat{\theta}_k)$ directly according to formula (3) contrary to Meng and Spall (2017) who used an iterative method. We compute the empirical mean of the 10000 estimated matrices leading to:

$$\frac{1}{10000}\sum_k I_{n,sco}(\hat{\theta}_k) = \begin{pmatrix} 2685.184 & -211.068 & -251.808 \\ -211.068 & 170.927 & -61.578 \\ -251.808 & -61.578 & 392.859 \end{pmatrix}$$

Comparison with the results of Meng and Spall (2017) is delicate since their numerical illustration of their method is based on a single simulated dataset thus potentially sensitive to sampling variations. However, they provide an estimation of the Fisher information matrix from this unique dataset

$$I_{Meng} = \begin{pmatrix} 2591.3 & -237.9 & -231.8 \\ -237.9 & 155.8 & -86.7 \\ -231.8 & -86.7 & 394.5 \end{pmatrix}.$$

Our results are coherent with their ones. To check the reliability of our results, we then compute as above the 10000 asymptotic confidence intervals of the three model parameters. We obtain for the three parameters (π, μ_1, μ_2) empirical covering rates of $0.953, 0.949, 0.951$ respectively for a nominal covering rate of 0.95.

5 Conclusion and discussion

In this work, we address the estimation of the Fisher information matrix in general latent variable models. We consider the moment estimate of the covariance matrix of the score whereas the observed FIM, equal to the moment estimate based on the expression of the FIM equal to minus the expectation of the Hessian of the log-likelihood, is usually used in practice. We detailed the theoretical properties of both estimates. We propose a stochastic approximation algorithm to compute the proposed estimate of the FIM when it can not be calculated analytically and establish its theoretical convergence properties. We carry out a simulation study in mixed effects model and a Poisson mixture model to compare the performances of both estimates and of the proposed algorithm. We emphasize that the moment estimate of the covariance matrix of the score requires less regularity assumptions than the observed FIM, leading in the same time to less derivative calculus. From a computational point of view, the implementation of the algorithm for the moment estimate of the covariance matrix of the score only involves the first derivatives of the log-likelihood, in contrary to the other moment estimate which involves the second derivatives of the log-likelihood.

The main perspective of this work is to adapt the procedure for statistical models whose derivatives of the log-likelihood have no tractable expressions, coupling the algorithm with numerical derivative procedures. It would be particularly interesting to consider mechanistic models such as crop models for example (Technow et al. (2015)).

References

Baey, C., Cournède, P.-H. and Kuhn, E. (2019) Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, **135**, 107–122.

Delattre, M., Savic, R., Miller, R., Karlsson, M. and Lavielle, M. (2012) Analysis of exposure-response of ci-945 in patients with epilepsy: application of novel mixed hidden markov modelling methodology. *Journal of Pharmacokinetics and Pharmacodynamics*, **39**, 263–271.

Delyon, B., Lavielle, M. and Moulines, E. (1999) Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, **27**, 94–128.

Gloaguen, P., Mahévas, S., Rivot, E., Woillez, M., Guitton, J., Vermard, Y. and Etienne, M.-P. (2014) An autoregressive model to describe fishing vessel movement and activity. *Environmetrics*, **26**, 17–28.

Kuhn, E. and Lavielle, M. (2004) Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM P&S*, 115–131.

Louis, T. A. (1982) Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**.

Meng, L. and Spall, J. (2017) Efficient computation of the fisher information matrix in the em algorithm. Annual Conference on Information Sciences and Systems (CISS).

Meng, X.-L. and Rubin, D. B. (1991) Using em to obtain asymptotic variance-covariance matrices: the sem algorithm. *Journal of the American Statistical Association*, **86**, 899–909.

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

Oakes, D. (1999) Direct calculation of the information matrix via the em algorithm. *J.R. Static. Soc. B*, **61**, 479–482.

Orchard, T. and Woodbury, M. A. (1972) A missing information principle: theory and applications. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability.

Picard, F., Robin, S., Lebarbier, E. and Daudin, J. (2007) A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, **63**, 758–766.

Silvapulle, M. J. and Sen, P. K. (2011) *Constrained statistical inference: Order, inequality, and shape constraints*, vol. 912. John Wiley & Sons.

