



# MONOLIX METHODOLOGY

Version 4.2.1

FEBRUARY 2012

A software for the analysis of nonlinear mixed effects models

## Maximum likelihood estimation

## Model selection

## Hypothesis testing

## Graphical analysis

## Data simulation

• • •

	I	M	P	S T O C H S T I C E M	R	T	A	N	M C M C	E	S	A	H M M	P	L	I	N	G
S	I	M	U	L	A	E	D	A	N	N	S A E M	A	L	I	N	G		
				M		T	R	O	P	O	L	I	S					

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Estimation of the parameters . . . . .	4
1.1.1	The SAEM algorithm . . . . .	4
1.1.2	The MCMC-SAEM algorithm . . . . .	7
1.1.3	The Simulated Annealing SAEM algorithm . . . . .	8
1.2	Some extensions . . . . .	10
1.2.1	Model with censored data . . . . .	10
1.2.2	Estimation with a prior distribution . . . . .	10
1.2.3	Modeling the inter-occasion variability . . . . .	12
1.2.4	Mixture models and mixture of models . . . . .	12
1.3	Estimation of the Fisher Information matrix . . . . .	14
1.3.1	Linearization of the model . . . . .	14
1.3.2	A stochastic approximation of the Fisher Information Matrix . . . . .	14
1.4	Estimation of the individual parameters . . . . .	15
1.5	Estimation of the likelihood . . . . .	16
1.5.1	Linearization of the model . . . . .	16
1.5.2	Estimation using importance sampling . . . . .	17
1.6	Estimation of the weighted residuals . . . . .	18
1.6.1	The Population Weighted Residuals . . . . .	18

---

1.6.2	The Individual Weighted Residuals . . . . .	18
1.6.3	The Normalized Prediction Distribution . . . . .	18
1.7	Inputs and outputs . . . . .	19
1.7.1	The inputs . . . . .	19
1.7.2	The outputs . . . . .	20
<b>2</b>	<b>Statistical models</b>	<b>22</b>
2.1	The nonlinear mixed effects model . . . . .	22
2.2	Individual parameters model . . . . .	23
2.2.1	Examples of transformations . . . . .	23
2.2.2	Example of continuous covariate model . . . . .	24
2.2.3	Example of categorical covariate model . . . . .	24
2.3	The residual error model . . . . .	25
2.4	Multi-responses model . . . . .	26
2.5	Model with censored data . . . . .	26
2.5.1	BLQ data . . . . .	26
2.5.2	Interval censored data . . . . .	27
2.6	Inter-occasion variability . . . . .	27
2.7	Discrete data models . . . . .	27
2.8	Mixture models and model mixtures . . . . .	28
2.8.1	Mixture models . . . . .	28
2.8.2	Model mixtures . . . . .	28
2.9	Prior models . . . . .	29

# Chapter 1

## Introduction

### 1.1 Estimation of the parameters

#### 1.1.1 The SAEM algorithm

We are in a classical framework of incomplete data: the observed data is  $y = (y_{ij}; 1 \leq i \leq N, 1 \leq j \leq n_i)$ , whereas the random parameters  $(\psi = \psi_i; 1 \leq i \leq N)$  are the non observed data. Then, the complete data of the model is  $(y, \psi)$ . Our purpose is to compute the maximum likelihood estimator of the unknown set of parameters  $\theta = (\mu, \Omega, a, b, c)$ , by maximizing the likelihood of the observations  $\ell(y; \theta)$ .

In the case of a linear model, the estimation of the unknown parameters can be treated with the usual EM algorithm. At iteration  $k$  of EM, the E-step consists in computing the conditional expectation of the complete log-likelihood  $Q_k(\theta) = \mathbb{E}(\log p(y, \psi; \theta) | y, \theta_{k-1})$  and the M-step consists in computing the value  $\theta_k$  that maximizes  $Q_k(\theta)$ .

Following [4, 10], the EM sequence  $(\theta_k)$  converges to a stationary point of the observed likelihood (*i.e* a point where the derivative of  $\ell$  is 0) under general regularity conditions. In cases where the regression function  $f$  does not linearly depend on the random effects, the E-step cannot be performed in a closed-form.

The stochastic approximation version of the standard EM algorithm, proposed by [3] consists in replacing the usual E-step of EM by a stochastic procedure. At iteration  $k$  of SAEM:

- *Simulation-step* : draw  $\psi^{(k)}$  from the conditional distribution  $p(\cdot | y; \theta_k)$ .
- *Stochastic approximation* : update  $Q_k(\theta)$  according to

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\log p(y, \psi^{(k)}; \theta) - Q_{k-1}(\theta)) \quad (1.1)$$

where  $(\gamma_k)$  is a decreasing sequence of positive numbers with  $\gamma_1 = 1$ .

- *Maximization-step* : update  $\theta_k$  according to

$$\theta_{k+1} = \text{Arg max}_{\theta} Q_k(\theta).$$

It is shown in [3] that SAEM converges to a maximum (local or global) of the likelihood of the observations under very general conditions.

Here, the complete log-likelihood can be written

$$\begin{aligned} \log p(y, \psi; \theta) &= \log p(y, h(\varphi); \theta) \\ &= - \sum_{i,j} \log(g(x_{ij}, \psi_i, \xi)) - \frac{1}{2} \sum_{i,j} \left( \frac{y_{ij} - f(x_{ij}, \psi_i)}{g(x_{ij}, \psi_i, \xi)} \right)^2 \\ &\quad - \frac{N}{2} \log(|\Omega|) - \frac{1}{2} \sum_{i=1}^N (\varphi_i - C_i \mu)' \Omega^{-1} (\varphi_i - C_i \mu) - \frac{N_{tot} + Nd}{2} \log(2\pi) \end{aligned}$$

where  $N_{tot} = \sum_{i=1}^N n_i$  is the total number of observations.

First, consider a constant residual error model ( $g = a$ ). The set of parameters to estimate is  $\theta = (\mu, \Omega, a)$ . Then, the complete model belongs to the exponential family and the approximation step reduces to only updating the sufficient statistics of the complete model:

$$\begin{aligned} s_{1,i,k} &= s_{1,i,k-1} + \gamma_k (\varphi_{i,k} - s_{1,i,k-1}), \quad i = 1, \dots, N \\ s_{2,k} &= s_{2,k-1} + \gamma_k \left( \sum_{i=1}^N \varphi_{i,k} \varphi'_{i,k} - s_{2,k-1} \right) \\ s_{3,k} &= s_{3,k-1} + \gamma_k \left( \sum_{i,j} \left( y_{ij} - f(x_{ij}, \psi_i^{(k)}) \right)^2 - s_{3,k-1} \right). \end{aligned}$$

Then,  $\theta_{k+1}$  is obtained in the maximization step as follows:

$$\mu_{k+1} = \left( \sum_{i=1}^N C_i' \Omega_k^{-1} C_i \right)^{-1} \sum_{i=1}^N C_i' \Omega_k^{-1} s_{1,i,k} \quad (1.2)$$

$$\Omega_{k+1} = \frac{1}{N} \left( s_{2,k} - \sum_{i=1}^N (C_i \mu_{k+1}) s'_{1,i,k} - \sum_{i=1}^N s_{1,i,k} (C_i \mu_{k+1})' + \sum_{i=1}^N (C_i \mu_{k+1}) (C_i \mu_{k+1})' \right) \quad (1.3)$$

$$a_{k+1} = \sqrt{\frac{s_{3,k}}{N_{tot}}} \quad (1.4)$$

**Remark 1:** The sequence of step sizes used in MONOLIX decreases as  $k^{-a}$ . More precisely, for any sequence of integers  $K_1, K_2, \dots, K_J$  and any sequence  $a_1, a_2, \dots, a_J$  of real numbers such

that  $0 \leq a_1 < a_2 < \dots < a_J \leq 1$ , we define the sequence of step sizes  $(\gamma_k)$  as follows:

$$\gamma_k = \frac{1}{k^{a_1}} \quad \text{for any } 1 \leq k \leq K_1 \quad (1.5)$$

and for  $2 \leq j \leq J$ ,

$$\gamma_k = \frac{1}{\left(k - K_{j-1} + \gamma_{K_{j-1}}^{-1/a_j}\right)^{a_j}} \quad \text{for any } \sum_{i=1}^{j-1} K_i + 1 \leq k \leq \sum_{i=1}^j K_i \quad (1.6)$$

Here,  $K = \sum_{j=1}^J K_j$  is the total number of iterations.

