JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN-LOUIS FOULLEY

# Algorithme EM : théorie et application au modèle mixte

*Journal de la société française de statistique*, tome 143, n° 3-4 (2002), p. 57-109

<http://www.numdam.org/item?id=JSFS_2002__143_3-4_57_0>

# ALGORITHME EM : THÉORIE ET APPLICATION AU MODÈLE MIXTE

Jean-Louis FOULLEY *

RÉSUMÉ

Cet article présente une synthèse pédagogique sur la théorie de l'algorithme EM et son utilisation dans les modèles linéaires mixtes en vue du calcul des estimations des composantes de variance par maximum de vraisemblance.

Dans une première partie, on rappelle les bases théoriques de l'algorithme. L'accent est mis sur l'importance du concept de données manquantes et sur l'identité de Fisher qui permet bien de comprendre pourquoi l'algorithme présente deux phases : la 1re consiste en la détermination de l'espérance de la logvraisemblance des données complètes (dite de ce fait E) par rapport à la distribution conditionnelle des données manquantes sachant les données observées et les valeurs courantes du paramètre ; la seconde (dite M) permet d'actualiser ces valeurs courantes par maximisation de la fonction obtenue précédemment. On présente ensuite les principales propriétés de l'algorithme notamment l'accroissement monotone de la vraisemblance des données observées, la convergence vers un point stationnaire de cette vraisemblance et ses caractèristiques. On termine par un inventaire des principales variantes : Gradient-EM, ECM , ECME, EM stochastique, EM supplémenté et PX-EM.

La deuxième partie traite plus spécifiquement du calcul des estimations du maximum de vraisemblance dans le cadre du modèle linéaire mixte gaussien. On considère tout d'abord l'application de l'algorithme EM standard à un ou plusieurs facteurs aléatoires indépendants en envisageant une estimation par maximum de vraisemblance standard (ML) ou résiduelle (REML). On introduit ensuite une variante de cet algorithme dit EM normalisé dans laquelle les effets aléatoires sont standardisés. On montre comment cette technique de standardisation permet d'obtenir des estimations REML (ou ML) de paramètres décrivant un modèle structurel de variances hétérogènes. On termine enfin par le cas du modèle linéaire mixte à plusieurs facteurs aléatoires corrélés en décrivant l'algorithme EM à la fois sous sa forme classique et sous la version dite PX d'expansion paramétrique qui améliore nettement les performances de l'algorithme.

*Mots clés* : Algorithme EM ; Modèles mixtes ; Composantes de variance ; ML ; REML ; PX-EM.

ABSTRACT

This paper presents a pedagogical review about the theory of the EM algorithm and its use in mixed linear models for computing variance component estimations by maximum likelihood. In the first part of the paper, emphasis is put on the concept of missing information and the Fisher identity which helps to explain why the algorithm deals with two steps. The first one (the so called E step) consists in

* INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy en Josas Cedex
foulley@jouy.inra fr

evaluating the expectation of the complete loglikelihood function with respect to the conditional distribution of the missing data given the observed data and the current value of the parameters; the second one, known as the M step, updates those parameter values by maximising the function defined at the previous stage. The most important properties of the algorithm are then presented such as the monotonicity of the algorithm, its convergence and rate of convergence to stationary values of the observed likelihood. Then modifications and extensions of the algorithm are presented and discussed such as gradient-EM, ECM, ECME, Stochastic EM, Supplemented EM and PX-EM.

The second part is concerned more specifically with maximum likelihood estimations under the mixed linear Gaussian model. The application of standard EM in the case of a single or several independent random factors is first considered using either both regular (ML) or residual (REML) maximum likelihood estimation of variance components. We deal next with a modification of this algorithm, the so called "Scaled EM" in which random effects are standardised. It is shown how this technique provides REML (or ML) estimation of dispersion parameters in a structural model for heterogeneous variances. Finally, we consider the case of the linear mixed model with several correlated random effects by describing the EM algorithm both under its standard and "parameter expanded" (PX) versions which greatly improves its performance.

*Keywords* : EM algorithm; Mixed models; Variance components; ML; REML; PX-EM.

## Plan de l'article

## Introduction

Le modèle linéaire mixte est un domaine de prédilection pour l'application de l'algorithme EM. Un développement particulier lui était déjà consacré dans le chapitre « Exemples » de l'article séminal de Dempster, Laird et Rubin (§ 4.4, pages 17-18) et la tendance s'est poursuivie par la suite (Laird, 1982 ; Laird et Ware, 1982 ; Laird, Lange et Stram, 1987 ; Meng et van Dyk, 1998 ; van Dyk, 2000). Une mention particulière est à attribuer au monde de la statistique appliquée qui a très largement contribué par le nombre de ses publications à la vulgarisation et au succès de l'algorithme EM (Meng et van Dyk, 1997).

En fait, Henderson anticipait EM dès 1973 en proposant un algorithme de calcul des estimations du maximum de vraisemblance des composantes de variance d'un modèle linéaire mixte qui s'avérera ultérieurement très proche de la solution EM standard.

Mais l'algorithme EM a une portée beaucoup plus générale. C'est effectivement un algorithme qui permet d'obtenir les estimations du maximum de vraisemblance dans les modèles où apparaissent des données manquantes ou qui peuvent être formalisés comme tels. Dans l'algorithme EM, le concept de données manquantes dépasse son acception classique (observations initialement planifiées mais qui ne sont pas effectuées) pour englober le cas de variables (ou processus) aléatoires de tout modèle théorique sous jacent aux observations réelles (Meng, 2000).

De fait, EM tient naturellement sa réputation et son succès, en tant qu'algorithme, de ses qualités intrinsèques de généralité, stabilité et simplicité, mais il dépasse ce cadre strictement numérique pour faire partie intégrante du mode de pensée statistique comme l'illustrent ses liens avec les techniques dites d'augmentation de données (Tanner and Wong, 1987 ; Van Dyk and Meng, 2001), avec le concept de variables cachées (ou auxiliaires ou latentes) et les méthodes de simulation de Monte Carlo par chaînes de Markov (Robert et Casella, 1999).

Dans ce contexte, il nous est paru utile de consacrer un développement spécifique au domaine du calcul des estimations ML et REML des composantes de la variance. Cela dit, un tel développement nécessite des connaissances élémentaires sur l'algorithme en général. C'est la raison pour laquelle nous avons fait précéder l'application au modèle mixte d'une présentation théorique générale de l'algorithme, de ses propriétés et de ses principales variantes. Ces rappels de théorie devraient également permettre d'aborder d'autres secteurs d'application de l'algorithme tels que, par exemple, celui des mélanges ou celui des modèles de Markov cachés.

# 1. Théorie

Avant de définir formellement l'algorithme et ses deux étapes E « Expectation » et M « Maximisation », nous allons tout d'abord montrer, à travers un exemple simple, comment on peut appréhender empiriquement les principes de base de l'EM, puis nous établirons, à partir des règles du calcul différentiel, un résultat théorique élémentaire dont la lecture conduit immédiatement à la formulation de l'algorithme.

## 1.1. Exemple

Celui-ci a trait à l'estimation des fréquences alléliques au locus de groupe sanguin humain ABO qui est un problème classique de génétique statistique (Rao, 1973 ; Weir, 1996). Il s'agit d'un locus autosomal à 3 allèles A, B et O, ce dernier étant récessif par rapport aux deux premiers qui sont codominants entre eux : on observe donc les phénotypes [A] (génotypes AA et AO), [B] (génotypes BB et BO), [AB] (génotype AB) et [O] (génotype OO). Sous l'hypothèse d'une population panmictique de grande taille en équilibre de Hardy-Weinberg, les fréquences des génotypes AA, AO, BB, BO, AB et OO sont respectivement de $p^2$, $2pr$, $q^2$, $2qr$, $2pq$ et $r^2$ si l'on désigne respectivement par $p$, $q$ et $r$ les fréquences des allèles A,B et O. L'estimation par maximum de vraisemblance de ces fréquences peut être abordée classiquement en exprimant la logvraisemblance des données et les dérivées premières et secondes de celle-ci par rapport aux paramètres.

Soit $\mathrm{L}(\boldsymbol{\phi};\mathbf{y}) = \ln p(\mathbf{y}|\boldsymbol{\phi})$ la logvraisemblance où $\mathbf{y} = (y_A, y_B, y_{AB}, y_O)'$ est le vecteur des nombres observés des différents phénotypes, $\boldsymbol{\Pi} = (\pi_A, \pi_B, \pi_{AB}, \pi_0)'$ celui homologue de leurs probabilités et $\boldsymbol{\phi} = (p, q, r)'$ celui des paramètres qui se réduit à $\boldsymbol{\phi} = (p, q)'$ puisque $p + q + r = 1$. Comme il s'agit d'un échantillonnage multinomial typique, $\mathrm{L}(\boldsymbol{\phi};\mathbf{y})$ s'écrit

$$\mathrm{L} = \sum_{j=1}^{4} y_i \ln \pi_j + Cste, \tag{1}$$

où $\pi_j$ est le $j^{\text{ème}}$ élément de $\boldsymbol{\Pi} = (\pi_A, \pi_B, \pi_{AB}, \pi_0)'$.

On en tire les expressions des scores $\mathbf{S} = \{s_k = \partial \mathrm{L}/\partial \phi_k\}$

$$s_k = \sum_{j=1}^{4} \frac{y_j}{\pi_j} \frac{\partial \pi_j}{\partial \phi_k}, \tag{2}$$

et des éléments de la matrice d'information de Fisher $\mathbf{I} = \{I_{kl}\} = \mathrm{E}\left(-\dfrac{\partial \mathrm{L}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right)$

$$I_{kl} = N \sum_{j=1}^{4} \frac{1}{\pi_j} \frac{\partial \pi_j}{\partial \phi_k} \frac{\partial \pi_j}{\partial \phi_l}, \tag{3}$$

où $N = \sum_{j=1}^{4} y_j$.

Comme les $\pi_j$ ne sont pas des fonctions élémentaires des paramètres $p$ et $q$, les expressions des $s_k$ et $I_{kl}$ ne sont pas immédiates et leur obtention s'avère quelque peu fastidieuse.

À l'inverse, les choses deviennent beaucoup plus simples si l'on suppose que tous les génotypes sont observés. En désignant par $x_k$ le nombre d'individus de génotype $k$, les estimateurs du maximum de vraisemblance (ML) de $p$ et $q$ s'obtiennent classiquement par les fréquences des gènes A et B dans l'échantillon soit :

$$p' = (2x_{AA} + x_{AB} + x_{AO})/2N; \quad q' = (2x_{BB} + x_{AB} + x_{BO})/2N, \tag{4}$$

avec ici $x_{AB} = y_{AB}$.

