A new method for the estimation of variance matrix with prescribed zeros in nonlinear mixed effects models

Djalil Chafai, Didier Concordet

A new method for the estimation of variance matrix with prescribed zeros in nonlinear mixed effects models

Djalil Chafaï and Didier Concordet

August, 2007. Revised March, 2008.
Accepted for publication in Statistics and Computing

Abstract

We propose a new method for the Maximum Likelihood Estimator (MLE) of nonlinear mixed effects models when the variance matrix of Gaussian random effects has a prescribed pattern of zeros (PPZ). The method consists of coupling the recently developed Iterative Conditional Fitting (ICF) algorithm with the Expectation Maximization (EM) algorithm. It provides positive definite estimates for any sample size, and does not rely on any structural assumption concerning the PPZ. It can be easily adapted to many versions of EM.

Keywords: nonlinear mixed effects models; maximum likelihood; expected maximisation algorithm; logitudinal data analysis; repeated measurements; iterated proportional fitting algorithm; Gaussian graphical models; stochastic inverse problems; pharmacokinetic/pharmacodynamics analysis.

1 Introduction

Nonlinear mixed effects models are widely used in population Pharmacology for Pharmacokinetics & Pharmacodynamics (PK/PD) modelling. Such models can be seen as special cases of repeated measurement data, for which the asymptotics concern the number of individuals rather than the number of measures per individual, see for instance [8] and references therein. In this article, we show that for nonlinear mixed effects models with Gaussian random effects, Expectation Maximization (EM) like algorithms for the computation of the Maximum Likelihood Estimator (MLE) can be coupled with Iterative Conditional Fitting (ICF) like algorithms in order to take into account a prescribed pattern of zeros (PPZ) in the variance matrix of the random effect. The ICF algorithm has been developed very recently [4] in the context of directly observed Gaussian graphical models. Finding an adequate approach for generic PPZ in the context of nonlinear mixed effects models is a long standing problem. Our approach provides a true solution for the M step of EM in this context, for any PPZ. It is thus far more satisfactory than the standard approaches used in the existing software packages such as NLMIXED (SAS), NONMEM, nlme (S-Plus and GNU-R), or Monolix.

For instance, a traditional model used in population PK/PD is of the form

$$Y_i = F(X_i) + g(X_i, \theta)\varepsilon_i, \quad 1 \leq i \leq N, \tag{1}$$

where Y_i is the vector of concentrations/effects observed on the i^{th} individual for the drug of interest. Here F is a known function, often nonlinear. The q-vectors X_i represent the unobserved individual parameters assumed independent and identically distributed $\mathcal{N}(m, \Sigma)$, the ε_i are $\mathcal{N}(0, I)$, unobserved, independent of the X_i and independent. The matrix $g(X_i, \theta)$ is the Cholesky transform of a positive definite matrix that depends on a parameter $\theta \in \Theta \subset \mathbb{R}^p$.

For instance, Figure 1 represents the maximum concentrations of cortisol (Y_i) obtained after giving fixed doses of ACTH to $N = 30$ horses. Each individual has its own curve described by the parameter X_i. The $g(X_i, \theta)$ matrix defines the variance heterogeneity of concentrations obtained with different doses. This matrix is often assumed to be equal to $\theta \text{diag}(|F(X_i|)$. The reader will later find a study of a model like (1) for this cortisol data set.

One of the main goals in population PK/PD is to describe the distribution of the X_i's by observation of the Y_i's. This amounts to the estimation of (m, Σ, θ) from the Y_i's. Recall that the X_i's are not observed. A natural approach is to compute the MLE by maximizing

$$L(m, \Sigma, \theta) = \prod_{i=1}^{N} \int \frac{1}{|g(x_i, \theta)|} \phi\left(g^{-1}(x_i, \theta)(Y_i - F(x_i))\right) \phi_{m,\Sigma}(x_i) dx_i \tag{2}$$

where $\phi_{m,\Sigma}(.)$ is the probability density function of the $\mathcal{N}(m, \Sigma)$ distribution and $|g(x_i, \theta)|$ denotes the determinant of the matrix $|g(x_i, \theta)|$. Except for specific models, such as *Gaussian linear* mixed effects models, maximum likelihood estimators have no closed form. Several methods have been proposed for estimation of the parameter (m, Σ, θ) in these models. The methods suggested by Beal and Sheiner [2] or Lindstrom and Bates [14] are based on a linearization of the conditional model (1) with respect to the vector X_i about 0 or about a posterior mode. Pinheiro and Bates [15], Vonesh and Carter [18], and Wolfinger [20] proposed Laplacian approximations of the likelihood. Importance sampling approximations [15], Gaussian quadratures [8], and pseudo-likelihood methods [5] have also been investigated. The reader will find a detailed analysis of these methods in the book by Davidian and Giltinan [8]. More recently, stochastic versions of the EM algorithm have been proposed, see for instance [13] and [19]. These EM like algorithms converge to the MLE under some regularity and identifiability conditions.

In many real situations, the kineticist's knowledge of the drug mechanism imposes a specific independence pattern on some components of X_i. This means that the variance matrix Σ contains a PPZ. The estimation of (m, Σ, θ) in the presence of a PPZ in Σ is problematic due to the positive definiteness constraint in the optimization. Pinheiro and Bates [15] studied different parameterizations of Σ that ensure the definite positiveness of the estimate. In particular they suggested the usage of a Cholesky like parameterization. Unfortunately, except for the case where Σ is a block diagonal matrix up to coordinates permutation, Cholesky like parametrizations do not preserve the structure of the PPZ and are thus useless. Kuhn and Lavielle proposed estimating Σ in two steps in the implementation of their EM like algorithm. First, Σ is estimated without any constraint, then zeros are plugged according to the PPZ into the estimate provided by the first step. This method is widely used in practice. Unfortunately, by "forcing the zeros" in this way, nothing guarantees that the obtained estimate is still a positive matrix, and even when it is positive definite, it is not the maximum likelihood in general.