We recommend to use  $a_1 = 0$  (that is  $\gamma_k = 1$ ) during the first iterations, and  $a_J = 1$  during the last iterations. Indeed, the initial guess  $\theta_0$  may be far from the maximum likelihood value we are looking for and the first iterations with  $\gamma_k = 1$  allow to converge quickly to a neighborhood of the maximum likelihood estimator. Then, smaller step sizes ensure the almost sure convergence of the algorithm to the maximum likelihood estimator.

In the case where  $J = 2$  with  $a_1 = 0$  and  $a_2 = 1$ , the sequence of step sizes is

$$\begin{aligned} \gamma_k &= 1 && \text{for } 1 \leq k \leq K_1 \\ &= \frac{1}{k - K_1 + 1} && \text{for } K_1 + 1 \leq k \leq K_1 + K_2 \end{aligned}$$

**Remark 2:** The estimated covariance matrix  $\Omega_{k+1}$  defined in (1.3) is a full covariance matrix. However, the covariance matrix  $\Omega$  of the random effects can have any covariance structure. If we assume, for example, that there is no correlation between the random effects, we will set to 0 the non diagonal elements of  $\Omega_{k+1}$  defined in (1.3).

We can also assume that a random effect has no variance. If the  $\ell$ th random effect has a variance equal to 0, then the  $\ell$ th individual parameter is no longer random and the simulation step of SAEM needs some modification. During the first  $K_0$  iterations, we use SAEM as it was described above, considering that all the effects are random and assuming that there is no correlation between the  $\ell$ th random effect and the other ones ( $\omega_{\ell\ell'}^2 = 0$  for any  $\ell \neq \ell'$ ). Then, during the next iterations, we use again SAEM, but the variance of this random effect is no longer estimated: it is forced to decrease at each iteration by setting

$$\omega_{\ell\ell,k+1}^2 = \alpha \omega_{\ell\ell,k}^2, \quad K_0 \leq k \leq K \quad (1.7)$$

where  $\alpha$  is chosen between 0 and 1 such that  $\omega_{\ell\ell,K}^2 = 10^{-6} \omega_{\ell\ell,K_0}^2$ .

**Remark 3:** - For a residual variance model of the form  $g = b f^c$ , where  $c$  is fixed, the complete model also belongs to the exponential family and the estimation of  $b$  is straightforward: the sufficient statistics sequence  $(s_{3,k})$  is defined by

$$s_{3,k} = s_{3,k-1} + \gamma_k \left( \sum_{i,j} \left( \frac{y_{ij} - f(x_{ij}, \psi_i^{(k)})}{f^c(x_{ij}, \psi_i^{(k)})} \right)^2 - s_{3,k-1} \right)$$

and  $b_{k+1} = \sqrt{s_{3,k}/N_{tot}}$ .

- For a general residual variance model  $g = a + b f^c$ , the complete model does not belong to the exponential family and the estimates of the residual variance parameters  $(a, b, c)$  cannot be expressed as a function of some sufficient statistics. Then, let  $(A_k, B_k, C_k)$  that minimize the complete log-likelihood:

$$(A_k, B_k, C_k) = \text{Arg min}_{(a,b,c)} \left\{ \sum_{i,j} \log(a + b f^c(x_{ij}, \psi_i^{(k)})) + \frac{1}{2} \sum_{i,j} \left( \frac{y_{ij} - f(x_{ij}, \psi_i^{(k)})}{a + b f^c(x_{ij}, \psi_i^{(k)})} \right)^2 \right\}$$

We update the residual variance parameters as follows:

$$a_{k+1} = a_k + \gamma_k (A_k - a_k) \quad (1.8)$$

$$b_{k+1} = b_k + \gamma_k (B_k - b_k) \quad (1.9)$$

$$c_{k+1} = c_k + \gamma_k (C_k - c_k) \quad (1.10)$$

The estimation of  $\mu$  and  $\Omega$  remains unchanged.

### 1.1.2 The MCMC-SAEM algorithm

For model (1.1), the simulation step cannot be directly performed. Kuhn and Lavielle [5] propose to combine the SAEM algorithm with a MCMC (Markov Chain Monte Carlo) procedure. This procedure consists in replacing the Simulation-step at iteration  $k$  by  $m$  iterations of the Hastings-Metropolis algorithm.

Here, we will consider the Gaussian parameters  $(\varphi_i)$ . For  $i = 1, 2, \dots, N$

- let  $\varphi_{i,0} = \varphi_i^{(k-1)}$
- for  $p = 1, 2, \dots, m$ ,
  1. draw  $\tilde{\varphi}_{i,p}$  using the proposal kernel  $q_{\theta_k}(\varphi_{i,p-1}, \cdot)$
  2. set  $\varphi_{i,p} = \tilde{\varphi}_{i,p}$  with probability

$$\alpha(\varphi_{i,p-1}, \tilde{\varphi}_{i,p}) = \min \left( 1, \frac{p(\tilde{\varphi}_{i,p} | y_i; \theta_k) q_{\theta_k}(\tilde{\varphi}_{i,p}, \varphi_{i,p-1})}{p(\varphi_{i,p-1} | y_i; \theta_k) q_{\theta_k}(\varphi_{i,p-1}, \tilde{\varphi}_{i,p})} \right)$$

and  $\varphi_{i,p} = \varphi_{i,p-1}$  with probability  $1 - \alpha(\varphi_{i,p-1}, \tilde{\varphi}_{i,p})$ .

- let  $\varphi_i^{(k)} = \varphi_{i,m}$ .

Several transition kernels, associated to different proposals can be successively used. We use the four following proposal kernels:

1.  $q_{\theta_k}^{(1)}$  is the prior distribution of  $\varphi_i$  at iteration  $k$ , that is the Gaussian distribution  $\mathcal{N}(C_i\mu_k, \Omega_k)$  and then

$$\alpha(\varphi_{i,p-1}, \tilde{\varphi}_{i,p}) = \min \left( 1, \frac{p(y_i | \tilde{\varphi}_{i,p}; \theta_k)}{p(y_i | \varphi_{i,p-1}; \theta_k)} \right)$$

2.  $q_{\theta_k}^{(2)}$  is a random permutation of the  $\varphi_i$ : generate a random permutation  $\sigma$  of  $\{1, 2, \dots, N\}$  and set  $\tilde{\varphi}_{i,p} = \varphi_{\sigma(i),p-1}$ . This kernel is not used anymore (deprecated).
3.  $q_{\theta_k}^{(3)}$  is a succession of  $d$  unidimensional Gaussian random walks: each component of  $\varphi_i$  are successively updated.
4.  $q_{\theta_k}^{(4)}$  is a multidimensional random walk  $\mathcal{N}(\varphi_{i,p-1}, \kappa\Omega_k)$ . The dimension changes for each iteration of the algorithm between 2 and the dimension of  $\varphi_i$  cyclicly, and it iterates and updates consecutive subvector of  $\varphi_i$ . This kernel is symmetric and then

$$\alpha(\varphi_{i,p-1}, \tilde{\varphi}_{i,p}) = \min \left( 1, \frac{p(y_i, \tilde{\varphi}_{i,p}; \theta_k)}{p(y_i, \varphi_{i,p-1}; \theta_k)} \right)$$

Then, the simulation-step at iteration  $k$  consists in running  $m_1$  iterations of the Hasting-Metropolis with proposal  $q_{\theta_k}^{(1)}$ ,  $m_2$  iterations with proposal  $q_{\theta_k}^{(2)}$ ,  $m_3$  iterations with proposal  $q_{\theta_k}^{(3)}$  and  $m_4$  iterations with proposal  $q_{\theta_k}^{(4)}$ .

**Remark 1 :** During the first  $K_b$  iterations (“burning” iterations) of SAEM, we only run the MCMC algorithm but the parameters are not updated.

**Remark 2 :** When the number  $N$  of subjects is small, convergence of the algorithm can be improved by running  $L$  Markov Chain instead of only one. The simulation step requires to draw  $L$  sequences  $\varphi^{(k,1)}, \dots, \varphi^{(k,L)}$  at iteration  $k$  and to combine stochastic approximation and Monte Carlo in the approximation step:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left( \frac{1}{L} \sum_{\ell=1}^L \log p(y, \varphi^{(k,\ell)}; \theta) - Q_{k-1}(\theta) \right) \quad (1.11)$$

### 1.1.3 The Simulated Annealing SAEM algorithm

Convergence of SAEM can strongly depend on the initial guess if the likelihood  $\ell$  possesses several local maxima. The Simulated Annealing version of SAEM improves the convergence of the algorithm toward the global maximum of  $\ell$ .