Il est naturel de remplacer dans ces expressions les observations manquantes $x_{AA}$, $x_{AO}$ et $x_{BB}$, $x_{BO}$ par des prédictions de celles-ci compte-tenu des observations faites ($\mathbf{y}$) et du modèle adopté (équilibre de Hardy-Weinberg) soit $x^{\#}_{AA} = \dfrac{p^2}{p^2 + 2pr}\, y_A$ et $x^{\#}_{AO} = \dfrac{2pr}{p^2 + 2pr}\, y_A$ ou après simplification :

$$x^{\#}_{AA} = \frac{p}{p + 2r}\, y_A; \quad x^{\#}_{AO} = \frac{2r}{p + 2r}\, y_A. \tag{5}$$

On procède de même par symétrie pour $x_{BB}$ et $x_{BO}$. En reportant ces quantités dans (4), on obtient les estimations suivantes :

$$p'' = (2x^{\#}_{AA} + y_{AB} + x^{\#}_{AO})/2N; \quad q'' = (2x^{\#}_{BB} + y_{AB} + x^{\#}_{BO})/2N \tag{6}$$

Les prédictions en (5) dépendant des valeurs des paramètres, le procédé va donc être appliqué de façon itérative : on va utiliser les valeurs actualisées des paramètres en (6) pour remettre à jour les prédictions des observations « manquantes » en (5), et celles-ci obtenues, on les reporte en (6) pour obtenir de nouvelles estimations des paramètres et ainsi de suite. On a, de cette façon, construit un algorithme itératif qui comporte deux étapes :

1) prédiction des données manquantes en fonction des valeurs courantes des paramètres et des observations ;

2) estimation des paramètres en fonction des prédictions actualisées et des observations, et qui préfigurent à la lettre respectivement les étapes E et M de l'algorithme de Dempster, Laird et Rubin.

On peut appliquer ce raisonnement à l'échantillon suivant : $y_A = 179$, $y_B = 35$, $y_{AB} = 6$ et $y_O = 202$. Les estimations du maximum de vraisemblance obtenues directement sont $\widehat{p} = 0.251560$, $\widehat{q} = 0.050012$ et $\widehat{r} = 0.698428$. Les résultats de l'algorithme EM figurent au tableau 1. La convergence s'effectue en quelques itérations y compris pour des valeurs de départ très éloignées de la solution.

TABLEAU 1. – Exemple de séquences EM dans le calcul des estimations ML des fréquences géniques $p, q$ et $r$ des allèles A,B et O

| Itération | $p$ | $q$ | $r$ |
|---|---|---|---|
| | Valeurs initiales égales | | |
| 0 | 0.33333333 | 0.33333333 | 0.33333333 |
| 1 | 0.28988942 | 0.06240126 | 0.64770932 |
| 2 | 0.25797623 | 0.05048400 | 0.69153977 |
| 3 | 0.25253442 | 0.05003857 | 0.69742702 |
| 4 | 0.25170567 | 0.05001433 | 0.69827999 |
| 5 | 0.25158173 | 0.05001197 | 0.69840630 |
| 6 | 0.25156326 | 0.05001165 | 0.69842509 |
| 7 | 0.25156051 | 0.05001161 | 0.69842788 |
| 8 | 0.25156010 | 0.05001160 | 0.69842830 |
| 9 | 0.25156004 | 0.05001160 | 0.69842836 |
| | Valeurs initiales quelconques | | |
| 0 | 0.92000000 | 0.07000000 | 0.01000000 |
| 1 | 0.42676717 | 0.08083202 | 0.49240082 |
| 2 | 0.28331520 | 0.05172378 | 0.66496102 |
| 3 | 0.25644053 | 0.05002489 | 0.69768439 |
| 5 | 0.25166896 | 0.05001344 | 0.69831761 |
| 6 | 0.25157625 | 0.05001187 | 0.69841188 |
| 7 | 0.25156244 | 0.05001164 | 0.69842592 |
| 8 | 0.25156039 | 0.05001160 | 0.69842801 |
| 9 | 0.25156009 | 0.05001160 | 0.69842832 |
| 10 | 0.25156004 | 0.05001160 | 0.69842837 |

## 1.2. Résultat préliminaire

Soit $\mathbf{y}$ une variable aléatoire $(N \times 1)$ dont la densité notée $g(\mathbf{y}|\boldsymbol{\phi})$ dépend du vecteur de paramètres $\boldsymbol{\phi} \in \Phi$ et $\mathbf{z}$ un vecteur de variables aléatoires auxiliaires, qualifiées de données manquantes [1], et ayant avec $\mathbf{y}$ une densité conjointe notée $f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})$ dépendant elle aussi de $\boldsymbol{\phi}$. Dans ces conditions très générales, on peut établir le résultat suivant, connu sous le nom d'identité de Fisher (1925), cité par Efron (1977), (*cf.* annexe A) :

$$\boxed{\frac{\partial \ln g(\mathbf{y}|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \mathrm{E}_C\left[\frac{\partial \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right]} \tag{7}$$

formule qui traduit simplement le fait que la dérivée de la logvraisemblance $\mathrm{L}(\boldsymbol{\phi}; \mathbf{y}) = \ln g(\mathbf{y}|\boldsymbol{\phi})$ de $\boldsymbol{\phi}$ basée sur $\mathbf{y}$ par rapport au paramètre est l'espérance conditionnelle de la dérivée de la logvraisemblance $\mathrm{L}(\boldsymbol{\phi}; \mathbf{x}) = \ln f(\mathbf{x}|\boldsymbol{\phi})$ des données dites augmentées $(\mathbf{x} = (\mathbf{y}', \mathbf{z}')')$. Cette espérance, notée $\mathrm{E}_C(\cdot)$, est prise par rapport à la distribution conditionnelle des données supplémentaires $\mathbf{z}$ sachant les données observées $\mathbf{y}$ et le paramètre $\boldsymbol{\phi}$.

Ce résultat étant acquis, admettons qu'on veuille résoudre par un procédé itératif l'équation :

$$\frac{\partial \mathrm{L}(\boldsymbol{\phi}; \mathbf{y})}{\partial \boldsymbol{\phi}} = 0, \tag{8}$$

ainsi qu'on est conduit classiquement à le faire en vue de l'obtention des estimations du maximum de vraisemblance.

On dispose donc à l'itération $[t]$ d'une valeur courante $\boldsymbol{\phi}^{[t]}$ du paramètre ; si l'on fait appel au résultat précédent en (7), on va s'intéresser à l'espérance conditionnelle de $\partial[\ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})]/\partial\boldsymbol{\phi}$ par rapport à la densité de $\mathbf{z}|\mathbf{y}$, $\boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}$ qu'on note $\mathrm{E}_C^{[t]}\left[\dfrac{\partial \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right]$. Cette espérance s'écrit :

$$\mathrm{E}_C^{[t]}\left[\frac{\partial \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right] = \int_Z \frac{\partial \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\, h(\mathbf{z}|\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})\mathrm{d}\mathbf{z}$$

où Z désigne l'espace d'échantillonnage de $\mathbf{z}$ et $h(\mathbf{z}|\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})$ est la densité de la loi conditionnelle des données manquantes $\mathbf{z}$ sachant $\mathbf{y}$ et $\boldsymbol{\phi} = \boldsymbol{\phi}^{(t)}$. Cette précision sera omise par la suite pour simplifier la notation, le domaine d'intégration étant implicitement spécifié par le symbole différentiel correspondant sous le signe somme, ici $\mathrm{d}\mathbf{z}$.

Comme $h(\mathbf{z}|\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})$ ne dépend pas de $\boldsymbol{\phi}$, on peut sortir l'opérateur de dérivation d'où

$$\boxed{\mathrm{E}_C^{[t]}\left[\frac{\partial \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right] = \frac{\partial}{\partial \boldsymbol{\phi}}\left\{\mathrm{E}_C^{[t]}\left[\ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})\right]\right\}.} \tag{9}$$

1. ou variables latentes, supplémentaires ou cachées selon les circonstances et les auteurs.

La résolution itérative de (8) peut donc se ramener à celle de l'équation

$$\frac{\partial}{\partial \boldsymbol{\phi}} \left\{ \mathrm{E}_C^{[t]} \left[ \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi}) \right] \right\} = \mathbf{0}, \tag{10}$$

qu'avaient mentionnée Foulley *et al.* (1987) et Foulley (1993) à propos de l'estimation du maximum de vraisemblance des composantes de la variance dans un modèle linéaire mixte. En fait, la simple lecture de cette équation préfigure la description de l'algorithme EM et de ses deux étapes.

Le terme $\ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi})$ représente la logvraisemblance des données augmentées (dites aussi « complètes » dans la terminologie de Dempster, Laird et Rubin). $\mathrm{E}_C^{[t]} \left[ \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\phi}) \right]$ désigne l'espérance conditionnelle de cette logvraisemblance par rapport à la densité des données supplémentaires $\mathbf{z}$ (ou « manquantes ») sachant les données observées $\mathbf{y}$ (ou « incomplètes » selon Dempster, Laird et Rubin) et la valeur courante $\boldsymbol{\phi}^{[t]}$ du paramètre. C'est donc une fonction de $\mathbf{y}$, $\boldsymbol{\phi}^{[t]}$ et du paramètre $\boldsymbol{\phi}$ que Dempster, Laird et Rubin notent $\mathrm{Q}(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})$ et son établissement correspond précisément à l'étape E (dite « Expectation ») de l'algorithme. L'annulation de sa dérivée première $\frac{\partial}{\partial \boldsymbol{\phi}} \left[ \mathrm{Q}(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) \right] = 0$ correspond à la phase de recherche de l'extremum : c'est l'étape dite M « Maximisation » de l'algorithme.

### 1.3. Formulation de l'algorithme

Dans la présentation de Dempster, Laird et Rubin, on oppose les données dites incomplètes représentées par la variable aléatoire $\mathbf{y}$ de densité $g(\mathbf{y}|\boldsymbol{\phi})$ aux données dites complètes $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$ formées de la concaténation des données incomplètes $\mathbf{y}$ et des données manquantes $\mathbf{z}$ et de densité $f(\mathbf{x}|\boldsymbol{\phi})$. Aux variables aléatoires $\mathbf{x}$ et $\mathbf{y}$ correspondent respectivement les espaces d'échantillonnage $\mathcal{X}$ et $\mathcal{Y}$ qui sont liés entre eux par une application de $\mathcal{X}$ dans $\mathcal{Y}$. Comme l'on n'observe pas $\mathbf{x} \in \mathcal{X}$, mais seulement $\mathbf{y} = \mathbf{y}(\mathbf{x}) \in \mathcal{Y}$, on peut spécifier de façon générale la relation entre les deux types de variables (complètes et incomplètes) par :

$$g(\mathbf{y}|\boldsymbol{\phi}) = \int_{\mathcal{X}_y} f(\mathbf{x}|\boldsymbol{\phi}) \mathrm{d}\mathbf{x}, \tag{11}$$

où $\mathcal{X}_y$ est un sous-espace observable de $\mathcal{X}$ défini par l'équation $\mathbf{y} = \mathbf{y}(\mathbf{x})$ (espace dit antécédent de $\mathcal{Y}$), soit

$$\mathcal{X}_y = \{\mathbf{x} \in \mathcal{X}; \mathbf{y} = \mathbf{y}(\mathbf{x})\} \subset \mathcal{X}. \tag{12}$$

Pour illustrer cette notion un peu abstraite, on peut prendre l'exemple du modèle dit « animal » des généticiens quantitatifs le plus simple : $\mathbf{y} = \mu\mathbf{1} + \mathbf{a} + \mathbf{e}$ où $\mathbf{a} = \{a_i\} \sim (0, \mathbf{A}\sigma_a^2)$ est le vecteur des effets génétiques additifs des individus indicés par $i$ ($\mathbf{A}$ étant la matrice de parenté) et $\mathbf{e} = \{e_i\} \sim (0, \mathbf{I}\sigma_e^2)$ est celui des effets génétiques non additifs et des effets environnementaux. Dans ce cas, on pourra définir les données complètes

directement par $\mathbf{x} = (\mathbf{a}', \mathbf{e}')'$ et on a $g(\mathbf{y}|\mu, \sigma_a^2, \sigma_e^2) = \int_{X_y} f(\mathbf{a}, \mathbf{e}|\mu, \sigma_a^2, \sigma_e^2)\mathbf{dade}$ avec $X_y = \{X; \mathbf{a} + \mathbf{e} = \mathbf{y} - \mathbf{1}\mu\}$. On peut aussi, plus classiquement, définir les données complètes sous la forme $\mathbf{x} = (\mathbf{y}', \mathbf{a}')'$ ou $\mathbf{x} = (\mathbf{y}', \mathbf{e}')'$.