For *Gaussian linear* mixed effects models, the algorithm of Anderson [1] deals with any linear hypothesis on the variance matrix of the random effect (a PPZ for instance).

Unfortunately, the estimate is not necessarily positive definite, see for instance [4]. To our knowledge, no method is available for MLE of *nonlinear* models such as (1) when the variance matrix Σ of the random effect has a PPZ.

The aim of this article is to propose a general method for the estimation of (m, Σ, θ) in the presence of a PPZ in Σ. The method uses the ICF algorithm to perform the Maximization step of the EM algorithm. In other words, we couple EM and ICF in order to compute the MLE (or at least a stationary point of the likelihood) of (m, Σ, θ) when Σ has a PPZ.

The ICF algorithm was developed recently by Chaudhuri *et al.* in [4] to estimate a variance matrix with PPZ of *observed Gaussian* random variables. In contrast, the random effects X_i's in (1) are Gaussian but *not observed*, and that is why we couple ICF with EM. The ICF converges towards positive definite saddle-points or local maxima of the likelihood function irrespective of the PPZ. The idea behind ICF is not new in the framework of graphical models, and is inspired by the famous Iterative Proportional Fitting (IPF) algorithm. We refer to [4] for a review. Some alternative algorithms to ICF are available for specific PPZ, such as chain graph models [6] or non-chordal graph models [7]. The ICF algorithm is attractive because it does not rely on a specific structure of the PPZ.

The rest of the article is organized as follows. In section 2, we give some of the properties and drawbacks of the popular "zero forced" estimator, that consists of plugging zeros according to the PPZ into a full variance matrix. In section 3, we recall the main properties of the EM algorithm for models such as (1). Section 4 is devoted to the ICF algorithm, and to the coupling of ICF with EM. Section 5 contains the step by step analysis of a model like (1) for the cortisol data set depicted in Figure 1. In the last section, we perform a simulation study that quantifies the benefit of our EM+ICF approach on the model used for the cortisol data set.

2 The Zero forced estimator

Assume for example that for some specific model we get the following MLE for Σ:

$$\widehat{\Sigma}_{uc} = \begin{pmatrix} 4 & -3 & 3 \\ -3 & 4 & -3 \\ 3 & -3 & 4 \end{pmatrix},$$

without taking into account the PPZ in Σ. We will refer to this as the *unconstrained* estimation. If the PPZ consists of $\Sigma_{13} = \Sigma_{31} = 0$, the "zero forced" estimation of Σ is simply given by

$$\widehat{\Sigma}_{zf} = \begin{pmatrix} 4 & -3 & 0 \\ -3 & 4 & -3 \\ 0 & -3 & 4 \end{pmatrix}.$$

The unconstrained estimate $\widehat{\Sigma}_{uc}$ is a positive definite matrix but the "zero forced" estimate $\widehat{\Sigma}_{zf}$ is not. However we know that for a regular model, the unconstrained MLE is consistent. Therefore $\widehat{\Sigma}_{uc}$ converges componentwise towards the true matrix Σ with PPZ. Consequently, there exists a random sample size from which the "zero forced" estimator is a positive definite matrix but this sample size is somewhat difficult to obtain.

A possible ploy allowing to build a positive definite consistent estimator of Σ could be as follows. Compute the unconstrained estimator and denote it by $\widehat{\Sigma}_{zf}$, the corresponding

"zero forced" estimator. Nothing guarantees that its lower eigenvalue λ_{min} is positive but since $\widehat{\Sigma}_{zf}$ is a consistent estimator of Σ, the quantity

$$(\lambda_{min})_- \triangleq \max\{-\lambda_{min}, 0\}$$

is a random sequence of positive numbers that converges almost-surely to zero. Now, consider some auxiliary sequence of positive real numbers (u_N) that goes to zero with the sample size N (e.g. $U_N = 1/N^2$), then, for any sample size N, the matrix

$$\widehat{\Sigma}_{zf} + ((\lambda_{min})_- + u_N)\, I$$

is a positive definite consistent estimator of Σ, and features the same PPZ. Its main drawback is that its diagonal terms are biased and that the choice of the (u_N) sequence is arbitrary. A better way to proceed is to directly consider the MLE of Σ with PPZ, which is precisely our aim in the next sections.

3 The EM algorithm

The EM algorithm [9] is a popular method to estimate parameters of a model with non-observed or incomplete data. Let us briefly recall how its general form works as introduced by Dempster et al. The EM algorithm consists of iterations of an Expectation and a Maximization step. At the k^{th} iteration, the E step computes the conditional expectation of the log-likelihood of the complete data (Y, X) with respect to the distribution of the missing, or non-observed, data X given the observed data Y at the current estimated parameter value $\psi^{(k)}$:

$$Q\left(\psi, \psi^{(k)}\right) = E\left[\log P(Y, X)|Y, \psi^{(k)}\right].$$

The M step finds $\psi^{(k+1)}$ so that for all ψ in the parameter space Ψ

$$\psi^{(k+1)} = \arg\sup_{\psi\in\Psi} Q\left(\psi, \psi^{(k)}\right).$$

These two-step iterations are repeated until convergence. The essential property of the EM algorithm is that the likelihood increases monotonically along the iterations. Under some identifiability and regularity conditions, this algorithm converges to a stationary point of the likelihood, see for instance [22].

Let us now describe more precisely this algorithm for model (1). We need first to define the parameter space on which the M step is to be performed. In this model, the parameter to be estimated is $\psi = (m, \Sigma, \theta)$. The variance matrix Σ lives in a subset of the set S_q^+ of $q \times q$ symmetric positive definite matrices. More precisely, let Π be the set of subsets of $\{(i, j); 1 \leq i < j \leq q\}$. For any $\pi \in \Pi$, the set

$$S_q^+(\pi) \triangleq \{A \in S_q^+; \forall (i, j) \in \pi, A_{ij} = 0\}$$

is formed by the symmetric positive definite matrices that have zeros located in π. The PPZ in Σ is represented by an element π of Π. We thus assume that for some $\pi \in \Pi$, $\psi = (m, \Sigma, \theta) \in \Psi \triangleq M \times S_q^+(\pi) \times \Theta$ where M and Θ are open subsets of $\mathbb{R}^q$ and $\mathbb{R}^p$ respectively.