For the sake of simplicity, we will consider here a constant residual error model  $g = a$ . Let

$$U(y, \varphi; \theta) = \frac{1}{2a^2} \sum_{i,j} (y_{ij} - f(x_{ij}, h(\varphi_i)))^2 + \frac{1}{2} \sum_{i=1}^N (\varphi_i - C_i\mu)' \Omega^{-1} (\varphi_i - C_i\mu)$$



Then, we can write the complete likelihood:

$$p(y, \varphi; \theta) = C(\theta) e^{-U(y, \varphi; \theta)}$$

where  $C(\theta)$  is a normalizing constant that only depends on  $\theta$ .

For any *temperature*  $T \geq 0$ , we consider the complete model

$$p_T(y, \varphi; \theta) = C_T(\theta) e^{-\frac{1}{T}U(y, \varphi; \theta)}$$

where  $C_T(\theta)$  is a normalizing constant. This model consists in replacing the variance matrix  $\Omega$  by  $T\Omega$  and the residual variance  $a^2$  by  $Ta^2$ . In other words, a model “with a large temperature” is a model with large variances.

We introduce a decreasing temperature sequence  $(T_k, 1 \leq k \leq K)$  and use the MCMC-SAEM algorithm considering the complete model  $p_{T_k}(y, \varphi; \theta)$  at iteration  $k$  (while the usual version of MCMC-SAEM uses  $T_k = 1$  at each iteration). The sequence  $(T_k)$  is large during the first iterations and decreases to 1 with exponential rate. This is done by choosing large initial variances  $\Omega_0$  and  $a_0^2$  and setting

$$\tilde{\Omega}_{k+1} = \frac{1}{N} \left( s_{2,k} - \sum_{i=1}^N (C_i \mu_{k+1}) s'_{1,i,k} - \sum_{i=1}^N s_{1,i,k} (C_i \mu_{k+1})' + \sum_{i=1}^N (C_i \mu_{k+1})(C_i \mu_{k+1})' \right) \quad (1.12)$$

$$a_{k+1} = \sqrt{\frac{s_{3,k}}{N_{tot}}} \quad (1.13)$$

$$\Omega_{k+1} = \max \left( \tau \Omega_k, \tilde{\Omega}_{k+1} \right) \quad (1.14)$$

$$a_{k+1}^2 = \max \left( \tau a_k^2, \frac{s_{3,k}}{N} \right) \quad (1.15)$$

during the first iterations of the algorithm and where  $0 \leq \tau \leq 1$ .

These large values of the variances make the conditional distribution  $p(\phi|y; \theta)$  less concentrated around its mode. This procedure allows the sequence  $(\theta_k)$  to escape from the local maxima of the likelihood and to converge to a neighborhood of the global maximum of  $\ell$ . After that, the usual MCMC-SAEM algorithm is used, estimating the variances at each iteration.

**Remark 1:** The Simulated Annealing version of SAEM is performed during the first  $K_{sa}$  iterations. Of course, SAEM without any simulated annealing can be run by setting  $\tau = 0$ . On the other hand, simulated annealing is obtained with  $\tau$  close to 1.

**Remark 2:** We can use two different coefficients  $\tau_1$  and  $\tau_2$  for  $\Omega$  and  $a^2$  in MONOLIX. It is possible, for example, to choose  $\tau_1 < 1$  and  $\tau_2 > 1$ , with a small initial residual variance and large initial inter-subject variances. In this case, SAEM tries to obtain the best possible fit during the first iterations, allowing a large inter-subject variability. During the next iterations, this variability is reduced and the residual variance increases until reaching the best possible trade-off between these two criteria.

## 1.2 Some extensions

### 1.2.1 Model with censored data

The statistical models are described in [Section 2.5](#).

The maximum likelihood estimation is based on the log-likelihood function  $L(y^{obs}; \theta)$  of the response  $y^{obs}$  with  $\theta = (\mu, \Omega, a, b, c)$  the vector of all the parameters of the model

$$L(y^{obs}; \theta) = \log \left( \prod_{i=1}^N \int p(y_i^{obs}, y_i^{cens}, \varphi_i; \theta) d\varphi_i dy_i^{cens} \right), \quad (1.16)$$

where  $p(y_i^{obs}, y_i^{cens}, \varphi_i; \theta)$  is the likelihood of the complete data  $(y_i^{obs}, y_i^{cens}, \varphi_i)$  of the  $i$ -th subject. The complete likelihood of the  $i$ -th subject is equal to:

$$p(y_i^{obs}, y_i^{cens}, \varphi_i; \theta) = \prod_{(i,j) \in I_{obs}} p(y_{ij}^{obs} | \varphi_i; \theta) p(\varphi_i; \theta) \prod_{(i,j) \in I_{cens}} p(y_{ij}^{cens} | y_i^{obs}, \varphi_i; \theta) p(\varphi_i; \theta),$$

Samson *et al.* proposed in [\[9\]](#) an extension of the SAEM algorithm to handle left-censored data in NLMEM as an exact Maximum Likelihood estimation method. The simulation of the censored data with a truncated Gaussian distribution is included in the MCMC procedure. The convergence of this extended SAEM algorithm is proved under general conditions.

In this case,

$$\begin{aligned} p(y_{ij}^{obs} | \varphi_i; \theta) &= \pi(y_{ij}^{obs}; f(\varphi_i, t_{ij}), g^2(\varphi_i, t_{ij})), & \text{if } (i, j) \in I_{obs} \text{ and} \\ p(y_{ij}^{cens} | y_i^{obs}, \varphi_i; \theta) &= \pi(y_{ij}^{cens}; f(\varphi_i, t_{ij}), g^2(\varphi_i, t_{ij})) \mathbb{1}_{y_{ij} \leq LOQ}, & \text{if } (i, j) \in I_{cens}, \end{aligned}$$

where  $\pi(x; m, v)$  is the probability density function of the Gaussian distribution with mean  $m$  and variance  $v$ , evaluated at  $x$ .

Interval censored data are treated analogously. In this case, as explained in [Section 2.5.2](#), it is considered

$$p(y_{ij}^{cens} | y_i^{obs}, \varphi_i; \theta) = \pi(y_{ij}^{cens}; f(\varphi_i, t_{ij}), g^2(\varphi_i, t_{ij})) \mathbb{1}_{y_{ij} \in (LOD, LOQ)}$$

where  $(LOD, LOQ)$  is the censoring interval.

### 1.2.2 Estimation with a prior distribution

It can be incorporated prior distributions on the fixed effect parameters  $\mu$  to be handled by the SAEM algorithm.

The parameter  $\mu$  is considered as a random Gaussian variable. Let us denote  $\mu_\star$  the mean of this prior distribution and  $V_\star$  its diagonal variance matrix:

$$\mu \sim \mathcal{N}(\mu_\star, V_\star) \quad (1.17)$$

The parameters  $\mu_\star$  and  $V_\star$  are fixed.

Let  $\theta_\star = (\mu_\star, V_\star, \Omega, \xi)$  and  $d_\star$  be the length of  $\mu_\star$ . Then, the complete log-likelihood of  $(y, \psi, \mu)$  can be written

$$\begin{aligned} \log p(y, \varphi, \mu; \theta_\star) = & - \sum_{i,j} \log(g(x_{ij}, h(\varphi_i); \xi)) - \frac{1}{2} \sum_{i,j} \left( \frac{y_{ij} - f(x_{ij}, h(\varphi_i))}{g(x_{ij}, h(\varphi_i); \xi)} \right)^2 \\ & - \frac{N}{2} \log(|\Omega|) - \frac{1}{2} \sum_{i=1}^N (\varphi_i - C_i \mu)' \Omega^{-1} (\varphi_i - C_i \mu) \\ & - \frac{1}{2} \log(|V_\star|) - \frac{1}{2} (\mu - \mu_\star)' V_\star^{-1} (\mu - \mu_\star) - \frac{N_{tot} + Nd + d_\star}{2} \log(2\pi) \end{aligned}$$

SAEM allows two estimation method for parameters with prior distributions:

- maximum a posteriori or MAP:

The simulation step and the approximation step are similar to the ones of [Section 1.1.1](#). Then,  $\mu_{k+1}$  is obtained in the maximization step as follows:

$$\mu_{k+1} = \left( V_\star^{-1} + \sum_{i=1}^N C_i' \Omega_k^{-1} C_i \right)^{-1} \left( V_\star^{-1} \mu_\star + \sum_{i=1}^N C_i' \Omega_k^{-1} s_{1,i,k} \right)$$

while  $\Omega_{k+1}$  and  $\sigma_{k+1}^2$  are computed as before.