Dans son acception générale, l'algorithme EM se définit par les deux phases suivantes.

*1) Phase E dite « Expectation» (ou Espérance)*

Sachant la valeur courante du paramètre $\boldsymbol{\phi}^{[t]}$ à l'itération $[t]$, la phase **E** consiste en la détermination de la fonction

$$\boxed{Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = \mathrm{E}_C^{[t]}[L(\boldsymbol{\phi}; \mathbf{x})]}. \tag{13}$$

Avec $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$, $Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})$ est l'espérance conditionnelle de la logvraisemblance des données complètes par rapport à la distribution des données manquantes $\mathbf{z}$ sachant les données incomplètes $\mathbf{y}$ et la valeur courante $\boldsymbol{\phi}^{[t]}$ du paramètre soit

$$Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = \int L(\boldsymbol{\phi}; \mathbf{y}, \mathbf{z}) h(\mathbf{z}|\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})\mathrm{d}\mathbf{z}. \tag{14a}$$

Avec une spécification générale des données complètes, cette fonction s'écrit

$$Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = \int L(\boldsymbol{\phi}; \mathbf{x}) k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})\mathrm{d}\mathbf{x}, \tag{14b}$$

où

$$k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}) = f(\mathbf{x}|\boldsymbol{\phi})/g(\mathbf{y}|\boldsymbol{\phi}). \tag{15}$$

*2) Phase M dite « Maximisation»*

On actualise la valeur courante du paramètre en maximisant la fonction obtenue à la phase E par rapport à $\boldsymbol{\phi}$, soit

$$\boxed{\boldsymbol{\phi}^{[t+1]} = \mathrm{argmax}_{\Phi}\, Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})}. \tag{16}$$

Il existe une version généralisée de l'algorithme dite GEM dans laquelle la valeur actualisée ne maximise pas nécessairement $Q$ mais l'augmente simplement c'est-à-dire satisfait $Q(\boldsymbol{\phi}^{[t+1]}; \boldsymbol{\phi}^{[t]}) \geqslant Q(\boldsymbol{\phi}^{[t]}; \boldsymbol{\phi}^{[t]})$, $\forall t$.

### 1.4. Cas d'un mélange gaussien

Un exemple particulièrement illustratif des potentialités de l'algorithme EM réside dans son application au cas d'un mélange de distributions (Dempster *et al.*, 1977; Titterington *et al.*, 1985; Celeux et Diebolt, 1985; McLachlan et Basford, 1985; McLachlan et Peel, 2000).

Pour simplifier, nous considérerons le cas d'un mélange d'un nombre fixé de lois gaussiennes univariées $\mathcal{N}(\mu_j, \sigma_j^2)$ d'espérance $\mu_j$ et de variance $\sigma_j^2$ en proportion $p_j$ pour chacune des composantes $j = 1, \ldots, J$ du mélange.

Soit $\mathbf{y}_{N\times 1} = \{y_i\}$ le vecteur des $N$ observations $y_i$ supposées indépendantes et de densité

$$f_{Y_i}(y; \boldsymbol{\Phi}) = \sum_{j=1}^{J} p_j f_j(y; \boldsymbol{\theta}_j) \tag{17}$$

où $\mathbf{p}_{J\times 1} = \{p_j\}$, $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2)'$, $\boldsymbol{\Phi} = (\mathbf{p}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_j', \ldots, \boldsymbol{\theta}_J')'$ représentent les paramètres et $f_j(y; \boldsymbol{\theta}_j)$ est la densité de la loi $\mathcal{N}(\mu_j, \sigma_j^2)$ relative à la composante $j$ du mélange.

Compte tenu de (17) et de l'indépendance des observations, la logvraisemblance des données observées s'écrit :

$$L(\boldsymbol{\Phi}; \mathbf{y}) = \sum_{i=1}^{N} \ln \left[ \sum_{j=1}^{J} p_j f_j(y_i; \boldsymbol{\theta}_j) \right], \tag{18}$$

expression qui ne se prête pas aisément à la maximisation.

Une façon de contourner cette difficulté est d'avoir recours à l'algorithme EM. On introduit alors des variables $z_i$ non observables indiquant l'appartenance de l'observation $i$ à une certaine composante $j$ du mélange et donc telle que $\Pr(z_i = j) = p_j$. Par définition, cette appartenance étant exclusive, la densité $g(x_i; \boldsymbol{\Phi})$ du couple $x_i = (y_i, z_i)'$ peut alors s'écrire

$$g(x_i; \boldsymbol{\Phi}) = \prod_{j=1}^{J} [g(y_i, z_i = j; \boldsymbol{\Phi})]^{a_{ij}}, \tag{19a}$$

($a_{ij}$ désignant l'indicatrice $a_{ij} = I_{[z_j=j]}$), soit encore, en décomposant la loi conjointe de $y_i$ et $z_i$,

$$g(x_i; \boldsymbol{\Phi}) = \prod_{j=1}^{J} [p_j f_j(y_i; \boldsymbol{\theta}_j)]^{a_{ij}}. \tag{19b}$$

Les couples $x_i$ étant indépendants entre eux, la densité des données complètes $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_i', \ldots, \mathbf{x}_N')'$ est le produit des densités élementaires soit

$$g(\mathbf{x}; \boldsymbol{\Phi}) = \prod_{i=1}^{N} \prod_{j=1}^{J} [p_j f_j(y_i; \boldsymbol{\theta}_j)]^{a_{ij}}. \tag{20}$$

On en déduit immédiatement l'expression de la logvraisemblance correspondante

$$L(\boldsymbol{\Phi}; \mathbf{x}) = \sum_{i=1}^{N} \sum_{j=1}^{J} a_{ij} \left[ \ln p_j + \ln f_i(y_i; \boldsymbol{\theta}_j) \right].$$

En prenant l'espérance de $L(\boldsymbol{\phi};\mathbf{x})$ par rapport à la distribution des données manquantes $a_{ij}$ sachant les données observées et les paramètres pris à leurs valeurs courantes, on obtient l'expresion de la fonction $Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})$ à la phase E

$$\boxed{Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) = \sum_{i=1}^{N}\sum_{j=1}^{J}\alpha_{ij}^{[t]}\left[\ln p_j + \ln f_j(y_i;\boldsymbol{\theta}_j)\right]}, \tag{21}$$

où $\alpha_{ij}^{[t]} = \mathrm{E}\left(a_{ij}|y_i, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}\right)$ s'interprète comme la probabilité conditionnelle d'appartenance de l'observation $i$ à la composante $j$ du mélange, soit

$$\alpha_{ij}^{[t]} = \Pr(z_i = j|y_i, \boldsymbol{\phi}) = \frac{p_j^{[t]} f_j(y_i;\theta_j^{[t]})}{\sum_{j=1}^{J} p_j^{[t]} f_j(y_i;\boldsymbol{\theta}_j^{[t]})}. \tag{22}$$

Il ne reste plus maintenant (phase M) qu'à maximiser la fonction $Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})$ par rapport à $\boldsymbol{\phi}$, ou plus précisèment $Q^{\#}(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) = Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) - \lambda(\sum_{j=1}^{J} p_j - 1)$ pour prendre en compte, grâce au multiplicateur de Lagrange $\lambda$, la relation d'exhaustivité qui lie les probabilités d'appartenance. Les dérivées partielles s'écrivent :

$$\frac{\partial Q^{\#}}{\partial p_j} = \sum_{i=1}^{N}\frac{\alpha_{ij}^{[t]}}{p_j} - \lambda; \quad \frac{\partial Q^{\#}}{\partial \mu_j} = \sum_{i=1}^{N}\alpha_{ij}^{[t]}(y_i - \mu_j)/\sigma_j^2;$$

$$\frac{\partial Q^{\#}}{\partial \sigma_j^2} = -1/2\left\{\sum_{i=1}^{N}\alpha_{ij}^{[t]}\left[\frac{1}{\sigma_j^2} - \frac{(y_i-\mu_j)^2}{\sigma_j^4}\right]\right\}$$

Par annulation, on obtient les solutions à savoir

$$p_j^{[t+1]} = \left(\sum_{i=1}^{N}\alpha_{ij}^{[t]}\right)\Big/N, \tag{23a}$$

$$\mu_j^{[t+1]} = \left(\sum_{i=1}^{N}\alpha_{ij}^{[t]}y_i\right)\Big/\left(\sum_{i=1}^{N}\alpha_{ij}^{[t]}\right) \tag{23b}$$

$$\sigma_j^{2[t+1]} = \left[\sum_{i=1}^{N}\alpha_{ij}^{[t]}(y_i - \mu_j^{[t+1]})^2\right]\Big/\left(\sum_{i=1}^{N}\alpha_{ij}^{[t]}\right). \tag{23c}$$

Les résultats précédents se généralisent sans problème à la situation multivariée $\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Ici on s'est placé dans la situation où les observations $y_i$ étaient indépendantes, mais cette hypothèse peut être levée. Grimaud *et al.* (2002) ont ainsi traité un modèle de mélange de deux modèles mixtes gaussiens.

En définitive, le traitement d'un mélange par l'algorithme EM rentre dans un cadre très général qui est mis à profit dans mainte application. Citons à titre d'exemple la recherche et la localisation de loci à effets quantitatifs (dits QTL en anglais) utilisant des marqueurs moléculaires dans des dispositifs de croisement (backcross par ex). Dans ce cas, les composantes du mélange sont les génotypes possibles au QTL putatif et les probabilités d'appartenance *a priori* sont données par les règles de ségrégation sachant l'ascendance et l'information procurée par les marqueurs moléculaires (Wu *et al.*, 2002). Dans ce genre de problème, l'algorithme EM a permis de substituer à l'expression classique de la vraisemblance (18) une forme plus aisée à maximiser (21) par le biais de la prise en compte d'informations cachées. Le traitement des distributions de mélange par l'algorithme EM est également à la base de certaines techniques de classification, *cf.* par exemple l'algorithme CEM (C pour classification) (Celeux et Govaert, 1992).

## 1.5. Cas particuliers

### *1.5.1. Famille exponentielle régulière*

On considère ici le cas où la distribution des données complètes appartient à la famille exponentielle régulière qu'on peut mettre sous la forme générale suivante :

$$f(\mathbf{x}|\boldsymbol{\phi}) = b(\mathbf{x}) \exp[\boldsymbol{\phi}'\mathbf{t}(\mathbf{x})]/a(\boldsymbol{\phi}), \tag{24}$$

où $\boldsymbol{\phi}$ est le vecteur $(k \times 1)$ des paramètres dits canoniques, $\mathbf{t}(\mathbf{x})$ le vecteur $(k \times 1)$ de la statistique exhaustive correspondante, et $a(\boldsymbol{\phi})$ et $b(\mathbf{x})$ des fonctions scalaires.