At the k^{th} iteration the Expectation step consists of the computation of

$$Q\left(m,\Sigma,\theta|m_k,\Sigma_k,\theta_k\right)=\sum_i \mathbf{E}\left(\log\frac{1}{|g(X_i,\theta)|}\phi\left(g^{-1}(X_i,\theta)(Y_i-F(X_i))\right)\phi_{m,\Sigma}(X_i)|Y_i,\Sigma_k,m_k,\theta_k\right)$$

where

$$\mathbf{E}\left(f(X,Y)|Y,\Sigma,m,\theta\right)\triangleq\int f(x,Y)\frac{\phi\left(g^{-1}(x,\theta)(Y-F(x))\right)\phi_{m,\Sigma}(x)}{|g(x,\theta)|\int\phi\left(g^{-1}(u,\theta)(Y-F(u))\right)/|g(u,\theta)|\phi_{m,\Sigma}(u)du}dx.$$

The maximization step computes

$$(m_{k+1},\Sigma_{k+1},\theta_{k+1})=\arg\sup_{M\times S_q^+(\pi)\times\Theta} Q\left(m,\Sigma,\theta|m_k,\Sigma_k,\theta_k\right).$$

For model (1), the integral that appears in the E step can be split into two parts . The E step reduces to calculate

$$Q\left(m,\Sigma,\theta|m_k,\Sigma_k,\theta_k\right)=Q_1\left(m,\Sigma|m_k,\Sigma_k,\theta_k\right)+Q_2\left(\theta|m_k,\Sigma_k,\theta_k\right)$$

where

$$\begin{aligned}Q_1\left(m,\Sigma|Y_i,m_k,\Sigma_k,\theta_k\right)&\triangleq\sum_i\mathbf{E}\left(\log\phi_{m,\Sigma}(X_i)|Y_i,m_k,\Sigma_k\right)\\&=-\frac{1}{2}\sum_i\mathbf{E}\left(\left(X_i-m\right)'\Sigma^{-1}\left(X_i-m\right)|Y_i,m_k,\Sigma_k\right)-\frac{N}{2}\log|\Sigma|\\&=-\frac{N}{2}\mathrm{tr}\left(\frac{1}{N}\sum_i\mathbf{E}\left(\left(X_i-m\right)\left(X_i-m\right)'|Y_i,m_k,\Sigma_k\right)\Sigma^{-1}\right)\\&\quad-\frac{N}{2}\log|\Sigma|,\end{aligned}$$

and

$$Q_2\left(\theta|m_k,\Sigma_k,\theta_k\right)\triangleq\sum_i\mathbf{E}\left(\log\phi\left(g^{-1}(X_i,\theta)(Y_i-F(X_i))\right)-\log(|g(X_i,\theta)|)|Y_i,m_k,\Sigma_k,\theta_k\right).$$

It follows that the M step can also be decomposed into two parts

$$\sup_{M\times S_q^+(\pi)\times\Theta} Q\left(m,\Sigma,\theta|m_k,\Sigma_k,\theta_k\right)=\sup_{M\times S_q^+(\pi)} Q_1\left(m,\Sigma|m_k,\Sigma_k,\theta_k\right)+\sup_{\Theta} Q_2\left(\theta|m_k,\Sigma_k,\theta_k\right).$$

Note that in the M step the maximization with respect to (m,Σ) is separated from that of θ. The function Q_2 depends only on θ via g.

Remark 1. *In most applications, h is the probability density function of a standard Gaussian distribution, and θ is a variance matrix that can possibly contain a PPZ. In that case, its maximization can be performed using the ICF method, as for Σ, as described hereafter.*

Maximization of Q_1 leads to

$$m_{k+1}=\frac{1}{N}\sum_i\mathbf{E}\left(X_i|Y_i,\Sigma_k,m_k,\theta_k\right)\tag{3}$$

and

$$\Sigma_{k+1} = \arg \inf_{\Sigma \in S_q^+(\pi)} \operatorname{tr}\left(\tilde{X}\Sigma^{-1}\right) + \log|\Sigma| \tag{4}$$

where

$$\tilde{X} = \frac{1}{N}\sum_i \mathbf{E}\left((X_i - m_{k+1})(X_i - m_{k+1})' \,|Y_i, \Sigma_k, m_k, \theta_k\right). \tag{5}$$

The matrix $\tilde{X}$ is an empirical conditional variance matrix. The random vectors X_i are not observed, however, the matrix $\tilde{X}$ can be evaluated at each iteration of the algorithm. When $\pi = \emptyset$, that is when Σ has no PPZ, Σ_{k+1} reduces to $\tilde{X}$. When $\pi \neq \emptyset$ the maximum of $Q_1(\Sigma)$ must be sought in $S_q^+(\pi)$. The next section deals with this problem.

4 Estimating the variance matrix with ICF

We have seen in the previous section that the M step of the EM algorithm involves a maximization problem such as

$$\Sigma_{k+1} = \arg \inf_{\Sigma \in S_q^+(\pi)} \operatorname{tr}\left(\tilde{X}\Sigma^{-1}\right) + \log|\Sigma|. \tag{6}$$

The difficulty here is that the optimization is not performed on the entire cone of symmetric definite matrices but only on a sub-cone that contains the matrices with PPZ. It is clear that a standard gradient-like algorithm does not fit these constraints. The usual method to get rid of the definite-positiveness constraint is to use a Cholesky like decomposition. Unfortunately, these decompositions do not preserve the PPZ when Σ is not a block diagonal matrix up to a permutation of the coordinates. The algorithm described hereafter allows one to move within $S_q^+(\pi)$ whatever π may be.