- posterior distribution:  $\mu$  is treated as unobserved variables so, the simulation step includes the simulation of  $\mu_{k+1}$  by using random walks as proposal kernels  $q_{\theta_\star}^{(1)}$  and  $q_{\theta_\star}^{(2)}$  (like kernel 3 and 4 for  $\varphi$  described above).

Then approximation and maximization steps remain the same, just that the simulated  $\mu_{k+1}$  are supposed as known (or fixed).

Of course, it is possible to combine several methods to estimate the complete set of parameters: we can fix some parameters, use maximum likelihood estimation for some other parameters and combine methods for parameters with prior information for the other ones.

Also, as for individual parameters, it is possible to define prior distributions as a transform of a Gaussian random variable

$$H^{-1}(\mu) \sim \mathcal{N}(H^{-1}(\mu_\star), V_\star)$$

where  $H$  is a monotonically increasing function defined in  $\mathbb{R}$ . In this case,  $\mu_\star$  is the typical value of the corresponding  $\mu$ .

**Remark:** M.A.P estimators are only possible for Gaussian prior distributions on covariate coefficients. For the intercepts, M.A.P is only available when the corresponding individual parameter holds  $\varphi = H^{-1}(\psi) \sim \mathcal{N}$ .

### 1.2.3 Modeling the inter-occasion variability

Mixed effects models with IOV are described in [Section 2.6](#).

An extension of the SAEM algorithm for models with two levels of random effects can be found in [\[7\]](#). The methodology proposed in this paper is limited to only two periods and assumes IOV on each parameter ( $\Gamma$  is a diagonal matrix with non zero element on the diagonal). We have extended this methodology to any number of occasions and also to any structure of the IOV covariance matrix  $\Gamma$ . MONOLIX 4.2.1 includes also the extension to any number of levels of IOV.

### 1.2.4 Mixture models and mixture of models

Mixture models and model of mixtures are described in [Section 2.8](#).

#### Mixture models

The mixture models are modeled by using latent covariates, it means, non-observed covariates  $L_i$  having probabilities  $p(L_i = m) = \pi_{im}$ .

In this case the complete log-likelihood can be written as

$$\log p(y, \psi; \theta) = \sum_i \log \left( \sum_m \pi_{im} p(y_i, \psi_i | L_i = m; \theta) \right)$$

it means that  $\psi$  follows a mixture model, so the simulation step must take this into account when simulating the individual parameters.

If we put together  $L_{im}$  with the other covariates  $C_i$  into the matrix  $C_{im}$ , then  $\theta_{k+1}$  can be

obtained in the maximization step as follows:

$$\mu_{k+1} = \left( \sum_{i,m} \pi_{im} C_{im}' \Omega_k^{-1} C_{im} \right)^{-1} \sum_{i,m} \pi_{im} C_{im}' \Omega_k^{-1} s_{1,i,k} \quad (1.18)$$

$$\Omega_{k+1} = \frac{1}{N} \left( s_{2,k} - \sum_{i,m} \pi_{im} (C_{im} \mu_{k+1}) s'_{1,i,k} - \sum_{i,m} \pi_{im} s_{1,i,k} (C_{im} \mu_{k+1})' \right. \quad (1.19)$$

$$\left. + \sum_{i,m} \pi_{im} (C_{im} \mu_{k+1}) (C_{im} \mu_{k+1})' \right) \quad (1.20)$$

### Mixture of models

Let us denote by  $p_{im}(\psi_i)$  the proportions for within subject mixture models (WSMM) and  $\pi_{im}(\psi_i)$  the probabilities for between subject mixture models (BSMM), for subject  $i$  and group  $m$  knowing the individual parameters  $\psi_i$ . Let  $f_m$  be the regression function for group  $m$  and  $g_m$  the standard deviation of the corresponding error model.

For WSMM, supposing for example that there are 2 groups, we have

$$f(x_{ij}, \psi_i) = p_{i1}(\psi_i) f_1(x_{ij}, \psi_i) + p_{i2}(\psi_i) f_2(x_{ij}, \psi_i)$$

and so

$$\log p(y|\psi; \theta) = - \sum_{i,j} \log(g(x_{ij}, \psi_i, \xi)) - \frac{1}{2} \sum_{i,j} \left( \frac{y_{ij} - f(x_{ij}, \psi_i)}{g(x_{ij}, \psi_i, \xi)} \right)^2 - \frac{N_{tot}}{2} \log(2\pi)$$

can be computed as before.

For BSMM, we have that

$$\log p(y|\psi; \theta) = \sum_i \log \left( \sum_m \pi_{im}(\psi_i) p(y_i|\psi_i, G_i = m; \theta) \right)$$

where  $p(y_i|\psi_i, G_i = m; \theta)$  are the probabilities for subject  $i$  to be in group  $m$ . They satisfy then

$$\log p(y_i|\psi_i, G_i = m; \theta) = - \sum_j \log(g_m(x_{ij}, \psi_i, \xi)) - \frac{1}{2} \sum_j \left( \frac{y_{ij} - f_m(x_{ij}, \psi_i)}{g_m(x_{ij}, \psi_i, \xi)} \right)^2 - \frac{n_i}{2} \log(2\pi)$$

with  $n_i$  representing the number of observations of subject  $i$ .

That means that the approximation step must be adapted in both cases, but the maximization step must be changed only for the estimations of the residual error model parameters  $\xi$ .

## 1.3 Estimation of the Fisher Information matrix

Let  $\theta^*$  be the true unknown value of  $\theta$ , and let  $\hat{\theta}$  be the maximum likelihood estimate of  $\theta$ . If the observed likelihood function  $\ell$  is sufficiently smooth, asymptotic theory for maximum-likelihood estimation holds and

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{N \rightarrow \infty} \mathcal{N}(0, I(\theta^*)^{-1}) \quad (1.21)$$

where  $I(\theta^*) = -\partial_{\theta}^2 \log \ell(y; \theta^*)$  is the true Fisher information matrix. Thus, an estimate of the asymptotic covariance of  $\hat{\theta}$  is the inverse of the Fisher information matrix  $I(\hat{\theta}) = -\partial_{\theta}^2 \log \ell(y; \hat{\theta})$ .

### 1.3.1 Linearization of the model

The Fisher information matrix of the nonlinear mixed effects model defined in (1) cannot be computed in a closed-form.

An alternative is to approximate this information matrix by the Fisher information matrix of the Gaussian model deduced from the nonlinear mixed effects model after linearization of the function  $f$  around the conditional expectation of the individual Gaussian parameters  $(\mathbb{E}(\phi_i|y; \hat{\theta}), 1 \leq i \leq N)$ . The Fisher information matrix of this Gaussian model is a block matrix (no correlations between the estimated fixed effects and the estimated variances). The gradient of  $f$  is numerically computed.

**Remark 1:** We do not recommend the linearization of the model to estimate the parameters of the model, as it is done with the FO and FOCE algorithms. On the other hand, many numerical experiments have shown that this approach can be used to estimate the Fisher information matrix.

**Remark 2:** Obviously, this approach cannot be used with discrete data models nor mixture models ...

### 1.3.2 A stochastic approximation of the Fisher Information Matrix

It is possible to obtain an estimation of the Fisher information matrix using the Louis's missing information principle [6]:

$$\partial_{\theta}^2 \log \ell(y; \theta) = \mathbb{E}(\partial_{\theta}^2 \log p(y, \varphi; \theta)|y; \theta) + \text{Cov}(\partial_{\theta} \log p(y, \varphi; \theta)|y; \theta) \quad (1.22)$$

where

$$\begin{aligned} \text{Cov}(\partial_\theta \log p(y, \varphi; \theta) | y; \theta) &= \text{E}(\partial_\theta \log p(y, \varphi; \theta) \partial_\theta \log p(y, \varphi; \theta)' | y; \theta) \\ &\quad - \text{E}(\partial_\theta \log p(y, \varphi; \theta) | y; \theta) \text{E}(\partial_\theta \log p(y, \varphi; \theta) | y; \theta)' \end{aligned}$$

and

$$\partial_\theta \log g(y; \theta) = \text{E}(\partial_\theta \log p(y, \varphi; \theta) | y; \theta)$$

Here,  $\partial_\theta u$  is the gradient of  $u$  (*i.e.* the vector of first derivatives of  $u$  with respect to  $\theta$ ) and  $\partial_\theta^2 u$  is the hessian of  $u$  (*i.e.* the matrix of second derivatives of  $u$  with respect to  $\theta$ ).