La statistique exhaustive $\mathbf{t}(\mathbf{x})$ du paramètre canonique $\boldsymbol{\phi}$ se caractérise par

$$\mathrm{E}[\mathbf{t}(\mathbf{x})|\boldsymbol{\phi}] = \partial \ln[a(\boldsymbol{\phi})]/\partial\boldsymbol{\phi}, \tag{25a}$$

$$\mathrm{Var}[\mathbf{t}(\mathbf{x})|\boldsymbol{\phi}] = \partial^2 \ln[a(\boldsymbol{\phi})]/\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'. \tag{25b}$$

Eu égard à la forme de la densité en (24), la phase E conduit à la fonction $Q$ suivante

$$Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = \boldsymbol{\phi}'\mathrm{E}_C^{[t]}[\mathbf{t}(\mathbf{x})] - \ln[a(\boldsymbol{\phi})] + cste. \tag{26}$$

Par annulation de la dérivée de Q, on obtient à la phase M l'équation suivante

$$\mathrm{E}[\mathbf{t}(\mathbf{x})] = \mathrm{E}_C^{[t]}[\mathbf{t}(\mathbf{x})],$$

que l'on peut écrire aussi, à l'instar de Dempster, Laird et Rubin, sous la forme

$$\boxed{\mathrm{E}\left[\mathbf{t}(\mathbf{x})|\boldsymbol{\phi}^{[t+1]}\right] = \mathrm{E}\left[\mathbf{t}(\mathbf{x})|\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}\right]}, \tag{27}$$

qui apparaît comme l'équation clé de l'algorithme EM dans la famille exponentielle.

Si cette équation a une solution dans l'espace des paramètres $\Phi$, elle est unique, puisque, dans la famille exponentielle régulière, moins deux fois la logvraisemblance est une fonction convexe.

L'exemple précédent de l'estimation de la fréquence allélique au locus de groupe sanguin humain ABO fournit une très bonne illustration de cette propriété. Une statistique exhaustive des fréquences alléliques de $p$ et $q$ consiste, à effectif total $N = y_+$ fixé, en les nombres d'allèles respectifs soit $t_A = 2x_{AA} + x_{AB} + x_{AO}$ et $t_B = 2x_{BB} + x_{AB} + x_{BO}$. À la phase M, on résout l'équation (27) $\mathrm{E}(t_A|p^{[t+1]}, q^{[t+1]}) = \mathrm{E}(t_A|\mathbf{y}, p^{[t]}, q^{[t]})$ soit

$$2Np^{[t+1]} = 2\mathrm{E}(x_{AA}|\mathbf{y}, p^{[t]}, q^{[t]}) + y_{AB} + \mathrm{E}(x_{AO}|\mathbf{y}, p^{[t]}, q^{[t]}), \qquad (28)$$

avec

$$\mathrm{E}(x_{AA}|y, p^{[t]}, q^{[t]}) = \frac{p^{[t]}}{p^{[t]} + 2r^{[t]}}\, y_A, \qquad (29)$$

puisque, conditionnellement à $y_A$, $x_{AA}$ a une distribution binomiale de paramètres $y_A$ et $p/(p+2r)$. On fait de même pour $q^{[t+1]}$. On retrouve ainsi les expressions (5) et (6) établies empiriquement au début.

Une autre illustration consiste en l'estimation des composantes de la variance dans le modèle linéaire mixte gaussien comme nous le verrons dans la deuxième partie de ce chapitre.

*1.5.2. Mode* a posteriori

L'algorithme EM peut être également utilisé dans un cadre bayésien en vue de l'obtention du mode de la distribution *a posteriori* $p(\boldsymbol{\Phi}|\mathbf{y})$. Il existe pour la logdensité *a posteriori* l'homologue de la formule (7) pour la logvraisemblance,

$$\boxed{\frac{\partial \ln p(\boldsymbol{\Phi}|\mathbf{y})}{\partial \boldsymbol{\Phi}} = \mathrm{E}_C\left[\frac{\partial \ln p(\boldsymbol{\Phi}|\mathbf{y}, \mathbf{z})}{\partial \boldsymbol{\Phi}}\right]}, \qquad (30)$$

où $\mathrm{E}_C(\cdot)$ indique comme précédemment une espérance conditionnelle prise par rapport à $\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi}$.

Sur cette base on déduit immédiatement les deux phases de l'algorithme EM correspondant au calcul du mode *a posteriori* de $\boldsymbol{\Phi}$.

Sachant la valeur courante du paramètre $\boldsymbol{\Phi}^{[t]}$ à l'itération $[t]$, la phase E consiste en la spécification de la fonction

$$Q^*(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]}) = \mathrm{E}_C^{[t]}\left[\ln p(\boldsymbol{\Phi}|\mathbf{y}, \mathbf{z})\right], \qquad (31)$$

qui, du fait du théorème de Bayes $p(\boldsymbol{\Phi}|\mathbf{y}, \mathbf{z}) \propto p(\mathbf{y}, \mathbf{z}|\boldsymbol{\Phi})p(\boldsymbol{\Phi})$, se réduit à

$$Q^*(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]}) = Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]}) + \ln p(\boldsymbol{\Phi}) + Cste, \qquad (32)$$

où $Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]})$ est défini comme précédemment : *cf.* (13) et (14ab).

À la Phase M, on actualise la valeur courante de $\boldsymbol{\Phi}$ en recherchant $\boldsymbol{\Phi}^{[t+1]}$ qui maximise la fonction $Q^*(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]})$ par rapport à $\boldsymbol{\Phi}$, soit $\boldsymbol{\Phi}^{[t+1]} = \mathrm{argmax}_{\boldsymbol{\Phi}} Q^*(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]})$.

## 1.6. Quelques propriétés

### *1.6.1. Accroissement monotone de la vraisemblance*

Soit une suite d'itérations EM : $\boldsymbol{\Phi}^{[0]}, \boldsymbol{\Phi}^{[1]}, \boldsymbol{\Phi}^{[2]}, \ldots, \boldsymbol{\Phi}^{[t]}, \boldsymbol{\Phi}^{[t+1]}, \ldots$, on peut établir le théorème suivant :

$$\boxed{L(\boldsymbol{\Phi}^{[t+1]}; \mathbf{y}) \geqslant L(\boldsymbol{\Phi}^{[t]}; \mathbf{y}),\ \forall t}, \tag{33}$$

l'égalité n'intervenant que, si et seulement si, à partir d'un certain rang, $Q(\boldsymbol{\Phi}^{[t+1]}; \boldsymbol{\Phi}^{[t]}) = Q(\boldsymbol{\Phi}^{[t]}; \boldsymbol{\Phi}^{[t]})$ et $h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi}^{[t+1]}) = h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi}^{[t]})$ ou $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\Phi}^{[t+1]}) = k(\mathbf{x}|\mathbf{y}, \boldsymbol{\Phi}^{[t]})$.

C'est une propriété fondamentale de l'algorithme qui garantit à l'utilisateur une bonne évolution des valeurs de la logvraisemblance.

La démonstration est intéressante pour éclairer la compréhension des mécanismes sous-jacents à EM. Elle se décline comme suit.

Par définition de la densité conjointe, on a : $f(\mathbf{y}, \mathbf{z}|\boldsymbol{\Phi}) = g(\mathbf{y}|\boldsymbol{\Phi})h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi})$ et, en passant aux logarithmes, $\ln g(\mathbf{y}|\boldsymbol{\Phi}) = \ln f(\mathbf{y}, \mathbf{z}|\boldsymbol{\Phi}) - \ln h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi})$. Si l'on intègre les deux membres par rapport à la densité de $\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi}^{(t)}$, il vient :

$$L(\boldsymbol{\Phi}; \mathbf{y}) = Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]}) - H(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]}), \tag{34}$$

où

$$H(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(t)}) = \int \ln[h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi})]h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi} = \boldsymbol{\Phi}^{[t]})\mathrm{d}\mathbf{z}. \tag{35}$$

Exprimons maintenant la variation de la logvraisemblance $L(\boldsymbol{\Phi}^{[t+1]}; \mathbf{y}) - L(\boldsymbol{\Phi}^{[t]}; \mathbf{y})$ quand on passe d'une itération EM à la suivante. Compte tenu de (34), cette variation s'écrit :

$$\begin{aligned} &L(\boldsymbol{\Phi}^{[t+1]}; \mathbf{y}) - L(\boldsymbol{\Phi}^{[t]}; \mathbf{y}) \\ &= \left[Q(\boldsymbol{\Phi}^{[t+1]}; \boldsymbol{\Phi}^{[t]}) - Q(\boldsymbol{\Phi}^{[t]}; \boldsymbol{\Phi}^{[t]})\right] - \left[H(\boldsymbol{\Phi}^{[t+1]}; \boldsymbol{\Phi}^{[t]}) - H(\boldsymbol{\Phi}^{[t]}; \boldsymbol{\Phi}^{[t]})\right] \end{aligned} \tag{36}$$

Par définition de la phase M de l'algorithme, la quantité $Q(\boldsymbol{\Phi}^{[t+1]}; \boldsymbol{\Phi}^{[t]}) - Q(\boldsymbol{\Phi}^{[t]}; \boldsymbol{\Phi}^{[t]})$ est positive ou nulle qu'il s'agisse d'un EM classique ou généralisé. Quant au deuxième terme, considérons la quantité $H(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]}) - H(\boldsymbol{\Phi}^{[t]}; \boldsymbol{\Phi}^{[t]})$ comme une fonction de $\boldsymbol{\Phi}$ ; elle s'écrit, au vu de la définition donnée en (35) :

$$H(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{[t]}) - H(\boldsymbol{\Phi}^{[t]}; \boldsymbol{\Phi}^{[t]}) = \int \ln\left[\frac{h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi})}{h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi} = \boldsymbol{\Phi}^{[t]})}\right] h(\mathbf{z}|\mathbf{y}, \boldsymbol{\Phi} = \boldsymbol{\Phi}^{[t]})\mathrm{d}\mathbf{z}. \tag{37}$$

Le logarithme étant une fonction concave, on peut majorer cette quantité par application de l'inégalité de Jensen [2].

$$\int \ln\left[\frac{h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi})}{h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]})}\right] h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]})\mathrm{d}\mathbf{z}$$

$$\leqslant \ln \int \frac{h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi})}{h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]})}\, h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]})\mathrm{d}\mathbf{z} = 0,$$

l'égalité ne se produisant que si $h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi}) = h(\mathbf{z}|\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]})$, $\forall\boldsymbol{\phi}$ (*cf.* Rao, 1973, page 59, formule 1e6.6) d'où

$$H(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) - H(\boldsymbol{\phi}^{[t]};\boldsymbol{\phi}^{[t]}) \leqslant 0,\ \forall\boldsymbol{\phi}, \tag{38}$$

ce qui établit le théorème de départ (33).