First, note that in the absence of PPZ in Σ, *i.e.* when $\pi = \emptyset$, the solution of (6) is $\Sigma_{k+1} = \tilde{X}$. We assume from now on that $\pi \neq \emptyset$. The case $q = 2$ is trivial since the only possible zero is $\Sigma_{12} = 0$. In this case, the "zero forced" estimator is always positive definite and is the solution of (6). We assume in the sequel that $q > 2$ and that $\pi \neq \emptyset$.

The method we propose is the core of the ICF algorithm presented in [4]. Even if Chaudhuri & al. did not express it at such, it is mainly based on the specific properties of the Schur complement of a matrix. Let us recall the following classical result, which can be found in [11] or [24] for instance.

Theorem 1 (Schur). *Let l be an integer in $\{1, \ldots, q\}$. Consider two vectors U and V that respectively belong to $\mathbb{R}^{q-l}$ and $\mathbb{R}^l$, the q-vector $Z' = (U', V')$ and a matrix $\Sigma \in S_q^+$ that admits the following block decomposition*

$$\Sigma = \begin{pmatrix} A & B \\ B' & C \end{pmatrix}$$

where $A \in S_{q-l}^+$, B is a $(q-l) \times l$ matrix and $C \in S_l^+$. The Schur complement of A is the matrix $S \triangleq C - A'^{-1}B$. It belongs to S_l^+ and moreover

i) $\det(\Sigma) = \det(A)\det(S)$,

ii) $Z'\Sigma^{-1}Z = U'A^{-1}U + (V - B'A^{-1}U)' S^{-1}(V - B'A^{-1}U)$.

The Schur complement appears naturally when the random vector Z described in the previous Theorem is distributed according to $\mathcal{N}(0, \Sigma)$. More precisely we have:

$$\mathcal{L}(V|U) = \mathcal{N}(B'A^{-1}U, S).$$

Note that properties i) and ii) remain true after permutation of the rows of Σ. In the particular case where $l = 1$, we can easily derive the following property.

Corollary 1. *We keep the same notations as in the previous proposition. For any $j \in \{1 \dots q\}$, let $A = \Sigma_{-j,-j}$ be the submatrix of Σ obtained by removing its j^{th} row and column , $B = \Sigma_{-j,j}$ the j^{th} column vector of Σ in which the j^{th} row has been removed, $C = \Sigma_{j,j}$. The column vector U and the positive real number V are respectively obtained by removing the j^{th} row of Z and as Z_j. Then, using these notations, the Schur complement of $A = \Sigma_{-j,-j}$ is the real positive number S given by the previous proposition and properties i) and ii) hold.*

We are now able to solve the optimization problem (4) by running iteratively the decomposition of the Corollary over the columns of Σ. Set $T_i = X_i - m_{k+1}$ and note that from (6) the function that has to be minimized can be rewritten as

$$K(\Sigma, T_1, \dots, T_N) = \frac{1}{N} \sum_i \mathbf{E}\left(T_i' \Sigma^{-1} T_i | Y_i, m_k, \Sigma_k\right) + \log|\Sigma|. \tag{7}$$

For the j^{th} column of Σ we set $A = \Sigma_{-j,-j}$, $B = \Sigma_{-j,j}$, $C = \Sigma_{j,j}$ and $S = C - B'A^{-1}B$ so that we can now write (7) as

$$K(\Sigma, T_1, \dots, T_N) = K(A, U_1, \dots, U_N) + K(S, V_1 - B'A^{-1}U_1, \dots, V_N - B'A^{-1}U_N)$$

where V_i is the j^{th} component of T_i and U_i is obtained by removing the j^{th} component of T_i. Therefore, if A is fixed, the partial optimization of $K(\Sigma, T_1, \dots, T_N)$ with respect to (B, S) can be reduced to the global optimization of

$$\begin{aligned} K(S, V_1 - B'A^{-1}U_1 \quad , \quad \dots, V_N - B'A^{-1}U_N) = \\ \frac{1}{NS} \sum_i \mathbf{E}\left(\left(V_i - B'A^{-1}U_i\right)^2 | Y_i, m_k, \Sigma_k\right) + \log(S) \end{aligned}$$

which is a standard least-squares problem. The optimization with respect to B and S leads to

$$\begin{aligned} B_{opt} &= \left[\sum_i \mathbf{E}\left(\left((A^{-1}U_i)(A^{-1}U_i)'\right) | Y_i, m_k, \Sigma_k\right)\right]^{-1} \sum_i \mathbf{E}\left(V_i A^{-1} U_i | Y_i, m_k, \Sigma_k\right) \quad (8) \\ &= A\left[\sum_i \mathbf{E}\left(U_i U_i' | Y_i, m_k, \Sigma_k\right)\right]^{-1} \sum_i \mathbf{E}\left(V_i U_i | Y_i, m_k, \Sigma_k\right). \quad (9) \end{aligned}$$

and

$$S_{opt} = \frac{1}{N} \sum_i \mathbf{E}\left(\left(V_i - B'_{opt} A^{-1} U_i\right)^2 | Y_i, m_k, \Sigma_k\right). \tag{10}$$

The vector $B = \Sigma_{-j,j}$ may contain some PPZ. These components are not optimized and are thus left at zero. This only decreases the dimension of the optimization problem. We

deduce that after this step on the j^{th} column of Σ, $C = \Sigma_{j,j}$ and $B = \Sigma_{-j,j}$ must be respectively updated with

$$\Sigma_{j,j}^{new} = S_{opt} + B'_{opt} \left(\Sigma_{-j,-j}\right)^{-1} B_{opt} \quad \text{and} \quad \Sigma_{-j,j}^{new} = B_{opt}.$$

The striking property of this step is that it allows us to move within S_q^+ without affecting the prescribed null components of Σ: $A = \Sigma_{-j,-j}$ and the null components of $B = \Sigma_{-j,j}$ are left unchanged. Since Σ is assumed positive definite $C = \Sigma_{j,j}$ cannot be zero.