Then, using SAEM, the matrix  $\partial_\theta^2 \log \ell(y; \hat{\theta})$  can be approximated by the sequence  $(H_k)$  defined as follows:

$$\begin{aligned} \Delta_k &= \Delta_{k-1} + \gamma_k (\partial_\theta \log f(y, \phi_k; \theta_k) - \Delta_{k-1}) \\ D_k &= D_{k-1} + \gamma_k (\partial_\theta^2 \log f(y, \phi_k; \theta_k) - D_{k-1}) \\ G_k &= G_{k-1} + \gamma_k (\partial_\theta \log f(y, \phi_k; \theta_k) \partial_\theta \log f(y, \phi_k; \theta_k)^t - G_{k-1}) \\ H_k &= D_k + G_k - \Delta_k \Delta_k^t \end{aligned}$$

## 1.4 Estimation of the individual parameters

When the parameters of the model have been estimated, we can estimate the individual parameters  $(\psi_i)$ . To do that, we will estimate the individual normally distributed parameters  $(\varphi_i)$  and derive the estimates of  $(\psi_i)$  using the transformation  $\psi_i = h(\psi_i)$ .

Let  $\hat{\theta}$  be the estimated value of  $\theta$  computed with the SAEM algorithm and let  $p(\varphi_i | y_i; \hat{\theta})$  be the conditional distribution of  $\varphi_i$  for  $1 \leq i \leq N$ .

We use the MCMC procedure used in the SAEM algorithm to estimate these conditional distributions. More precisely, for  $1 \leq i \leq N$ , we empirically estimate:

- the conditional mode (or Maximum A Posteriori)  $m(\varphi_i | y_i; \hat{\theta}) = \text{Arg max}_{\varphi_i} p(\varphi_i | y_i; \hat{\theta})$ ,
- the conditional mean  $E(\varphi_i | y_i; \hat{\theta})$ ,
- the conditional standard deviation  $sd(\varphi_i | y_i; \hat{\theta})$ .

### Remarks:

1. The prior distribution of  $\varphi_i$  is a normal distribution, but not the conditional distribution  $p(\varphi_i | y_i; \hat{\theta})$  (remember that the structural model is not a linear function of  $\varphi_i \dots$ ). Then, the conditional mode  $m(\varphi_i | y_i; \hat{\theta})$  and the conditional expectation  $E(\varphi_i | y_i; \hat{\theta})$  are two different predictors of  $\varphi_i$ .

2. If the transformation  $h$  is not linear,

$$\begin{aligned}\mathbb{E}(\psi_i|y_i;\hat{\theta}) &= \mathbb{E}(h(\varphi_i|y_i;\hat{\theta})) \\ &\neq h(\mathbb{E}(\varphi_i|y_i;\hat{\theta}))\end{aligned}$$

In MONOLIX, we estimate  $\mathbb{E}(\varphi_i|y_i;\hat{\theta})$  and  $\mathbb{E}(\psi_i|y_i;\hat{\theta})$ .

The number of iterations of the MCMC algorithm used to estimate the conditional mean and standard deviation is adaptively chosen as follows:

1. The  $(\varphi_i)$  are initialized with the last value obtained in SAEM
2. We run the Hastings-Metropolis with kernel  $q^{(1)}$ ,  $q^{(3)}$  and  $q^{(4)}$  and compute at each iteration the empirical conditional mean and s.d. of  $\varphi_i$ :

$$e_{i,K} = \frac{1}{K} \sum_{k=1}^K \varphi_{i,k} \quad (1.23)$$

$$sd_{i,K} = \sqrt{\frac{1}{K} \sum_{k=1}^K \varphi_{i,k}^2 - e_{i,K}^2} \quad (1.24)$$

where  $\varphi_{i,k}$  is the value of  $\varphi_i$  at iteration  $k$  of the MCMC algorithm.

3. we stop the algorithm at iteration  $K$  and use  $e_{i,K}$  and  $sd_{i,K}$  to estimate the conditional mean and s.d. of  $\varphi_i$  if, for any  $K - L_{mcmc} + 1 \leq k \leq K$ ,

$$\begin{aligned}(1 - \rho_{mcmc})\bar{e}_K &\leq \bar{e}_k \leq (1 + \rho_{mcmc})\bar{e}_K \\ (1 - \rho_{mcmc})\bar{sd}_K &\leq \bar{sd}_k \leq (1 + \rho_{mcmc})\bar{sd}_K\end{aligned} \quad (1.25)$$

where  $0 < \rho_{mcmc} < 1$ . That means that the sequence of empirical means and s.d. must stay in a  $\rho_{mcmc}$ -confidence interval during  $L_{mcmc}$  iterations.

## 1.5 Estimation of the likelihood

### 1.5.1 Linearization of the model

The likelihood of the nonlinear mixed effects model defined in (1) cannot be computed in a closed-form.

An alternative is to approximate this likelihood by the likelihood of the Gaussian model deduced from the nonlinear mixed effects model after linearization of the function  $f$  around the predictions of the individual parameters  $(\varphi_i, 1 \leq i \leq N)$ .



### 1.5.2 Estimation using importance sampling

The likelihood of the observations can be estimated without any approximation using a Monte-Carlo approach. The likelihood  $\ell$  of the observations can be decomposed as follows

$$\begin{aligned}\ell(y; \theta) &= \int p(y, \varphi; \theta) d\varphi \\ &= \int h(y|\varphi; \theta) \pi(\varphi; \theta) d\varphi\end{aligned}$$

where  $\pi$  is the so-called *prior distribution* of  $\varphi$ . According to (2.2),  $\pi$  is a Gaussian distribution.

For any distribution  $\tilde{\pi}$  absolutely continuous with respect to the prior distribution  $\pi$ , we can write

$$\ell(y; \theta) = \int h(y|\varphi; \theta) \frac{\pi(\varphi; \theta)}{\tilde{\pi}(\varphi; \theta)} \tilde{\pi}(\varphi; \theta) d\varphi$$

Then,  $\ell(y; \theta)$  can be approximated via an *Importance Sampling* integration method:

1. draw  $\varphi^{(1)}, \varphi^{(2)}, \dots, \varphi^{(M)}$  with the distribution  $\tilde{\pi}(\cdot; \theta)$ ,
2. let

$$\ell_M(y; \theta) = \frac{1}{M} \sum_{j=1}^M h(y|\varphi^{(j)}; \theta) \frac{\pi(\varphi^{(j)}; \theta)}{\tilde{\pi}(\varphi^{(j)}; \theta)} \quad (1.26)$$

The statistical properties of the estimator  $\ell_M(y; \theta)$  of the likelihood  $\ell(y; \theta)$  strongly depend on the sampling distribution  $\tilde{\pi}$ . First, note that

$$\begin{aligned}\mathbb{E}(\ell_M(y; \theta)) &= \ell(y; \theta), \\ \text{Var}(\ell_M(y; \theta)) &= \mathcal{O}(1/M).\end{aligned}$$

Furthermore, if  $\tilde{\pi}$  is the conditional distribution  $p(\phi|y; \theta)$ , the variance of the estimator is null and  $\hat{\ell}_M(y; \theta) = \ell(y; \theta)$  for any value of  $M$ . That means that an accurate estimation of  $\ell(y; \theta)$  can be obtained with a small value of  $M$  if the sampling distribution is close to the conditional distribution  $p(\phi|y; \theta)$ .

In MONOLIX, for  $i = 1, 2, \dots, N$ , we empirically estimate the conditional mean  $\mathbb{E}(\varphi_i|y_i; \hat{\theta})$  and the conditional variance  $\text{Var}(\varphi_i|y_i; \hat{\theta})$  of  $\varphi_i$  as described above. Then, the  $\varphi_i^{(j)}$  are drawn with the sampling distribution  $\tilde{\pi}$  as follows:

$$\varphi_i^{(j)} = \mathbb{E}(\varphi_i|y_i; \hat{\theta}) + \text{Var}(\varphi_i|y_i; \hat{\theta})^{\frac{1}{2}} \times T_{ij}$$

where  $(T_{ij})$  is a sequence of *i.i.d.* random variables distributed with a *t*-distribution with  $\nu$  degrees of freedom.

It is possible to use the default value  $\nu = 5$ . It is also possible to automatically test different d.f in  $\{2, 5, 10, 20\}$  and to select the one that provides the smallest empirical variance for  $\ell_M(y; \theta)$ .