Technow, F., Messina, C. D., Totir, L. R. and Cooper, M. (2015) Integrating crop growth models with whole genome prediction through approximate bayesian computation. *Plos One*, **10**.

van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

6 Appendix

It is assumed that the random variables $s^0, z_1, z_2, \cdots$ are defined on the same probability space $(\Omega, \mathcal{A}, P)$. We denote $\mathcal{F} = \{\mathcal{F}_k\}_{k\geq 0}$ the increasing family of σ-algebras generated by the random variables $s_0, z_1, z_2, \cdots, z_k$. We assume the following conditions:

- **(M1')** The parameter space Θ is an open subset of $\mathbb{R}^p$. The individual complete data likelihood function is given for all $i = 1, \ldots, n$ by:

$$f_i(z_i; \theta) = \exp\left(-\psi_i(\theta) + \langle S_i(z_i), \phi_i(\theta)\rangle\right),$$

  where $\langle \cdot, \cdot \rangle$ denotes the scalar product, S_i is a Borel function on $\mathbb{R}^{d_i}$ taking its values in an open subset $\mathcal{S}_i$ of $\mathbb{R}^{m_i}$. Moreover, the convex hull of $S(\mathbb{R}^{\sum d_i})$ is included in $\mathcal{S}$ and for all $\theta \in \Theta$ $\int S(z) \prod p_i(z_i; \theta)\mu(dz) < \infty$

- **(M2')** Define for each i $L_i : \mathcal{S}_i \times \Theta \to \mathbb{R}$ as $L_i(s_i; \theta) \triangleq -\psi_i(\theta) + \langle s_i, \phi_i(\theta)\rangle$. The functions ψ_i and ϕ_i are twice continuously differentiable on Θ.

- **(M3)** The function $\bar{s} : \Theta \to \mathcal{S}$ defined as $\bar{s}(\theta) \triangleq \int S(z)p(z;\theta)\mu(dz)$ is continuously differentiable on Θ.
- **(M4)** The function $l : \Theta \to \mathbb{R}$ defined as $l(\theta) \triangleq \log g(\theta) = \log \int_{\mathbb{R}^{d_z}} f(z;\theta)\mu(dz)$ is continuously differentiable on Θ and $\partial_\theta \int f(z;\theta)\mu(dz) = \int \partial_\theta f(z;\theta)\mu(dz)$.
- **(M5)** There exists a continuously differentiable function $\widehat{\theta} : \mathcal{S} \to \Theta$, such that:

$$\forall s \in \mathcal{S}, \quad \forall \theta \in \Theta, \quad L(s;\widehat{\theta}(s)) \geq L(s;\theta).$$

In addition, we define:

- **(SAEM1)** For all k in $\mathbb{N}$, $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.
- **(SAEM2)** $l : \Theta \to \mathbb{R}$ and $\widehat{\theta} : \mathcal{S} \to \Theta$ are m times differentiable, where m is the integer such that $\mathcal{S}$ is an open subset of $\mathbb{R}^m$.
- **(SAEM3)** For all positive Borel functions Φ $E[\Phi(z_{k+1})|\mathcal{F}_k] = \int \Phi(z)p(z;\theta_k)\mu(dz)$.
- **(SAEM4)** For all $\theta \in \Theta$, $\int \|S(z)\|^2 p(z;\theta)\mu(dz) < \infty$, and the function

$$\Gamma(\theta) \triangleq \mathrm{Cov}_\theta[S(z)] \triangleq \int S(z)^t S(z) p(z;\theta)\mu(dz) - \left[\int S(z)p(z;\theta)\mu(dz)\right]^t \left[\int S(z)p(z;\theta)\mu(dz)\right]$$

is continuous w.r.t. θ.

We also define assumptions required for the normality result:

- **(SAEM1')** For all k in $\mathbb{N}$, $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$. There exists γ^* such that $\lim k^\alpha/\gamma_k = \gamma^*$, and $\gamma_k/\gamma_{k+1} = 1 + O(k^{-1})$.
- **(SAEM4')** For some $\alpha > 0$, $\sup_\theta \mathrm{E}_\theta(\|S(Z)\|^{2+\alpha}) < \infty$ and Γ is continuous w.r.t. θ.
- **(LOC1)** The stationary points of l are isolated: any compact subset of Θ contains only a finite number of such points.
- **(LOC2)** For every stationary point θ^*, the matrices $\mathrm{E}^*_\theta(\partial_\theta L(S(Z),\theta^*)(\partial_\theta L(S(Z),\theta^*))^t)$ and $\partial^2_\theta L(\mathrm{E}^*_\theta(S(Z)),\theta^*)$ are positive definite.
- **(LOC3)** The minimum eigenvalue of the covariance matrix $R(\theta) = \mathrm{E}_\theta((S(Z) - \bar{s}(\theta))(S(Z) - \bar{s}(\theta))^t)$ is bounded away from zero for θ in any compact subset of Θ.