As already mentioned, iterations of these steps converge to a local maximum of $K(\Sigma, T_1, \ldots, T_N)$, see for instance [4].

5 The cortisol data set

In this section, we use a practical example to illustrate the implementation of an EM algorithm coupled with the ICF algorithm. In order to explore the endocrine function of horse, a sample of horses ($N = 30$) was given eight doses of ACTH by intravenous route. The ACTH stimulates the adrenal gland that produces cortisol. The concentration profiles of cortisol in plasma were summarized by the maximal concentration reached after the ACTH administration (see Figure 1).

The seven doses of ACTH given to each animal were (in mg/kg)

$$0.005, \quad 0.01, \quad 0.1, \quad 0.5, \quad 1, \quad 2, \quad 10.$$

The production of cortisol is modelled as

$$Y_{ij} = \left(X_{1i} + \frac{X_{2i} d_j^{X_{3i}}}{X_{4i}^{X_{3i}} + d_j^{X_{3i}}} \right) (1 + \sigma \varepsilon_{ij}), \quad 1 \leq j \leq 7, \quad 1 \leq i \leq 30, \tag{11}$$

where Y_{ij} is the maximal cortisol concentration observed in the i^{th} horse after administration of a dose d_j of ACTH, $X_i' = (X_{i1}, \ldots, X_{i4})$ is a random vector that contains the individual parameters for the i^{th} animal. We assume that the random vectors X_i are independent and identically distributed $\mathcal{N}(m, \Sigma)$ and that the residual terms $\varepsilon_i' = (\varepsilon_{i1}, \ldots, \varepsilon_{i7})$ are independent and identically distributed $\mathcal{N}(0, I_7)$. Moreover, the X_i's and ε_i's are assumed to be mutually independent. In this example, $p = 4$, $\theta = \sigma^2$ and

$$g(X, \theta) = \sigma \text{diag} \left(X_1 + \frac{X_2 d_j^{X_3}}{X_{4i}^{X_3} + d_j^{X_3}} \right)_{j=1\ldots7}.$$

According to the kineticist, the correlations between X_{i1} and X_{i4} and X_{i3} and X_{i4} should be zero and thus Σ has the following structure

$$\Sigma = \begin{pmatrix} . & . & . & 0 \\ . & . & . & . \\ . & . & . & 0 \\ 0 & . & 0 & . \end{pmatrix}.$$

In some problems, no *a priori* information is available for the possible zero correlation between the components of X_i. The method of multiple testing of correlation, as described in Drton and Perlman [10], may be used in such cases to reveal the structure of Σ.

The estimation of model parameters requires evaluation of the conditional expectations of functions such as $\mathbf{E}\left(f(X_i, Y_i)|Y_i, \Sigma_k, m_k, \sigma_k^2\right)$. A standard approach is to use a stochastic version of EM that consists of the simulation of a Markov Chain $(X_i^{(l)})_l$ with $P\left(.|Y_i, \Sigma_k, m_k, \sigma_k^2\right)$ as unique stationary distribution by using a Metropolis-Hastings algorithm and the approximation of the conditional expectation by

$$\mathbf{E}\left(f(X_i, Y_i)|Y_i, \Sigma_k, m_k, \sigma_k^2\right) \approx \frac{1}{L}\sum_{l=1}^{l} f(X_i^{(l)}, Y_i).$$

For the analysis of the cortisol data we chose $L = 500$. We simulated the Markov chain with the Metropolis-Hastings algorithm with $\mathcal{N}(m_k, \Sigma_k)$ as the proposal distribution. In this case, the acceptance probability of the Metropolis-Hastings algorithm reduces to

$$\min\left(\frac{\phi\left(g^{-1}(X, \sigma_k^2)(Y - F(X))\right)|g(x^{(l)}, \sigma_k^2)|}{|g(X, \sigma_k^2)|\phi\left(g^{-1}(x^{(l)}, \sigma_k^2)(Y - F(x^{(l)}))\right)}, 1\right)$$

which only depends on the conditional distribution of the observation.

The algorithm to estimate the model parameters can be summarized in the following scheme:

1. Start from some initial guess $\Sigma_0, m_0, \sigma_0^2$ and set $k = 0$;
2. Compute m_{k+1} from (3) and
$$\sigma_{k+1}^2 = \frac{1}{N}\sum_i \mathbf{E}\left((Y_i - F(X_i))'C^{-1}(X_i)(Y_i - F(X_i))|Y_i, m_k, \Sigma_k, \sigma_k^2\right),$$
where $C^{-1}(X_i)$ is a diagonal matrix whose j^{th} term is $(1/(F(X_i))_j)^2$.
3. Set $\Sigma_k^{(0)} = \Sigma_k$ and and set $l = 0$
4. for j:=1 to $q = 4$
increment l
compute $\Sigma_{j,j}^{(l+1)} = S_{opt} + B'_{opt}\left(\Sigma_{-j,-j}^{(l)}\right)^{-1} B_{opt}$ and $\Sigma_{-j,j}^{(l+1)} = B_{opt}$ where S_{opt} and B_{opt} are respectively defined by (10) and (8) ;
5. if $\Sigma_k^{(l-4)}$ and $\Sigma_k^{(l)}$ are not close enough go to step 4. Otherwise set $\Sigma_{k+1} = \Sigma_k^{(l)}$;
6. Stop if $(\Sigma_k, m_k, \sigma_k^2)$ and $\left(\Sigma_{k+1}, m_{k+1}, \sigma_{k+1}^2\right)$ are close enough. Otherwise increment k and go to step 2.