## 1.6 Estimation of the weighted residuals

### 1.6.1 The Population Weighted Residuals

The Population Weighted Residuals are evaluated as

$$PWRES_{ij} = \frac{y_{ij} - \hat{y}_{ij}^{pop}}{\hat{\sigma}_{ij}^{pop}}$$

where  $\hat{y}_{ij}^{pop}$  is the population prediction of  $y_{ij}$  and  $(\hat{\sigma}_{ij}^{pop})^2 = Var_{\hat{\theta}}(y_{ij})$  is the variance of  $y_{ij}$ .

Two population predictions are proposed in MONOLIX :

1. the pop. param. prediction  $f(x_{ij}; h(\mathbb{E}_{\hat{\theta}}(\varphi_i))) = f(x_{ij}; h(C_i \hat{\mu}))$
2. the pop. mean prediction  $\mathbb{E}_{\hat{\theta}}(f(x_{ij}; \psi_i)) = \mathbb{E}_{\hat{\theta}}(f(x_{ij}; h(\varphi_i)))$ .

Here,  $\mathbb{E}_{\hat{\theta}}(f(x_{ij}; h(\varphi_i)))$  and  $Var_{\hat{\theta}}(y_{ij})$  are estimated with a Monte-Carlo procedure.

### 1.6.2 The Individual Weighted Residuals

The Individual Weighted Residuals are evaluated as

$$IWRES_{ij} = \frac{y_{ij} - \hat{y}_{ij}^{ind}}{\hat{\sigma}_{ij}^{ind}}$$

where  $\hat{y}_{ij}^{ind} = f(x_{ij}; \hat{\psi}_i)$  is the individual prediction of  $y_{ij}$  and  $(\hat{\sigma}_{ij}^{ind})^2 = g(x_{ij}; \hat{\psi}_i, \hat{\xi})^2$  is the residual variance of  $y_{ij}$ .

The  $\hat{\psi}_i$ 's are the individual estimates of the  $\psi_i$ 's described in [Section 1.4](#) (the conditional modes or the conditional means)

**Remark:** When a transformed residual error model is used (an exponential error model for instance), the weighted residuals are computed using  $t(y)$  instead of  $y$ .

### 1.6.3 The Normalized Prediction Distribution

The Normalized Prediction Distribution Errors are defined as follow

$$NPDE_{ij} = \Phi^{-1}(\hat{p}_{ij})$$

where  $\Phi$  is the  $\mathcal{N}(0, 1)$  cumulative distribution function and where  $\hat{p}_{ij}$  is an empirical estimator of

$$p_{ij} = \mathbb{P}(Y_{ij} < y_{ij})$$

obtained by Monte-Carlo.

## 1.7 Inputs and outputs

### 1.7.1 The inputs

To summarize, MONOLIX requires to define the model and to fix some parameters used for the algorithms. First, it is necessary to define:

- the structural model, that is the regression function  $f$  defined in (2.1),
- the covariate model, that is the structure of the matrix  $\mu$  defined in (2.2) and the covariates  $(c_i)$ .
- the variance-covariance model for the random effects, that is the structure of the variance-covariance matrix  $\Omega$  defined in (2.2).
- the residual variance model, that is the regression function  $g$ .

Then, it is necessary to specify several parameters for running the algorithms:

- the SAEM algorithm requires to specify
  - the initial values of the fixed effects  $\mu_0$ , the initial variance-covariance matrix  $\Omega_0$  of the random effects and the initial residual variance coefficients  $a_0$ ,  $b_0$  and  $c_0$ ,
  - the sequence of step sizes  $(\gamma_k)$ , that is the numbers of iterations  $(K_1, K_2)$  and the coefficients  $(a_1, a_2)$  defined in (1.5) and (1.6),
  - the number of burning iterations  $K_b$  used with the same value  $\theta_0$  before updating the sequence  $(\theta_k)$ .
- the MCMC algorithm requires to set
  - the number of Markov Chains  $L$ ,
  - the numbers  $m_1$ ,  $m_2$ ,  $m_3$  and  $m_4$  of iterations of the Hasting-Metropolis algorithm,
  - the probability of acceptance  $\rho$  for kernel  $q^{(3)}$  and  $q^{(4)}$ ,
- the algorithm to estimate the conditional distribution of the  $(\varphi_i)$  requires to set
  - the width of the confidence interval  $\rho_{mcmc}$  (see (1.25)),
  - the number of iterations  $L_{mcmc}$ .
- the Simulated Annealing algorithm requires to set
  - the coefficient  $\tau_1$  and  $\tau_2$  defining the decrease of the temperature (see (1.14,1.15))
  - the number of iterations  $K_{sa}$ .
- the Importance Sampling algorithm requires to set
  - the Monte Carlo number  $M$  used to estimate the observed likelihood (see (1.26)).

### 1.7.2 The outputs

#### a) Estimation of the parameters:

The SAEM algorithm computes the maximum likelihood estimate  $\hat{\theta}$  and estimates, the MAP for the corresponding parameters with priors. It computes also a rough estimation of the conditional expectation and s.d. of the individual parameters.

Recall that  $d$  is the number of individual parameters, then for  $j = 1, 2 \dots d$ , we estimate the vector of fixed effects  $\mu$  (intercept and coefficients of the covariates) by  $(\hat{\mu})$ .

Let  $\Omega = (\omega_{jl}, 1 \leq j, l \leq d)$ . Then, we estimate  $\omega_{jl}$  by  $\hat{\omega}_{jl}$ , for all  $1 \leq j, l \leq d$ . The residual error model parameters  $\xi$  are also estimated by  $\hat{\xi}$ .

#### a) Estimation of the Fisher information matrix:

It computes the estimators covariance matrix  $I(\hat{\theta})^{-1}/N$  defined in [Section 1.3](#).

With it we can

1. estimate the standard errors of  $\mu$ ,
2. test if some components of  $\mu$  are null by computing the significance level of the Wald test.
3. estimate the standard error of  $\hat{\omega}_{jl}$ , for all  $1 \leq j, l \leq d$ .
4. estimate the standard errors of  $\hat{\xi}$ .

#### b) Estimation of the conditional distributions:

The MCMC algorithm provides an estimation of the conditional means, conditional modes and conditional standard deviations of the individual parameters and of the random effects.

It allows also to simulate the population parameters with priors that where chosen to use the posterior distribution estimator.

#### c) Estimation of the likelihood:

The Importance Sampling algorithm computes an estimate  $\ell_M(y; \hat{\theta})$  of the observed likelihood together with its standard error.

Also, the individual contribution to the total log-likelihood is computed.

#### d) Hypothesis testing and model selection:

We can test the covariate model, the covariance model and the residual error model.

The AIC and BIC criteria are defined by

$$AIC = -2 \log \ell_M(y; \hat{\theta}) + 2P \quad (1.27)$$

$$BIC = -2 \log \ell_M(y; \hat{\theta}) + \log(N)P \quad (1.28)$$

where  $P$  is the total number of parameters to be estimated and  $N$  is the number of subjects.

When comparing two nested models  $\mathcal{M}_0$  and  $\mathcal{M}_1$  with dimensions  $P_0$  and  $P_1$  (with  $P_1 > P_0$ ), the Likelihood Ratio Test uses the test statistic

$$LRT = 2(\log \ell_{M,1}(y; \hat{\theta}_1) - \log \ell_{M,0}(y; \hat{\theta}_0))$$

According to the hypotheses to test, the limiting distribution of  $LRT$  under the null hypothesis is either a  $\chi^2$  distribution, or a mixture of a  $\chi^2$  distribution and a  $\delta - Dirac$  distribution. For example:

- to test whether some fixed effects are null, assuming the same covariance structure of the random effects, one should use

$$LRT \xrightarrow[N \rightarrow \infty]{} \chi^2(P_1 - P_0)$$

- to test whether some correlations of the covariance matrix  $\Omega$  are null, assuming the same covariate model, one should use

$$LRT \xrightarrow[N \rightarrow \infty]{} \chi^2(P_1 - P_0)$$

- to test whether the variance of one of the random effects is zero, assuming the same covariate model, one should use

$$LRT \xrightarrow[N \rightarrow \infty]{} \frac{1}{2} \chi^2(1) + \frac{1}{2} \delta_0$$

#### e) Estimation of the weighted residuals:

The Population Weighted Residuals ( $PWRES_{ij}$ ), the Individual Weighted Residuals ( $IWRES_{ij}$ ) and the Normalized Prediction Distribution Errors ( $NPDE_{ij}$ ) are computed as described [Section 1.6](#)

## Chapter 2

# Statistical models

### 2.1 The nonlinear mixed effects model

Detailed and complete presentations of the nonlinear mixed effects model can be found in [1, 2, 8]. See also the many references therein.