Remarquons que l'on aurait pu faire la même démonstration en partant de la relation $\ln[g(\mathbf{y}|\boldsymbol{\phi})] = \ln[f(\mathbf{x}|\boldsymbol{\phi})] - \ln[k(\mathbf{x}|\mathbf{y},\boldsymbol{\phi})]$ (*cf.* 15), $H$ étant définie alors par

$$H(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) = \int \ln[k(\mathbf{x}|\mathbf{y},\boldsymbol{\phi})]\, k(\mathbf{x}|\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]})\mathrm{d}\mathbf{z}. \tag{39}$$

*1.6.2. Cohérence interne*

Si $\boldsymbol{\phi}^*$ est un point stationnaire de $L(\boldsymbol{\phi};\mathbf{y})$, il annule aussi la dérivée de $Q(\boldsymbol{\phi};\boldsymbol{\phi}^*)$ par rapport à $\boldsymbol{\phi}$ et réciproquement :

$$\left.\frac{\partial L(\boldsymbol{\phi};\mathbf{y})}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} = \mathbf{0} \Longleftrightarrow \left.\frac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}^*)}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} = \mathbf{0}, \tag{40}$$

Ce théorème découle d'un corollaire de (7) et (9). En effet, par définition $Q(\boldsymbol{\phi};\boldsymbol{\phi}_0) = \mathrm{E}_C^0[L(\boldsymbol{\phi};\mathbf{x})]$ où $\mathrm{E}_C^0(\cdot)$ indique une espérance conditionnelle prise par rapport à la distribution de $\mathbf{z}|\mathbf{y}$, $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ et $\dfrac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}_0)}{\partial\boldsymbol{\phi}} = \dfrac{\partial}{\partial\boldsymbol{\phi}}\{\mathrm{E}_C^0[L(\boldsymbol{\phi};\mathbf{x})]\}$. Du fait de l'égalité (9), on peut intervertir les opérateurs de dérivation et d'espérance si bien que $\dfrac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}_0)}{\partial\boldsymbol{\phi}} = \mathrm{E}_C^0\left[\dfrac{\partial L(\boldsymbol{\phi};\mathbf{x})}{\partial\boldsymbol{\phi}}\right]$ et en évaluant ces deux fonctions de $\boldsymbol{\phi}$ au point $\boldsymbol{\phi}_0$, il vient compte tenu de (7)

$$\left.\frac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}_0)}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} = \left.\frac{\partial L(\boldsymbol{\phi};\mathbf{y})}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} \tag{41}$$

Le théorème (40) en découle par application de (41) à $\boldsymbol{\phi}_0 = \boldsymbol{\phi}^*$ point stationnaire de $L(\boldsymbol{\phi};\mathbf{y})$. Cette propriété de cohérence interne, dite de « self-consistency » dans le monde anglo-saxon, remonterait à Fisher et aurait fait

2. Si $X$ est une variable aléatoire d'espérance $\mu$ et si $f(x)$ est une fonction concave, alors $E[f(X)] \leqslant f(\mu)$.

l'objet de nombreuses redécouvertes depuis les années 1930. Remarquons que, du fait de (34), la propriété (41) implique

$$\left.\frac{\partial H(\boldsymbol{\phi};\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} = \mathbf{0}. \tag{42}$$

McLachlan et Krishnan (1997, page 85) établissent tout d'abord (42) à partir de (38) et en déduisent (41) et (40). Quoiqu'il en soit, ce résultat est fondamental pour établir les propriétés de convergence des itérations EM vers un point stationnaire de $L(\boldsymbol{\phi};\mathbf{y})$.

*1.6.3. Convergence vers un point stationnaire*

La question de la convergence de l'algorithme fait l'objet de plusieurs théorèmes correspondant aux différentes conditions qui sous-tendent cette propriété. Nous ne rentrerons pas dans tous ces développements, certes importants, mais d'accès difficile. Le lecteur est renvoyé à l'ouvrage de McLachlan and Krishan (1997) ainsi que, pour plus de détails, à l'article de Wu (1983). Nous nous restreindrons aux deux résultats suivants.

On note $\mathcal{L}(L_0) = \{\boldsymbol{\phi} \in \Phi; L(\boldsymbol{\phi};\mathbf{y}) = L_0\}$ le sous-ensemble de $\Phi$ dont les éléments ont pour logvraisemblance $L(\boldsymbol{\phi};\mathbf{y})$ une valeur donnée $L_0$.

**Théorème.** Soit une suite d'itérations EM ou GEM : $\boldsymbol{\phi}^{[0]}, \boldsymbol{\phi}^{[1]}, \boldsymbol{\phi}^{[2]}, \ldots, \boldsymbol{\phi}^{[t]}, \boldsymbol{\phi}^{[t+1]}$, qui vérifie la condition $\left.\dfrac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{[t+1]}} = \mathbf{0}$. Lorsque la fonction $\dfrac{\partial Q(\boldsymbol{\phi};\boldsymbol{\Psi})}{\partial \boldsymbol{\phi}}$ est continue en $\boldsymbol{\phi}$ et $\boldsymbol{\Psi}$, alors $\boldsymbol{\phi}^{[t]} \longrightarrow \boldsymbol{\phi}^*$ quand $t \longrightarrow +\infty$ où $\boldsymbol{\phi}^*$ est un point stationnaire (il vérifie $L'(\boldsymbol{\phi}^*;\mathbf{y}) = \mathbf{0}$) qui est tel que $L(\boldsymbol{\phi}^*;\mathbf{y}) = L^* = \operatorname{Lim} L(\boldsymbol{\phi}^{[t]};\mathbf{y})$ si l'une ou l'autre des conditions suivantes est remplie :

a) $\mathcal{L}(L^*)$ est un singleton où $\mathcal{L}(L_0) = \{\boldsymbol{\phi} \in \Phi : L(\boldsymbol{\phi};\mathbf{y}) = L_0\}$ ;

b) $\mathcal{L}(L^*)$ n'est pas un singleton mais est fini et $||\boldsymbol{\phi}^{[t+1]} - \boldsymbol{\phi}^{[t]}|| \longrightarrow 0$, quand $t \longrightarrow +\infty$.

La démonstration dans le cas b) repose sur le raisonnement suivant. Eu égard aux conditions de régularité, $L(\boldsymbol{\phi}^{[t]};\mathbf{y})$ converge vers une valeur $L^*$, le point limite $\boldsymbol{\phi}^*$ (du fait de $||\boldsymbol{\phi}^{[t+1]} - \boldsymbol{\phi}^{[t]}|| \longrightarrow 0$) correspondant dans $\mathcal{L}(L^*)$ va vérifier

$$\left.\frac{\partial L(\boldsymbol{\phi};\mathbf{y})}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} = \left.\frac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} = \operatorname*{Lim}_{t\to\infty}\left[\left.\frac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{[t+1]}}\right] = \mathbf{0}.$$

Ce théorème ne garantit donc pas la convergence vers un maximum global de la logvraisemblance $L(\boldsymbol{\phi};\mathbf{y})$. Si $L(\boldsymbol{\phi};\mathbf{y})$ a plusieurs points stationnaires, la convergence d'une suite d'itérations EM vers l'un d'entre eux (maximum local

ou global ou point selle) va dépendre de la valeur de départ. Quand le point stationnaire est un point selle, une très petite perturbation de cette valeur va détourner la suite des itérations EM du point selle.

Il est à remarquer que la convergence de $L(\boldsymbol{\Phi};\mathbf{y})$ vers $L^*$ n'implique pas automatiquement celle de $\boldsymbol{\Phi}^{[t]}$ vers $\boldsymbol{\Phi}^*$ ; il faut certaines conditions à cet effet comme la condition de continuité de la fonction $[\partial Q(\boldsymbol{\Phi};\boldsymbol{\Psi}]/\partial\boldsymbol{\Phi}$. Ainsi, Boyles (1983) décrit un exemple de convergence d'un GEM non pas vers un seul point mais vers les points d'un cercle.

**Corollaire.** Il a trait au cas où la fonction $L(\boldsymbol{\Phi};\mathbf{y})$ est unimodale avec un seul point stationnaire $\boldsymbol{\Phi}^*$ à l'intérieur de $\Phi$. On est donc dans le cas a) d'un singleton et, si la fonction $\dfrac{\partial Q(\boldsymbol{\Phi};\boldsymbol{\Psi})}{\partial\boldsymbol{\Phi}}$ est continue en $\boldsymbol{\Phi}$ et en $\boldsymbol{\Psi}$, toute suite d'itérations EM quelle que soit la valeur de départ converge vers l'unique maximum global de $L(\boldsymbol{\Phi};\mathbf{y})$.

*1.6.4. Partition de l'information*

On a montré que $f(\mathbf{x}|\boldsymbol{\Phi}) = g(\mathbf{y}|\boldsymbol{\Phi})k(\mathbf{x}|\mathbf{y},\boldsymbol{\Phi})$ (*cf.* 15). En passant au logarithme et en dérivant deux fois par rapport à $\boldsymbol{\Phi}$, il vient

$$-\frac{\partial^2 \ln g(\mathbf{y}|\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'} = -\frac{\partial^2 \ln f(\mathbf{x}|\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'} + \frac{\partial^2 \ln k(\mathbf{x}|\mathbf{y},\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}.$$

Le deuxième membre fait intervenir les données manquantes. Pour évaluer sa contribution réelle, nous en prendrons l'espérance par rapport à la distribution conditionnelle de ces données $\mathbf{z}$ sachant $\mathbf{y}$ et $\boldsymbol{\Phi}$, notée comme précédemment $\mathrm{E}_C(\cdot)$, d'où

$$-\frac{\partial^2 L(\boldsymbol{\Phi};\mathbf{y})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'} = -\mathrm{E}_C\left[\frac{\partial^2 L(\boldsymbol{\Phi};\mathbf{x})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right] + \mathrm{E}_C\left[\frac{\partial^2 \ln k(\mathbf{x}|\mathbf{y},\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]. \tag{43}$$

Cette formule peut s'écrire symboliquement sous la forme

$$\boxed{I(\boldsymbol{\Phi};\mathbf{y}) = \mathcal{I}_c(\boldsymbol{\Phi};\mathbf{x}) - \mathcal{I}_m(\boldsymbol{\Phi};\mathbf{y})}, \tag{44}$$

qui s'interprète comme une partition de l'information en ses composantes.

Le premier terme correspond à la matrice d'information (moins deux fois le hessien de la logvraisemblance) relative à $\boldsymbol{\Phi}$ procurée par les données observées $\mathbf{y}$,

$$I(\boldsymbol{\Phi};\mathbf{y}) = -\frac{\partial^2 L(\boldsymbol{\Phi};\mathbf{y})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}. \tag{45}$$

Le second terme représente la matrice d'information des données complètes $\mathbf{x}$ moyennée par rapport à la distribution conditionnelle des données manquantes $\mathbf{z}$ sachant les données observées $\mathbf{y}$ et le paramètre $\boldsymbol{\Phi}$, soit

$$\mathcal{I}_c(\boldsymbol{\Phi};\mathbf{x}) = -\mathrm{E}_C\left[\frac{\partial^2 L(\boldsymbol{\Phi};\mathbf{x})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]. \tag{46}$$

Le terme noté

$$\mathcal{I}_m(\boldsymbol{\Phi};\mathbf{y}) = -\mathrm{E}_C\left[\frac{\partial^2 \ln k(\mathbf{x}|\mathbf{y},\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right], \tag{47}$$

s'identifie, eu égard à (44), à la perte d'information $\mathcal{I}_c(\boldsymbol{\Phi};\mathbf{x}) - I(\boldsymbol{\Phi};\mathbf{y})$ consécutive au fait d'observer $\mathbf{y}$ et non $\mathbf{x}$ d'où son appellation d'information manquante.

Comme l'a montré initialement Louis (1982), on peut évaluer ce terme assez facilement. Soit $\mathbf{S}(\boldsymbol{\Phi};\mathbf{y}) = \dfrac{\partial L(\boldsymbol{\Phi};\mathbf{y})}{\partial\boldsymbol{\Phi}}$ et $\mathbf{S}(\boldsymbol{\Phi};\mathbf{x}) = \dfrac{\partial L(\boldsymbol{\Phi};\mathbf{x})}{\partial\boldsymbol{\Phi}}$ les fonctions de score relatives respectivement aux données observées et aux données complètes, on montre (*cf.* annexe A) que

$$\mathcal{I}_m(\boldsymbol{\Phi};\mathbf{y}) = \mathrm{Var}_C[\mathbf{S}(\boldsymbol{\Phi};\mathbf{x})] \tag{48}$$

c'est-à-dire que l'information manquante est la variance du score des données complètes, variance prise par rapport à la distribution conditionnelle de $\mathbf{z}$ sachant $\mathbf{y}$ et $\boldsymbol{\Phi}$.