For standard EM algorithms, *i.e.* when no constraint is imposed on the structure of Σ, steps 3), 4) and 5) of the previous algorithm should be replaced by the update of Σ_k according to equation (5).

It is well known that standard EM algorithms go quickly to a stationary point of the likelihood during the first iterations and then take time to converge. Since for stochastic EM algorithms the criterion being optimized changes randomly at each iteration, it is somewhat difficult to achieve and check convergence even when the length L of the simulated Markov Chain is large. Improvements have been proposed to overcome these problems. In particular Kuhn and Lavielle [13] suggested updates of the following form :

$$\begin{cases} \Sigma_{new} = (1 - \gamma_k)\Sigma_k + \gamma_k\Sigma_{k+1} \\ m_{new} = (1 - \gamma_k)m_k + \gamma_k m_{k+1} \\ \sigma_{new}^2 = (1 - \gamma_k)_k\sigma_k^2 + \gamma_k\sigma_{k+1}^2, \end{cases}$$

where $(\gamma_k)_k$ is a decreasing sequence of positive numbers, $\left(\Sigma_{k+1}, m_{k+1}, \sigma^2_{k+1}\right)$ is defined as in the previous algorithm and $(\Sigma_{new}, m_{new}, \sigma^2_{new})$ is the update of $(\Sigma_k, m_k, \sigma^2_k)$. Note that this update scheme forces the algorithm to converge and preserves the PPZ as well as the positive definiteness of Σ.

The sequence $(\gamma_k)_k$ should satisfy $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$. These two conditions are fulfilled when $\gamma_k = a/k^b$ with $a > 0$ and $b \in (0, 1)$. Choosing $\gamma_k = a/k$ speeds-up convergence of the algorithm but the choice of a has to be made sample by sample. Choosing the same a for all samples can lead to poor estimations. As practical advice, we suggest choosing $\gamma_k = 1/k^{0.8}$. The algorithm takes more time to converge but a fine tuning of a is unnecessary.

A well known drawback of the EM algorithm is that it does not produce standard errors as a by-product. We implemented the method proposed by Jamshidian and Jennrich [12]. This method relies on numerical derivation and seems well-suited to the method we propose. Even if standard errors are helpful for comparing the results obtained with these two models, the Fisher information matrix gives pertinent quantitative information only when the sample size is large enough. However, $N = 30$ is probably not a large sample size. In the next section we use simulations to weight the performance of the estimation proposed for the cortisol data.

The estimation of the model parameters for the cortisol data requires some initial estimates to be provided. Thanks to the model parametrization, we can directly read reasonable values for m_0 on Figure 1. Since the four components of X respectively represent the basal value of cortisol, the maximal increase, the "slope" of the sigmoid and the ACTH dose for which half the maximal increase is obtained we roughly get $m_0 = (50, 70, 1, 0.1)$. We initialize Σ with the following diagonal matrix: $\Sigma_0 = \mathrm{Diag}\left(0.01 m_0^2\right)$. Finally, for this heteroscedastic model, σ can be interpreted as the coefficient of variation of the cortisol for a given dose. We set it at 20% that is $\sigma_0^2 = 0.2^2$. We estimated Σ with EM alone (no constraint was imposed) and with EM+ICF that preserves the PPZ. In this example, the algorithm seems to converge in less than 400 iterations. We implemented this algorithm in C++ with a matrix library. Estimates of the parameters obtained with EM alone were:

$$\widehat{\Sigma} = \begin{pmatrix} 21.25(7.90) & & & \\ 6.28 & 4.25(1.05) & & \\ 0.13 & 0.33 & 0.047(0.0071) & \\ 0.015 & 0.000867 & -0.000267 & 0.0000213(0.000016) \end{pmatrix},$$

The figures between brackets are the standard errors for the variances. We have chosen to give only some standard errors to lighten the presentation. $\widehat{m} = (50.03, 69.81, 1.78, 0.0845)$, $se\left(\widehat{m}\right) = (0.87, 1.84, 0.085, 0.0069)$, $\widehat{\sigma^2} = 0.0145$ and $\log L(\widehat{\Sigma}, \widehat{m}, \widehat{\sigma^2}) = -750.25$. Estimates of the parameters obtained with the EM+ICF algorithm were:

$$\widehat{\Sigma} = \begin{pmatrix} 19.50(7.85) & & & \\ -4.66 & 2.33(1.12) & & \\ -0.29 & -0.095 & 0.058(0.0063) & \\ 0 & -0.0024 & 0 & 0.0000144(0.000012) \end{pmatrix},$$

$\widehat{m} = (48.84, 71.46, 1.47, 0.0840)$, $se\left(\widehat{m}\right) = (0.82, 1.81, 0.084, 0.0071)$ and $\widehat{\sigma^2} = 0.0151$ and $\log L(\widehat{\Sigma}, \widehat{m}, \widehat{\sigma^2}) = -754.23$. We can see that these likelihoods are about the same and a likelihood ratio test would not reject the PPZ proposed by the kineticist. The residual variance estimates are also very close. Surprisingly, there are quite large differences between the estimates of the third component of m and the non null components of Σ.

Remark 2 (Modelling). *The general problem of mean and variance modelling for longitudinal data is delicate, and several choices are possible, see for instance [8], [21], [16, 17], [25], and [3]. Our model* (11) *belongs to a standard family of models in PK/PD and was chosen with the kineticist. This relatively simple model is heteroscedastic with a constant coefficient of variation. An examination of the "individual" residuals shows that they are centered, which is quite satisfactory.*

6 Simulations

The aim of this section is to quantify the potential benefit of directly estimating a variance matrix with PPZ. We simulated 100 data sets using model (11) with parameters close to the estimate found in the cortisol data analysis : $N = 30$,$m' = (50, 70, 1.5, 0.08)$,

$$\sigma^2 = 0.015 \quad \text{and } \Sigma = \begin{pmatrix} 20 & & & \\ -4.5 & 2.5 & & \\ -0.3 & -0.1 & 0.05 & \\ 0 & -2 \times 10^{-3} & 0 & 10^{-5} \end{pmatrix}.$$

Both EM and EM+ICF estimates were calculated. Results are given in Table 1.