We consider the following general nonlinear mixed effects model for continuous outputs:

$$y_{ij} = f(x_{ij}, \psi_i) + g(x_{ij}, \psi_i, \xi)\varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \quad (2.1)$$

Here,

- $y_{ij} \in \mathbb{R}$  is the  $j$ th observation of subject  $i$ ,
- $N$  is the number of subjects,
- $n_i$  is the number of observations of subject  $i$ ,
- the regression variables, or design variables,  $(x_{ij})$  are assumed to be known,  $x_{ij} \in \mathbb{R}^{n_x}$ ,
- for subject  $i$ , the vector  $\psi_i = (\psi_{i,\ell}; 1 \leq \ell \leq n_\psi) \in \mathbb{R}^{n_\psi}$  is a vector of  $n_\psi$  individual parameters:

$$\psi_i = H(\mu, c_i, \eta_i) \quad (2.2)$$

where

- $c_i = (c_{im}; 1 \leq m \leq M)$  is a known vector of  $M$  covariates,
- $\mu$  is an unknown vector of fixed effects of size  $n_\mu$ ,
- $\eta_i$  is an unknown vector of normally distributed random effects of size  $n_\eta$ :

$$\eta_i \sim_{i.i.d.} \mathcal{N}(0, \Omega)$$

- the residual errors  $(\varepsilon_{ij})$  are random variables with mean zero and variance 1,
- the residual error model is defined by the function  $g$  and some parameters  $\xi$ .

Here, the parameters of the model are  $\theta = (\mu, \Omega, \xi)$ . We will denote  $\ell(y; \theta)$  the likelihood of the observations  $y = (y_{ij}; 1 \leq i \leq n, 1 \leq j \leq n_i)$  and  $p(y, \psi; \theta)$  the likelihood of the complete data  $(y, \psi) = (y_{ij}, \psi_i; 1 \leq i \leq n, 1 \leq j \leq n_i)$ . Thus,

$$\ell(y; \theta) = \int p(y, \psi; \theta) d\psi.$$

Let us see now the statistical model used in MONOLIX 4.2.1 more in details.

## 2.2 The statistical model for the individual parameters

In MONOLIX 4.2.1, we assume that  $\psi_i$  is a transformation of a Gaussian random vector  $\varphi_i$ :

$$\psi_i = h(\varphi_i) \quad (2.3)$$

where, by rearranging the covariates  $(c_{im})$  into a matrix  $C_i$ ,  $\varphi_i$  can be written as

$$\varphi_i = C_i \mu + \eta_i \quad (2.4)$$

### 2.2.1 Examples of transformations

Here, different transformations  $(h_\ell)$  can be used for the different components of  $\psi_i = (\psi_{i,\ell})$  where  $\psi_{i,\ell} = h_\ell(\varphi_{i,\ell})$  for  $\ell = 1, 2, \dots, \ell$ . Let us denote by  $\Phi(u)$  the cumulative distribution function of a Gaussian distributed random variable.

- $\psi_{i,\ell}$  has a log-normal distribution if  $h_\ell(u) = e^u$ ,
- assuming that  $\psi_{i,\ell}$  takes its values in  $(0, 1)$ , we can use a logit transformation  $h_\ell(u) = 1/(1 + e^{-u})$ , or a probit transformation  $h_\ell(u) = \Phi(u)$ .
- assuming that  $\psi_{i,\ell}$  takes its values in  $(A, B)$ , we can define  $h_\ell(u) = A + (B - A)/(1 + e^{-u})$ , or  $h_\ell(u) = A + (B - A)\Phi(u)$ .

In the following, we will use either the parameters  $\psi_i$  or the Gaussian transformed parameters  $\varphi_i = h^{-1}(\psi_i)$ .

The model can address continuous and/or categorical covariates.

### 2.2.2 Example of continuous covariate model

Consider a PK model that depends on volume and clearance and consider the following covariate model for these two parameters:

$$\begin{aligned} CL_i &= CL_{\text{pop}} \left( \frac{W_i}{W_{\text{pop}}} \right)^{\beta_{CL,W}} \left( \frac{A_i}{A_{\text{pop}}} \right)^{\beta_{CL,A}} e^{\eta_{i,1}} \\ V_i &= V_{\text{pop}} \left( \frac{W_i}{W_{\text{pop}}} \right)^{\beta_{V,W}} e^{\eta_{i,2}} \end{aligned}$$

Where  $W_i$  and  $A_i$  are the weight and the age of subject  $i$  and where  $W_{\text{pop}}$  and  $A_{\text{pop}}$  are some “typical” values of these two covariates in the population. Here,  $\psi_i$  will denote the PK parameters (clearance and volume) of subject  $i$  and  $\varphi_i$  its log-clearance and log-volume. Let

$$W_i^* = \log \left( \frac{W_i}{W_{\text{pop}}} \right) \quad ; \quad A_i^* = \log \left( \frac{A_i}{A_{\text{pop}}} \right)$$

Then,

$$\begin{aligned} \varphi_i &= \begin{pmatrix} \log(CL_i) \\ \log(V_i) \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & W_i^* & A_i^* & 0 \\ 0 & 1 & 0 & 0 & W_i^* \end{pmatrix} \begin{pmatrix} \log(CL_{\text{pop}}) \\ \log(V_{\text{pop}}) \\ \beta_{CL,W} \\ \beta_{CL,A} \\ \beta_{V,W} \end{pmatrix} + \begin{pmatrix} \eta_{i,1} \\ \eta_{i,2} \end{pmatrix} \\ &= C_i \mu + \eta_i \end{aligned}$$

### 2.2.3 Example of categorical covariate model

Assume that some categorical covariate  $G_i$  takes the values  $1, 2, \dots, K$ . Assume that if patient  $i$  belongs to group  $k$ , *i.e.*  $G_i = k$ , then

$$\log(CL_i) = \log(CL_{\text{pop},k}) + \eta_i$$

where  $CL_{\text{pop},k}$  is the population clearance in group  $k$ .

Let  $k^*$  be the reference group. Then, for any group  $k$ , we will decompose the population clearance  $CL_{\text{pop},k}$  as

$$\log(CL_{\text{pop},k}) = \log(CL_{\text{pop},k^*}) + \beta_k$$

where  $\beta_{k^*} = 0$ .



The variance of the random effects can also depend on this categorical covariate:

$$\eta_i \sim \mathcal{N}(0, \Omega_k) \quad \text{if } G_i = k$$

**Remark:** It is assumed in MONOLIX 4.2.1 that the correlation matrix of the random effect is the same for all the groups. In other words, only the variances of the random effects can differ from one group to another.

## 2.3 The residual error model

The within-group errors ( $\varepsilon_{ij}$ ) are supposed to be Gaussian random variables with mean zero and variance 1. Furthermore, we suppose that the  $\varepsilon_{ij}$  and the  $\eta_i$  are mutually independent.

Different error models can be used in MONOLIX 4.2.1 :

- the constant error model assumes that  $g = a$  and  $\xi = a$ ,
- the proportional error model assumes that  $g = b f$  and  $\xi = b$ ,
- a combined error model assumes that  $g = a + b f$  and  $\xi = (a, b)$ ,
- an alternative combined error model assumes that  $g = \sqrt{a^2 + b^2 f^2}$  and  $\xi = (a, b)$ ,
- a combined error model with power assumes that  $g = a + b f^c$  and  $\xi = (a, b, c)$ ,
- ...

Furthermore, all these error models can be applied to some transformation of the data:

$$t(y_{ij}) = t(f(x_{ij}, \psi_i)) + g(x_{ij}, \psi_i, \xi) \varepsilon_{ij} \quad (2.5)$$

For example:

- the exponential error model assumes that  $y > 0$ :

$$\begin{aligned} t(y) &= \log(y) \\ y &= f e^{g\varepsilon} \end{aligned}$$

- the logit error model assumes that  $0 < y < 1$ :

$$\begin{aligned} t(y) &= \log(y/(1-y)) \\ y &= \frac{f}{f + (1-f)e^{-g\varepsilon}} \end{aligned}$$

- the logit error model can be extended if we assume that  $A < y < B$ :

$$\begin{aligned} t(y) &= \log((y - A)/(B - y)) \\ y &= A + (B - A) \frac{f - A}{f - A + (B - f)e^{-g\varepsilon}} \end{aligned}$$

It is possible with MONOLIX to assume that the residual errors  $(\varepsilon_{ij})$  are correlated:

$$\text{corr}(\varepsilon_{i,j}, \varepsilon_{i,j+1}) = \rho^{(x_{i,j+1} - x_{i,j})} \quad (2.6)$$

Here, we assume that  $0 \leq \rho < 1$  and that for any  $i$ ,  $(x_{i,j}, 1 \leq j \leq n_i)$  is an increasing sequence of regression scalar variables.