Ce résultat découle directement du lemme suivant (*cf.* annexe A)

$$\frac{\partial^2 L(\boldsymbol{\Phi};\mathbf{y})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'} = \mathrm{E}_C\left[\frac{\partial^2 L(\boldsymbol{\Phi};\mathbf{x})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right] + \mathrm{Var}_C\left[\frac{\partial L(\boldsymbol{\Phi};\mathbf{x})}{\partial\boldsymbol{\Phi}}\right]$$

qui peut s'écrire aussi $-I(\boldsymbol{\Phi};\mathbf{y}) = -\mathcal{I}_c(\boldsymbol{\Phi};\mathbf{x}) + \mathrm{Var}_C[\mathbf{S}(\boldsymbol{\Phi};\mathbf{x})]$ QED.

Comme $\mathrm{E}_C[\mathbf{S}(\boldsymbol{\Phi};\mathbf{x})] = \mathbf{S}(\boldsymbol{\Phi};\mathbf{y})$ (*cf.* 7), l'expression se simplifie en

$$\mathcal{I}_m(\boldsymbol{\Phi};\mathbf{y}) = \mathrm{E}_C\left[\mathbf{S}(\boldsymbol{\Phi};\mathbf{x})\mathbf{S}'(\boldsymbol{\Phi};\mathbf{x})\right] - \mathbf{S}(\boldsymbol{\Phi};\mathbf{y})\mathbf{S}'(\boldsymbol{\Phi};\mathbf{y}). \tag{49a}$$

et, localement au point d'estimation ML $\boldsymbol{\Phi} = \widehat{\boldsymbol{\Phi}}$ tel que $\mathbf{S}(\widehat{\boldsymbol{\Phi}};\mathbf{y}) = \mathbf{0}$, on a

$$\mathcal{I}_m(\widehat{\boldsymbol{\Phi}};\mathbf{y}) = \mathrm{E}_C\left[\mathbf{S}(\boldsymbol{\Phi};\mathbf{x})\mathbf{S}'(\boldsymbol{\Phi};\mathbf{x})\right]\Big|_{\boldsymbol{\Phi}=\widehat{\boldsymbol{\Phi}}} \tag{49b}$$

d'où, un moyen de calcul de l'information observée

$$I(\widehat{\boldsymbol{\Phi}};\mathbf{y}) = \mathcal{I}_c(\widehat{\boldsymbol{\Phi}};\mathbf{x}) - \mathcal{I}_m(\widehat{\boldsymbol{\Phi}};\mathbf{y}). \tag{50}$$

### *1.6.5. Vitesse de convergence*

L'algorithme EM suppose implicitement l'existence d'une application $M$ de l'espace paramétrique $\Phi$ sur lui-même, puisque, par construction, on passe de façon univoque de $\boldsymbol{\Phi}^{[k]}$ à $\boldsymbol{\Phi}^{[k+1]}$. On peut donc écrire :

$$\boldsymbol{\Phi}^{[k+1]} = \mathbf{M}(\boldsymbol{\Phi}^{[k]}), \tag{51}$$

où $\mathbf{M}(\boldsymbol{\Phi})_{(r\times 1)} = \left[M_1(\boldsymbol{\Phi}), M_2(\boldsymbol{\Phi}), \ldots, M_i(\boldsymbol{\Phi}), \ldots, M_r(\boldsymbol{\Phi})\right]'$ et $\boldsymbol{\Phi}_{(r\times 1)} = \{\phi_i\}$.

En faisant un développement limité de $\mathbf{M}(\boldsymbol{\Phi}^{[k]})$ au premier ordre au voisinage de $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{[k-1]}$, on obtient :

$$\boldsymbol{\Phi}^{[k+1]} - \boldsymbol{\Phi}^{[k]} \approx \mathbf{J}(\boldsymbol{\Phi}^{[k-1]})(\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^{[k-1]}). \tag{52}$$

Dans cette formule, $\mathbf{J}(\boldsymbol{\Phi})$ est la matrice jacobienne $(r \times r)$ dont l'élément $(i,j)$ s'écrit

$$\mathbf{J}_{ij}(\boldsymbol{\Phi}) = \partial M_i / \partial \phi_j,$$

où $M_i$ est le ième élément de $\mathbf{M}$ et $\phi_j$ le jème élément du vecteur $\boldsymbol{\Phi}$. Si $\boldsymbol{\Phi}^{[k-1]} \longrightarrow \boldsymbol{\Phi}^*$ alors, sous les conditions de continuité habituelles, $\mathbf{J}(\boldsymbol{\Phi}^{[k-1]}) \longrightarrow \mathbf{J}(\boldsymbol{\Phi}^*)$ si bien qu'à partir d'un certain rang, on pourra écrire $\boldsymbol{\Phi}^{[k+1]} - \boldsymbol{\Phi}^{[k]} \approx \mathbf{J}(\boldsymbol{\Phi}^*)(\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^{[k-1]})$.

La vitesse de convergence

$$v = \lim_{k \to \infty} ||\boldsymbol{\Phi}^{[k+1]} - \boldsymbol{\Phi}^{[k]}|| / ||\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^{[k-1]}|| \tag{53}$$

est alors gouvernée par la plus grande valeur propre de $\mathbf{J}(\boldsymbol{\Phi}^*)$, $v = \max_{1 \leqslant i \leqslant r} \lambda_i$, une valeur élevée de cette valeur propre impliquant une convergence lente.

Dans le cas de la famille exponentielle, Dempster, Laird et Rubin ont montré que

$$\mathbf{J}(\boldsymbol{\Phi}^*) = \left\{\mathrm{var}[\mathbf{t}(\mathbf{x})|\boldsymbol{\Phi}^*]\right\}^{-1} \mathrm{var}[\mathbf{t}(\mathbf{x})|\boldsymbol{\Phi}^*, \mathbf{y}], \tag{54}$$

où $\mathbf{t}(\mathbf{x})$ est le vecteur des statistiques exhaustives de $\boldsymbol{\Phi}$ basées sur les données complètes $\mathbf{x}$.

De façon générale, ces mêmes auteurs ont établi que

$$\mathbf{J}(\boldsymbol{\Phi}^*) = \mathcal{I}_c^{-1}(\boldsymbol{\Phi}^*; \mathbf{x}) \mathcal{I}_m(\boldsymbol{\Phi}^*; \mathbf{y}), \tag{55}$$

quantité qui mesure la fraction de l'information complète qui est perdue du fait de la non observation de $\mathbf{z}$ en sus de $\mathbf{y}$. Si la perte d'information due à l'existence de données incomplètes est faible, la convergence sera rapide, cette perte d'information pouvant d'ailleurs varier selon les composantes de $\boldsymbol{\Phi}$.

Comme $\mathcal{I}_m(\boldsymbol{\Phi}; \mathbf{y}) = \mathcal{I}_c(\boldsymbol{\Phi}; \mathbf{x}) - I(\boldsymbol{\Phi}; \mathbf{y})$, la formule (55) peut s'écrire aussi

$$\mathbf{J}(\boldsymbol{\Phi}^*) = \mathbf{I}_r - \mathcal{I}_c^{-1}(\boldsymbol{\Phi}^*; \mathbf{x}) I(\boldsymbol{\Phi}^*; \mathbf{y}). \tag{56}$$

Pour être en conformité avec la littérature numérique, c'est la matrice $\mathbf{I}_r - \mathbf{J}(\boldsymbol{\Phi}^*) = \mathcal{I}_c^{-1}(\boldsymbol{\Phi}^*; \mathbf{x}) I(\boldsymbol{\Phi}^*; \mathbf{y})$ dont la valeur propre la plus petite définit les performances de l'algorithme qui, certaines fois, est qualifiée de matrice de vitesse de convergence.

L'expression (56) conduit aussi à exprimer la matrice d'information des données observées sous la forme

$$\mathrm{I}(\boldsymbol{\Phi}^*; \mathbf{y}) = \mathcal{I}_c(\boldsymbol{\Phi}^*; \mathbf{x}) \left[\mathbf{I}_r - \mathbf{J}(\boldsymbol{\Phi}^*)\right], \tag{57}$$

et, pour l'inverse :

$$\begin{aligned} \mathrm{I}^{-1}(\boldsymbol{\Phi}^*; \mathbf{y}) &= \left[\mathbf{I}_r - \mathbf{J}(\boldsymbol{\Phi}^*)\right]^{-1} \mathcal{I}_c^{-1}(\boldsymbol{\Phi}^*; \mathbf{x}) \\ &= \left\{\mathbf{I}_r + \left[\mathbf{I}_r - \mathbf{J}(\boldsymbol{\Phi}^*)\right]^{-1} \mathbf{J}(\boldsymbol{\Phi}^*)\right\} \mathcal{I}_c^{-1}(\boldsymbol{\Phi}^*; \mathbf{x}) \end{aligned}$$

$$\boxed{\mathrm{I}^{-1}(\boldsymbol{\Phi}^*; \mathbf{y}) = \mathcal{I}_c^{-1}(\boldsymbol{\Phi}^*; \mathbf{x}) + \left[\mathbf{I}_r - \mathbf{J}(\boldsymbol{\Phi}^*)\right]^{-1} \mathbf{J}(\boldsymbol{\Phi}^*) \mathcal{I}_c^{-1}(\boldsymbol{\Phi}^*; \mathbf{x})} \tag{58}$$

Cette formule est la base d'un algorithme dit « Supplemented EM » (Meng and Rubin, 1991) permettant de calculer la précision asymptotique des estimations ML obtenues via l'algorithme EM.

Au voisinage de $\boldsymbol{\Phi}^*$, on peut écrire, par un développement limité de $\boldsymbol{\Phi}^{[k+1]} = \mathbf{M}(\boldsymbol{\Phi}^{[k]})$ au premier ordre

$$\boldsymbol{\Phi}^{[k+1]} - \boldsymbol{\Phi}^* \approx \mathbf{J}(\boldsymbol{\Phi}^*)(\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^*), \tag{59}$$

formule qui indique le caractère linéaire de la convergence des itérations EM. Un algorithme ayant ce type de convergence peut être accéléré notamment par la version multivariée de la méthode d'accélération d'Aitken. On a

$$\boldsymbol{\Phi}^* - \boldsymbol{\Phi}^{[k-1]} = (\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^{[k-1]}) + (\boldsymbol{\Phi}^{[k+1]} - \boldsymbol{\Phi}^{[k]}) + (\boldsymbol{\Phi}^{[k+2]} - \boldsymbol{\Phi}^{[k+1]} + \dots + (\boldsymbol{\Phi}^{[k+h-1]} - \boldsymbol{\Phi}^{[k+h]}) + \dots$$

Or, du fait de l'expression (52),

$$\boldsymbol{\Phi}^{[k+h+1]} - \boldsymbol{\Phi}^{[k+h]} = \mathbf{J}^h(\boldsymbol{\Phi}^*)(\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^{[k-1]}),$$

et, en reportant dans l'expression précédente, il vient

$$\boldsymbol{\Phi}^* = \boldsymbol{\Phi}^{[k-1]} + \left[\sum_{h=0}^{\infty} \mathbf{J}^h(\boldsymbol{\Phi}^*)\right](\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^{[k-1]}),$$

soit encore, en utilisant la propriété de convergence de la série géométrique $\sum_{h=0}^{\infty} \mathbf{J}^h(\boldsymbol{\Phi}^*)$ vers $\left[\mathbf{I}_r - \mathbf{J}(\boldsymbol{\Phi}^*)\right]^{-1}$ lorsque ses valeurs propres sont comprises entre 0 et 1

$$\boxed{\boldsymbol{\Phi}^* = \boldsymbol{\Phi}^{[k-1]} + \left[\mathbf{I}_r - \mathbf{J}(\boldsymbol{\Phi}^*)\right]^{-1}(\boldsymbol{\Phi}^{[k]} - \boldsymbol{\Phi}^{[k-1]})}. \tag{60}$$

Laird *et al.* (1987) ont proposé une approximation numérique de $\mathbf{J}(\boldsymbol{\Phi}^*)$ à partir de l'historique des itérations EM et qu'ils appliquent au calcul des estimations REML des composantes de la variance pour des modèles linéaires mixtes d'analyse de données répétées. Ainsi, de l'itération $k$, on va pouvoir se projeter, si tout va bien, au voisinage de $\boldsymbol{\Phi}^*$, donc réduire les calculs et gagner du temps.