As expected, the standard errors given in the example are smaller than those of Table 1. For such sample sizes, which are often encountered in practice, asymptotic statistics should be interpreted with care.

The mean parameter m seems to be well estimated. At least on these simulations, the Σ structure influences the estimation of m. However, we notice that the estimates obtained with EM+ICF have a smaller standard error and mean quadratic error (M.Q.E.) than those obtained without any constraint. On the whole EM+ICF also gives estimates with lower bias. This suggests that the mean and variance estimations are heavily dependent. This sheds light on approaches, such as the 'zero forced" method, that rely on estimating the full variance matrix first and modify it by forcing the PPZ: since all the non zero entries are estimated with the assumption that the variance matrix does not have prescribed zeros, they could be poorly estimated. This is consistent with the results obtained by Ye and Pan [23] who concluded, in a different context, that misspecification of the working variance structure could lead to a large loss in efficiency of the estimators of the mean parameters.

Likelihood ratio tests were performed to test

$$\begin{cases} H_0 : & \Sigma \in S_4^+(\pi) \\ H_1 : & \Sigma \in S_4^+ \end{cases}$$

for $\pi = \{(1, 4); (3, 4)\}$. Note that whatever the value of π, H_0 is not on the boundary of S_4^+. Consequently, the likelihood ratio statistics follow asymptotically a Chi-square distribution under H_0. Since the data have been simulated under H_0, the P-values distribution should be close to a uniform law on $(0, 1)$ at least for large N. The Q-Qplot of the P-Values is represented in Figure 2. This figure shows that the P-Values are not distributed according to a uniform distribution and thus the distribution of the likelihood ratio statistics is not close to a χ^2 distribution. Consequently, $N = 30$ is probably not large enough to trust asymptotic statistics.

7 Conclusion

We have proposed a method for the estimation of the variance matrix with PPZ in nonlinear mixed effects models. This method, which consists of coupling an ICF like algorithm with an EM like algorithm gives more efficient estimates than standard EM that ignore the PPZ. For the sake of simplicity, we have only presented the estimation algorithm for independent and identically distributed observations. Extension to different numbers of observations per individual is straightforward. We also restricted our study to models with Gaussian ε. More general models in which the distribution of ε is not Gaussian and depends on a parameter θ_2 can also be considered. This simply requires the Metropolis-Hastings chain to be chosen accordingly. We deliberately chose to show in section 2 a columnwise ICF implementation that can be extended using theorem 1 to blocks of Σ.

Of course, our approach can be adapted without much effort to many versions of EM and many alternatives to ICF. For pedagogical reasons, we presented our EM+ICF coupling on a low dimensional example. The method of course is particularly suited to large variance matrices with a high percentage of prescribed zero entries.

Acknowledgements. We thank Dr. Alain Bousquet-Mélou who provided the cortisol data set and for valuable advice regarding the model and the specific structure of Σ. We also thank Dr. Mathias Drton for interesting discussions on the ICF algorithm and for providing the last version of [4] before publication. The final form of this article has greatly benefited from the questions and suggestions of two anonymous reviewers.

References

[1] T. W. Anderson. Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.*, 1:135–141, 1973.

[2] S. L. Beal and L. B. Sheiner. *NONMEM user's guide. Nonlinear Mixed Effects Models for Repeated Measures Data*. University of California, San Francisco, 1992.

[3] E. C. Cepeda and D. Gamerman. Bayesian modeling of joint regressions for the mean and covariance matrix. *Biom. J.*, 46(4):430–440, 2004.

[4] S. Chaudhuri, M. Drton, and T. S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216, 2007.

[5] D. Concordet and O. G. Nunez. A simulated pseudo-maximum likelihood estimator for nonlinear mixed models. *Computational Statistics & Data Analysis*, 39(2):187–201, 2002.

[6] D. R. Cox, N. Wermuth, and G. Marchetti. Decompositions and estimation of a chain of covariances. Technical report, Department of Mathematical Statistics, Chalmers Goteborgs Universitet, 2004.

[7] J. Dahl, L. Vandenberghe, and V. Roychowdhury. Covariance selection for non-chordal graphs via chordal embedding. Preprint, available on http://www.ee.ucla.edu/ vandenbe/covsel.html, 2006.

[8] M. Davidian and D. M. Giltinan. *Nonlinear Models for Repeated Measurement Data*. New York: Chapman and Hall, 1995.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion.

[10] M. Drton and M. D. Perlman. Model selection for gaussian concentration graphs. *Biometrika*, 91:591–602, 2004.

[11] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

[12] M. Jamshidian and R. I. Jennrich. Standard errors for em estimation. *J. R. Statist. Soc. B*, 62:257–270, 2000.

[13] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4):1020–1038, 2005.

[14] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

[15] J. C. Pinheiro and D. M. Bates. Approximation of the log-likelihood function in the nonlinear mixed effects models. *J. Comp. Graph. Statist.*, 4:12–35, 1995.

[16] M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

[17] M. Pourahmadi. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87(2):425–435, 2000.

[18] E. F. Vonesh and R. L. Carter. Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, 48:1–17, 1992.

[19] J. Wang. EM algorithms for nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 51(6):3244–3256, 2007.

[20] R. D. Wolfinger. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80(4):791–795, 1993.

[21] R. D. Wolfinger. Heterogeneous variance-covariance structures for repeated measures. *J. Agric. Biol. Environ. Stat.*, 1(2):205–230, 1996.

[22] C. F. Wu. On the convergence properties of the EM algorithm. *Ann. Statis.*, 11:95–103, 1983.

[23] H. Ye and J. Pan. Modelling of covariance structures in generalised estimating equations for longitudinal data. *Biometrika*, 93:927–941, 2006.