## 2.4 Multi-responses model

The basic model can be extended to multi-responses:

$$\begin{aligned} y_{ij}^{(1)} &= f_1(x_{ij}^{(1)}, \psi_i) + g_1(x_{ij}^{(1)}, \psi_i; \xi_1) \varepsilon_{ij}^{(1)}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_{i1} \\ &\vdots \\ y_{ij}^{(L)} &= f_L(x_{ij}^{(L)}, \psi_i) + g_L(x_{ij}^{(L)}, \psi_i; \xi_L) \varepsilon_{ij}^{(L)}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_{iL} \end{aligned}$$

This is useful, for example, for PKPD models in which the input of the PD model  $x_{ij}^{(2)}$  is the concentration, that is the output of the PK model  $f_1(x_{ij}^{(1)}, \psi_i)$ .

## 2.5 Model with censored data

### 2.5.1 BLQ data

In some context, because of assay limitation, when data  $y_{ij}$  are inferior to a limit of quantification ( $LOQ$ ), we do not observe  $y_{ij}$  but only the censored value  $LOQ$ . These data are usually named BLQ (Below the Limit of Quantification) data or left-censored data.

Let denote  $I_{obs} = \{(i, j) | y_{ij} \geq LOQ\}$  and  $I_{cens} = \{(i, j) | y_{ij} < LOQ\}$  the index sets of the uncensored and censored observations respectively. For  $(i, j) \in I_{cens}$ , let  $y_{ij}^{cens} = y_{ij}$  denote the unknown value of the censored observation  $j$  of subject  $i$ . Let denote  $y_i^{cens}$  the vector of censored observations of subject  $i$ . Finally, we observe

$$y_{ij}^{obs} = \begin{cases} y_{ij} & \text{if } (i, j) \in I_{obs}, \\ LOQ & \text{if } (i, j) \in I_{cens}. \end{cases}$$

We denote  $y_i^{obs} = (y_{i1}^{obs}, \dots, y_{in_i}^{obs})$  as the observations of subject  $i$  and  $y^{obs} = (y_1^{obs}, \dots, y_N^{obs})$  the total observations dataset.

### 2.5.2 Interval censored data

It is possible now also to model interval censored data, i.e data where it is only known that  $y_{ij}$  is above a limit of detection  $LOD_{ij}$  but below the limit of quantification  $y_{ij}^{cens} \in [LOD_{ij}, LOQ_{ij})$ . The intervals could be  $(-\infty, LOQ_{ij})$  (left censored data, as above) and  $[LOQ_{ij}, +\infty)$  (right censored data).

## 2.6 Modeling the inter-occasion variability

We will denote  $y_{ikj}$  the  $j$ th observation for subject  $i$  during occasion  $k$ :

$$y_{ikj} = f(\psi_{ik}, t_{ikj}) + g(\psi_{ik}, t_{ikj}, \xi) \varepsilon_{ikj} \quad (2.7)$$

Here,  $\psi_{ik} = h(\varphi_{ik})$  is the individual parameter of subject  $i$  at occasion  $k$ :

$$\varphi_{ik} = C_{ik} \mu + \eta_i + \kappa_{ik} \quad (2.8)$$

- $C_{ik}$  is the matrix of covariates of subject  $i$  at occasion  $k$ ,
- $\eta_i$  random effect of subject  $i$  (inter-subject variability):  $\eta_i \sim \mathcal{N}(0, \Omega)$ ,
- $\kappa_{ik}$  random effect of subject  $i$  at occasion  $k$  (inter-occasion variability):  $\kappa_{ik} \sim \mathcal{N}(0, \Gamma)$ ,
- $\eta_i$  and  $\kappa_{ik}$  are assumed to be independent,
- $\Omega$  inter-subject variability covariance matrix,
- $\Gamma$  inter-occasion variability covariance matrix.

## 2.7 Discrete data models

The basic model proposed in (2.1) is a regression model used for fitting continuous data that can be extended for categorical data or count data models. Assume that  $(y_{ij})$  takes its values in  $\{0, 1, 2, \dots\}$ . We define the conditional likelihood of the observations using a mixed effects model:

$$\mathbb{P}(y_{ij} = k | \psi_i) = f(k, x_{ij}, \psi_i) \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \quad (2.9)$$

In other words, for any  $i$ , the probability that  $y_{ij}$  takes the value  $k$  depends on some (unknown) individual parameter  $\psi_i$  and possibly on some (known) design variable  $x_{ij}$ .

A mixed hidden Markov models (mixed HMM, or MHMM) assumes that there exists some non observed sequences  $(z_{ij})$  (the states) that take their values in  $1, 2, \dots, L$  such that, for any  $i$ ,

- $(z_{ij}, j \geq 1)$  is a Markov Chain,
- conditionally to the sequence of states  $(z_{ij})$ , the  $(y_{ij})$  are independent random variables
- the transition probabilities  $\mathbb{P}(z_{i,j+1} = v | z_{ij} = u)$  and the emission probabilities (*i.e.* conditional probabilities)  $\mathbb{P}(y_{ij} = k | z_{ij} = u)$  depend on some individual parameters  $\psi_i$ .

## 2.8 Mixture models and model mixtures

### 2.8.1 Mixture models

In MONOLIX, a mixture model assume that there exist some “latent” categorical covariate  $G$  that takes  $K$  values. Then, the mixture model reduces to the categorical covariate model described [Section 2.2](#) but here, the categorical covariates are unknown: they are treated as random variables and the probabilities

$$\pi_k = \mathbb{P}(G_i = k)$$

are part of the statistical model and should be estimated as well.

### 2.8.2 Model mixtures

Let  $f_1, f_2, \dots, f_K$  be  $K$  different structural models,

- **Between Subject Model Mixture (BSMM)**

We assume that some categorical covariate  $G$  takes  $K$  values and that

$$y_{ij} = f_k(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \quad \text{if } G_i = k$$

In a BSMM model, the “latent” categorical covariates are unknown: they are treated as random variables and the probabilities

$$\pi_k = \mathbb{P}(G_i = k)$$

are part of the statistical model and should be estimated as well.

- **Within Subject Model Mixture (WSMM)**

For any patient  $i$ , let  $p_{i,1}, p_{i,2}, \dots, p_{i,K}$  be  $K$  proportions such that

$$\begin{aligned} y_{ij} &= f_i(x_{ij}, \psi_i) + \varepsilon_{ij} \\ f_i &= p_{i,1}f_1 + p_{i,2}f_2 + \dots + p_{i,K}f_K \end{aligned}$$

In a WSMM model, the proportions  $(p_{i,k})$  are additional individual parameters that should be modeled as well (under the constraint that the sum is 1).

## 2.9 Prior models on fixed effects parameters

It is possible to define prior distribution models on the fixed effects. The allowed distributions:

- log-normal
- logit-normal
- probit-normal
- user-defined: transformation of a gaussian distribution:

$$h^{-1}(\mu) \sim \mathcal{N}(\mu_0, \sigma_\mu^2)$$

where  $h$  is any increasing function defined for all real numbers.

# Bibliography

- [1] DAVIDIAN, M., AND GILTINAN, D. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, 1995.
- [2] DAVIDIAN, M., AND GILTINAN, D. Nonlinear models for repeated measurements: An overview and update. *JABES* 8 (2003), 387–419.
- [3] DELYON, B., LAVIELLE, M., AND MOULINES, E. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* 27, 1 (1999), 94–128.
- [4] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39, 1 (1977), 1–38. With discussion.
- [5] KUHN, E., AND LAVIELLE, M. Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM P&S* 8 (2004), 115–131.
- [6] LOUIS, T. A. Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 44, 2 (1982), 226–233.
- [7] PANHARD, X., AND SAMSON, A. Extension of the SAEM algorithm for nonlinear mixed models with two levels of random effects. *Biometrics (to appear)* (2008).
- [8] PINHEIRO, J. C., AND BATES, D. M. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000.
- [9] SAMSON, A., LAVIELLE, M., AND MENTRÉ, F. Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: application to HIV dynamics model. *Computational Statistics and Data Analysis* 51 (2006), 1562–1574.
- [10] WU, C.-F. J. On the convergence properties of the EM algorithm. *Ann. Statist.* 11, 1 (1983), 95–103.