### 1.7. Variantes

À partir de la théorie de base telle qu'elle fut formulée par Dempster, Laird et Rubin se sont développées maintes variantes qui répondent au besoin d'adapter celle-ci aux difficultés qui peuvent se rencontrer, soit dans la mise en œuvre des phases E et M, soit dans l'obtention de résultats supplémentaires ou de meilleures performances. Sans avoir la prétention d'être exhaustif, nous répertorierons les principales d'entre elles.

*1.7.1. « Gradient-EM »*

On fait appel à cette technique lorsqu'il n'y a pas de solution analytique à la phase M. Dans la version décrite par Lange (1995), celle-ci est réalisée par la méthode de Newton-Raphson. Sachant la valeur courante des paramètres $\boldsymbol{\Phi}^{[t]}$, on va initier une série d'itérations internes $\boldsymbol{\Phi}^{[t,k]}$ utilisant les expressions du gradient et du hessien de la fonction $Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})$ soit

$$-\ddot{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})\Big|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}^{[t,k]}}(\boldsymbol{\Phi}^{[t,k+1]}-\boldsymbol{\Phi}^{[t,k]})=\dot{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})\Big|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}^{[t,k]}}, \qquad (61)$$

où $\dot{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})=\partial Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})/\partial\boldsymbol{\Phi}$ et $\ddot{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})=\partial^2 Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})/\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'$. Il peut être avantageux numériquement de ne pas aller jusqu'à la convergence en réduisant le nombre d'itérations internes jusqu'à une seule $\boldsymbol{\Phi}^{[t,1]}=\boldsymbol{\Phi}^{[t+1;0]}$ comme l'envisage Lange. Dans ce cas, il importe toutefois de bien vérifier qu'on augmente la fonction $Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})$ et qu'on reste ainsi dans le cadre d'un EM dit généralisé.

Dans certaines situations, l'expression de $\mathrm{E}\big[\ddot{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})\big]$ prise par rapport à la distribution de $\mathbf{y}$ est beaucoup plus simple que celle de $\ddot{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})$ et l'on aura alors recours à un algorithme de Fisher (Titterington, 1984; Foulley *et al.*, 2000).

*1.7.2. ECM et ECME*

La technique dite ECM (« Expectation Conditional Maximisation ») a été introduite par Meng et Rubin (1993) en vue de simplifier la phase de maximisation quand celle-ci fait intervenir différents types de paramètres. On partitionne alors le vecteur des paramètres $\boldsymbol{\Phi}=(\boldsymbol{\gamma}',\boldsymbol{\theta}')'$ en sous vecteurs (par exemple $\boldsymbol{\gamma}$ et $\boldsymbol{\theta}$), puis on maximise la fonction $Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})$ successivement par rapport à $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ étant fixé, puis par rapport $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ étant fixé, soit

$$\boldsymbol{\gamma}^{[t+1]}=\operatorname{argmax} Q(\boldsymbol{\gamma},\boldsymbol{\theta}^{[t]};\boldsymbol{\Phi}^{[t]}), \qquad (62a)$$

$$\boldsymbol{\theta}^{[t+1]}=\operatorname{argmax} Q(\boldsymbol{\gamma}^{[t+1]},\boldsymbol{\theta};\boldsymbol{\Phi}^{[t]}). \qquad (62b)$$

Dans la version dite ECME (« Expectation Conditional Maximisation Either ») due à Liu et Rubin (1994), une des étapes de maximisation conditionnelle précédentes est réalisée par maximisation directe de la vraisemblance $L(\boldsymbol{\Phi};\mathbf{y})$ des données observées, soit, par exemple,

$$\boldsymbol{\theta}^{[t+1]}=\operatorname{argmax} L(\boldsymbol{\gamma}^{[t+1]},\boldsymbol{\theta};\mathbf{y}). \qquad (63)$$

*1.7.3. EM stochastique*

Cette méthode fut introduite par Celeux et Diebolt (1985) en vue de l'estimation ML des paramètres d'une loi de mélange. Le principe de cette méthode dite en abrégé SEM (« Stochastic EM ») réside dans la maximisation de la logvraisemblance $L(\boldsymbol{\Phi};\mathbf{x})=\ln[f(\mathbf{x}|\boldsymbol{\Phi})]$ des données complètes à partir, non pas de son expression analytique, mais grâce à une évaluation numérique

de celle-ci *via* le calcul de $\ln\big[f(\mathbf{y},\mathbf{z}^{[t]}|\boldsymbol{\Phi})\big]$ où $\mathbf{z}^{[t]}$ est un échantillon simulé de données manquantes tiré dans la distribution conditionnelle de celles-ci de densité $h(\mathbf{z}^{[t]}|\mathbf{y},\boldsymbol{\Phi}=\boldsymbol{\Phi}^{[t]})$. Outre la simplicité du procédé, celui-ci offre l'avantage d'éviter le blocage de l'algorithme en des points stationnaires stables mais indésirables (Celeux *et al.*, 1996).

Wei et Tanner (1990) reprennent cette idée pour calculer la fonction $Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})$ de la phase E quand celle-ci n'est plus possible analytiquement par le biais d'une approximation de Monte- Carlo classique d'une espérance (Robert et Casella, 1999 ; formule 5.3.4 page 208). Concrètement, on procède comme suit :

a) tirage de $m$ échantillons de $\mathbf{z}$ soit $\mathbf{z}_1,\ldots,\mathbf{z}_j,\ldots,\mathbf{z}_m$ extraits de la loi de densité ;

$$h(\mathbf{z}|\mathbf{y},\boldsymbol{\Phi}=\boldsymbol{\Phi}^{[t]});$$

b) approximation de $Q(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]}$ par

$$\widetilde{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})=\frac{1}{m}\sum_{j=1}^{m}\ln f(\mathbf{y},\mathbf{z}_j|\boldsymbol{\Phi}). \tag{64}$$

On remarque de suite que pour $m=1$, MCEM se ramène exactement à SEM, et que pour $m\longrightarrow\infty$, MCEM équivaut à EM. On gagnera donc à moduler les valeurs de $m$ au cours du processus itératif (Tanner, 1996 ; Booth and Hobert, 1999) ; en partant par exemple de $m_0=1$ et, en accroissant continûment et indéfiniment $m$ selon une progression adéquate, on mime ainsi un algorithme de recuit simulé où l'inverse de $m$ joue le rôle de la température (Celeux *et al.*, 1995). D'un point de vue théorique, les propriétés de SEM notamment les résultats asymptotiques ont été établis par Nielsen (2000).

Il y a des variantes autour de ces algorithmes de base. Mentionnons par exemple l'algorithme dit « SAEM » (« Stochastic Approximative EM »). Dans la version de Celeux et Diebolt (1992), l'actualisation du paramètre courant $\boldsymbol{\Phi}^{[t]}$ par SAEM s'effectue par combinaison des valeurs actualisées $\boldsymbol{\Phi}_{SEM}^{[t+1]}$ de SEM et $\boldsymbol{\Phi}_{EM}^{[t+1]}$ de EM selon la formule suivante :

$$\boldsymbol{\Phi}^{[t+1]}=\gamma_{t+1}\boldsymbol{\Phi}_{SEM}^{[t+1]}+(1-\gamma_{t+1})\boldsymbol{\Phi}_{EM}^{[t+1]}, \tag{65}$$

où les $\gamma_t$ forment une suite de nombres réels décroissant de $\gamma_0=1$ à $\gamma_\infty=0$ avec les deux conditions suivantes : $\mathrm{Lim}(\gamma_t/\gamma_{t+1})=1$ et $\sum_t\gamma_t\longrightarrow\infty$ quand $t\longrightarrow\infty$. Ces deux conditions assurent la convergence presque sûre de la suite des itérations SAEM vers un maximum local de la vraisemblance.

Ce faisant, on réalise à chaque étape un dosage entre une actualisation purement EM et une actualisation purement stochastique, cette dernière composante étant dominante au départ pour s'amenuiser au cours des itérations au profit de la composante EM.

Dans la version de Delyon *et al.* (1999), cette combinaison se fait à la phase E sur la base de la fonction $Q$ précédente notée ici $\underline{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]})$ et de la partie

simulée en (64) selon la formule

$$\underline{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{t+1]}) = \gamma_{t+1}\left[\frac{1}{m_{t+1}}\sum_{j=1}^{m_{t+1}} \ln f(\mathbf{y},\mathbf{z}_{j,t+1}|\boldsymbol{\Phi})\right] + (1-\gamma_{t+1})\underline{Q}(\boldsymbol{\Phi};\boldsymbol{\Phi}^{[t]}). \tag{66}$$

De la même façon, la composante purement simulée dominante au départ ira en s'amenuisant au fil des itérations. L'avantage par rapport à MCEM réside dans la prise en compte de toutes les valeurs simulées depuis le départ alors que seules les $m_t$ simulées à l'étape $t$ sont prises en compte dans l'algorithme MCEM. Les conditions de convergence de cet algorithme ont été discutées par Delyon *et al.* (1999) et par Kuhn et Lavielle (2002) quand le processus de simulation des données manquantes s'effectue *via* MCMC.

### *1.7.4. EM supplémenté*

Cet algorithme dit «EM supplémenté» (SEM en abrégé) fut introduit par Meng et Rubin (1991) pour compléter l'EM classique, en vue d'obtenir la précision des estimations ML de $\boldsymbol{\Phi}$ sous la forme de la matrice de variance covariance asymptotique de $\widehat{\boldsymbol{\Phi}}$.

Le point de départ de cet algorithme est la formule donnant l'expression de l'inverse de la matrice d'information de Fisher relative à $\boldsymbol{\Phi}$ vue précédemment (*cf.* 58),

$$I^{-1}(\widehat{\boldsymbol{\Phi}};\mathbf{y}) = \mathcal{I}_c^{-1}(\widehat{\boldsymbol{\Phi}};\mathbf{x}) + \left[\mathbf{I}_r - \mathbf{J}(\widehat{\boldsymbol{\Phi}})\right]^{-1}\mathbf{J}(\widehat{\boldsymbol{\Phi}})\mathcal{I}_c^{-1}(\widehat{\boldsymbol{\Phi}};\mathbf{x}),$$

en fonction de l'inverse $\mathcal{I}_c^{-1}(\widehat{\boldsymbol{\Phi}};\mathbf{x})$ de la matrice d'information des données complètes $\mathbf{x}$ moyennée par rapport à la distribution conditionnelle des données manquantes et de la matrice jacobienne $\mathbf{J}(\boldsymbol{\Phi})$ dont l'élément $ij$ se définit par $r_{ij} = \partial M_i/\partial\phi_j$.