[24] F. Zhang, editor. *The Schur Complement and its Applications*, volume 4 of *Numerical Methods and Algorithms*. Springer-Verlag, New York, 2005.

[25] D. L. Zimmerman and V. Núñez-Antón. Parametric modelling of growth curve data: an overview. *Test*, 10(1):1–73, 2001. With comments and a rejoinder by the authors.

Didier Concordet, d.concordet(AT)envt.fr
Djalil Chafaï, d.chafai(AT)envt.fr (corresponding author)

UMR181, INRA, ENVT,
23, Chemin des Capelles, B.P. 87614, F-31076 Toulouse CEDEX 3, France
and
UMR CNRS 5219, Institut de Mathématiques de Toulouse,
Université Paul Sabatier, 118 route de Narbonne, F-31062, Toulouse CEDEX 9, France.

		EM			EM+ICF		
	True value	Mean	S.E.	$\sqrt{M.Q.E.}$	Mean	S.E.	$\sqrt{M.Q.E.}$
Σ_{11}	20	17.3	10.74	11.07	18.88	9.09	9.16
Σ_{12}	-4.5	-3.64	3.16	3.28	-4.1	2.39	2.42
Σ_{22}	2.5	2.27	1.48	1.5	2.74	1.25	1.28
$\Sigma_{13} \times 10^2$	-30	-8.04	86.53	89.27	-12.27	39.85	43.62
$\Sigma_{23} \times 10^2$	-10	-7.57	19.95	20.1	-10.66	14.65	14.66
$\Sigma_{33} \times 10^3$	50	74.36	95.6	98.66	73.67	65.71	69.85
$\Sigma_{14} \times 10^4$	0	-7	91.62	91.89	0	0	0
$\Sigma_{24} \times 10^4$	-20	49.78	306.23	314.08	104.01	204.02	238.76
$\Sigma_{34} \times 10^5$	0	-43.69	63.29	76.9	0	0	0
$\Sigma_{44} \times 10^6$	10	17.84	14.15	16.18	18.57	13.68	16.15
m_1	50	51.18	1.38	1.82	48.96	1.23	1.61
m_2	70	70.73	2.60	2.70	69.52	2.44	2.48
m_3	1.5	1.54	0.22	0.22	1.49	0.21	0.22
$m_4 \times 10^3$	80	91.44	4.08	12.15	90.06	3.87	10.78
$\sigma^2 \times 10^3$	15	15.91	1.77	1.99	14.07	1.57	1.82
log like.		-749.32	11.39		-751.54	11.68	

Table 1: Empirical mean, standard error and square root of mean-quadratic-error of the estimates (M.Q.E.) obtained with EM and EM+ICF. The Mean Quadratic-Error is defined as bias^2+ Variance.

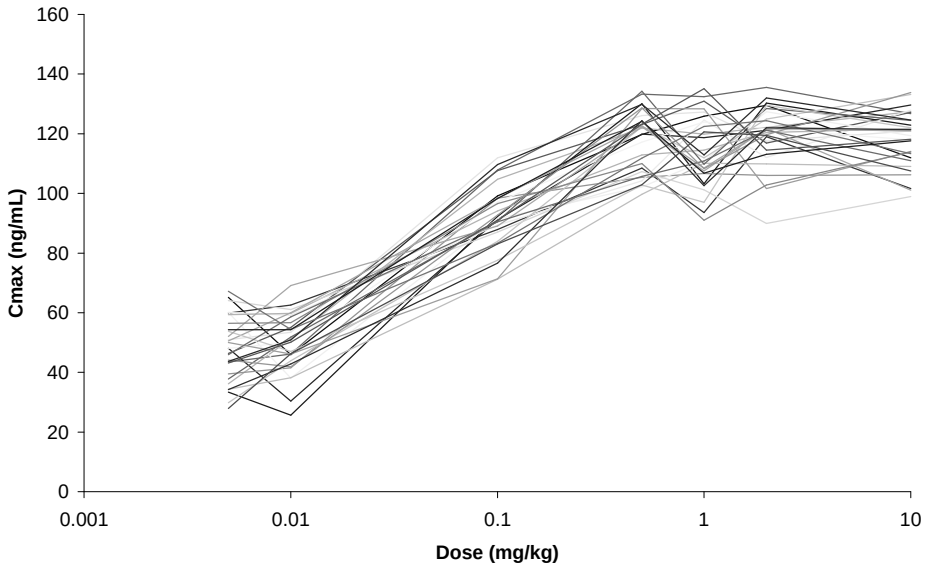

Figure 1: Maximum cortisol concentrations observed after IV administrations of ACTH in 30 horses.

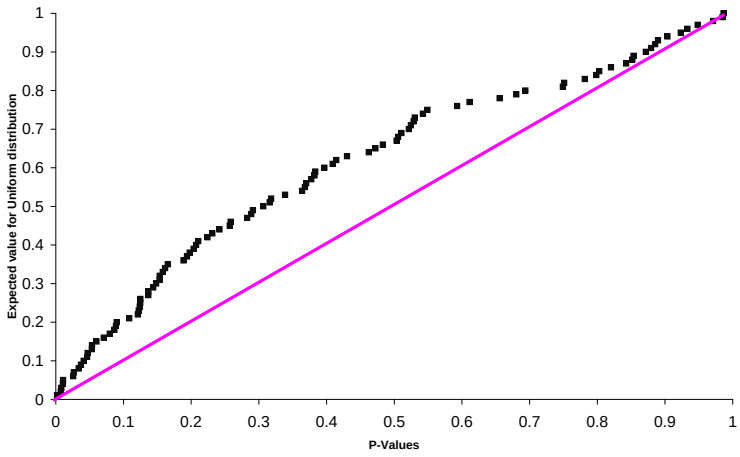

Figure 2: The Q-Q plot of the P-values of the likelihood ratio test versus the uniform distribution on $(0, 1)$.