Dans la famille exponentielle, il n'y a pas de difficulté particulière à l'obtention de $\mathcal{I}_c^{-1}(\widehat{\boldsymbol{\Phi}};\mathbf{x})$. L'apport crucial de Meng et Rubin (1991) est d'avoir montré comment on pouvait évaluer numériquement la matrice $\mathbf{J}(\widehat{\boldsymbol{\Phi}})$ à partir de la mise en œuvre de l'EM classique. Posons, à l'instar de McLachlan et Krishnan (1997) : $\boldsymbol{\Phi}_{(j)}^{[t]}(\widehat{\phi}_1,\widehat{\phi}_2,\ldots,\widehat{\phi}_{j-1},\phi_j^{[t]},\ldots,\widehat{\phi}_r)'$, $r_{ij}$ peut s'écrire comme suit :

$$r_{ij} = \lim_{t\to\infty}\frac{M_i(\boldsymbol{\Phi}_{(j)}^{[t]}) - \widehat{\boldsymbol{\Phi}}_i}{\phi_j^{[t]} - \widehat{\phi}_j}. \tag{67}$$

En fait, l'algorithme EM réalise l'application (*cf.* 51) lors du passage d'une itération à l'autre. En pratique, partant de $\boldsymbol{\Phi}_{(j)}^{[t]}$ comme valeur courante, l'itération suivante de EM relative à la composante $\phi_i^{[t+1]}$ procure donc la valeur de $M_i(\boldsymbol{\Phi}_{(j)}^{[t]})$ d'où l'on déduit la valeur de $r_{ij}$ à partir de la formule (67). Ce calcul est réalisé pour différentes valeurs de $t$ de façon à ne retenir *in fine* que les valeurs stables de $r_{ij}$. McLachlan et Krishnan (1997) notent que les caractéristiques de la matrice $\mathbf{I}_r - \mathbf{J}(\widehat{\boldsymbol{\Phi}})$ ainsi obtenues sont de bons outils

de diagnostic de la solution $\widehat{\boldsymbol{\phi}}$ obtenue. Ainsi, lorsque cette matrice n'est pas positive définie, on peut en inférer que l'algorithme a convergé vers un point selle indétectable par la procédure classique. Il conviendra alors de réamorcer une séquence EM à partir de ces valeurs affectées d'une perturbation adéquate.

Une autre façon d'obtenir la précision de l'estimation $\widehat{\boldsymbol{\phi}}$ est de repartir de la formule générale $I(\widehat{\boldsymbol{\phi}};\mathbf{y}) = \mathcal{I}_c(\widehat{\boldsymbol{\phi}};\mathbf{y}) - \mathcal{I}_m(\widehat{\boldsymbol{\phi}};\mathbf{y})$ et d'utiliser la formule de Louis vue en (49ab) soit $\mathcal{I}_m(\widehat{\boldsymbol{\phi}};\mathbf{y})\mathrm{E}_C\left[\mathbf{S}(\boldsymbol{\phi};\mathbf{x})\mathbf{S}'(\boldsymbol{\phi};\mathbf{x})\right]\Big|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}}$ qu'on peut évaluer par simulation en prenant la moyenne sur $m$ échantillons du produit du score $\mathbf{S}(\boldsymbol{\phi};\mathbf{y},\mathbf{z}_j)$ par son transposé (Tanner, 1996). On peut aussi avoir recours à des techniques de boostrap classique ou paramétrique.

### *1.7.5. PX-EM*

L'algorithme EM fait partie des standards de calcul des estimations de maximum de vraisemblance. Il doit son succès à sa simplicité de formulation, à sa stabilité numérique et à la diversité de son champ d'application. Toutefois, sa vitesse de convergence peut s'avérer lente dans certains types de problème d'où des tentatives pour y remédier. Dans le cas du modèle mixte, plusieurs auteurs ont proposé des procédures de « normalisation » des effets aléatoires (Foulley et Quaas, 1995 ; Lindström et Bates, 1988 ; Meng et van Dyk, 1998 ; Wolfinger et Tobias, 1998). Ce principe a été repris par Meng et van Dyk (1997) puis généralisé par Liu *et al.* (1998) dans le cadre d'une nouvelle version de l'algorithme qualifiée de « Parameter Expanded EM » (PX-EM en abrégé).

Cette théorie repose sur le concept d'extension paramétrique à un espace plus large $\boldsymbol{\varphi}$ que l'espace d'origine par adjonction d'un vecteur de paramètres de travail $\alpha$ tel que $\boldsymbol{\varphi} = (\boldsymbol{\phi}'_*, \boldsymbol{\alpha}')'$ où $\boldsymbol{\phi}_*$ joue le même rôle dans la densité des données complètes $p_X\left[\mathbf{x}|\boldsymbol{\varphi} = (\boldsymbol{\phi}'_*, \boldsymbol{\alpha}')'\right]$ du modèle étendu (noté X) que $\boldsymbol{\phi}$ dans celle $p(\mathbf{x}|\boldsymbol{\phi})$ du modèle d'origine (noté O). Cette extension doit satisfaire les deux conditions suivantes :

1) retour à l'espace d'origine sans ambiguïté par la fonction $\boldsymbol{\phi} = R(\boldsymbol{\varphi})$ ;

2) préservation du modèle des données complètes pour $\boldsymbol{\alpha}$ pris à sa valeur de référence $\boldsymbol{\alpha}_0$ c'est-à-dire que pour $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, la loi de $\mathbf{x}$ se réduit à celle définie sous le modèle O soit $p_0(\mathbf{x}|\boldsymbol{\phi}) = p_X(\mathbf{x}|\boldsymbol{\phi}^* = \boldsymbol{\phi}, \boldsymbol{\alpha} = \boldsymbol{\alpha}_0)$. Autrement dit, si l'on pose $\boldsymbol{\phi}^* = \boldsymbol{\phi}^*(\boldsymbol{\alpha})$ alors $\boldsymbol{\phi}^*(\boldsymbol{\alpha}_0) = \boldsymbol{\phi}$.

La première condition se traduit par le fait que la logvraisemblance reste inchangée $L(\boldsymbol{\gamma};\mathbf{y}) = L(\boldsymbol{\gamma}_*, \boldsymbol{\alpha};\mathbf{y})$ quelle que soit la valeur de $\boldsymbol{\alpha}$ choisie. La deuxième condition est mise à profit à la phase E en prenant l'espérance de la logvraisemblance des données complètes par rapport à la densité $h(\mathbf{z}|\mathbf{y}, \boldsymbol{\varphi} = \boldsymbol{\varphi}^{[t]})$ des données manquantes où $\boldsymbol{\alpha}^{[t]}$ est égalé à sa valeur de référence $\boldsymbol{\alpha}_0$ simplifiant ainsi grandement la mise en œuvre de cette étape qui devient identique à celle d'un algorithme classique sous le modèle d'origine O (dit EMO).

L'exemple de Liu *et al.* (1998) permet d'illustrer ces principes. Il s'agit d'un modèle linéaire aléatoire très simple généré par l'approche hiérarchique suivante à deux niveaux :

1) $y|z \sim \mathcal{N}(z,1)$ où $y$ désigne la variable observée et $z$ la variable manquante ;

2) $z|\theta,\ \sigma^2 \sim \mathcal{N}(\theta,\sigma^2)$ où l'espérance $\theta$ de la loi de $z$ est le paramètre inconnu et la variance $\sigma^2$ est supposée connue.

Remarquons que cela équivaut à écrire : 1) $y = z + e$; $e \sim \mathcal{N}(0,1)$, et 2) $z = \theta + u$; $u \sim \mathcal{N}(0,\sigma^2)$ soit encore, marginalement $y = \theta + u + e$, et l'on reconnaît là une structure de modèle linéaire aléatoire. Dans l'algorithme classique (dit EMO puisqu'il y s'appuie sur le modèle d'origine O) on procède comme suit.

Phase E : $z$ étant une statistique exhaustive de $\theta$, on remplace $z$ par son espérance conditionnelle $\widetilde{z}^{[t]} = \mathrm{E}(z|\theta^{[t]},\sigma^2,y)$. Du fait de l'hypothèse de normalité des distributions, cette espérance s'écrit comme l'équation de régression de $y$ en $z$,

$\widetilde{z}^{[t]} = \mathrm{E}(z|\theta^{[t]},\sigma^2) + \mathrm{Cov}(y,z)(\mathrm{var}\, y)^{-1}\left[y - \mathrm{E}(y|\theta^{[t]},\sigma^2)\right]$, soit compte tenu de 1) et 2),

$$\widetilde{z}^{[t]} = \theta^{[t]} + \frac{\sigma^2}{1+\sigma^2}(y - \theta^{[t]}) = \frac{\theta^{[t]} + \sigma^2 y}{1+\sigma^2}. \tag{68}$$

Phase M : On résout l'équation $\mathrm{E}(z|\theta^{[t+1]},\sigma^2) = \mathrm{E}(z|y,\theta^{[t]},\sigma^2)$ qui a pour solution $\theta^{[t+1]} = \dfrac{\theta^{[t]} + \sigma^2 y}{1+\sigma^2}$, d'où l'expression de l'écart entre cette itération EM0 et l'estimateur vrai $(y)$

$$\theta_{EMO}^{[t+1]} - \theta_{ML} = \frac{\theta^{[t]} - y}{1+\sigma^2}, \tag{69}$$

formule qui indique que la convergence va être d'autant plus lente que $\sigma^2$ sera petit.

Liu *et al.* (1998) formulent le modèle reparamétré (dit X) en y incluant un décentrage $\alpha$ : 1) $y \,|\, z \sim \mathcal{N}(z+\alpha,1)$ et 2) $z \,|\, \theta_*,\ \sigma^2 \sim \mathcal{N}(\theta_*,\sigma^2)$. Pour détailler le raisonnement, on peut expliciter la logvraisemblance des données complètes :

$$-2L(\theta_*,\alpha,\sigma^2;y,z) = \left[(z-\theta_*)^2/\sigma^2\right] + (y-z-\alpha)^2 + \ln\sigma^2. \tag{70}$$

On retrouve alors la propriété selon laquelle $z$ est une statistique exhaustive de $\theta_*$. La phase E reste inchangée puisque la loi de $z \,|\, \theta_*$, $\alpha = 0$, $\sigma^2$, $y$ est identique à la loi de $z \,|\, \theta$, $\sigma^2$, $y$. À la phase M, on résout $\mathrm{E}(z \,|\, \theta_*^{[t+1]},\sigma^2) = \widetilde{z}^{[t]}$ soit $\theta_*^{[t+1]} = \dfrac{\theta^{[t]} + \sigma^2 y}{1+\sigma^2}$. Quant à $\alpha$, on a, eu égard à l'expression (70), $\alpha^{[t+1]} = y - \widetilde{z}^{[t]}$ soit, compte tenu de (68), $\alpha^{[t+1]} = y - \dfrac{\theta^{[t]} + \sigma^2 y}{1+\sigma^2}$ et $\theta^{[t+1]} = \alpha^{[t+1]} + \theta_*^{[t+1]}$ c'est-à-dire $\theta^{t+1]} = y$ si bien que la convergence s'obtient dès la 1ère itération. On peut expliciter la relation entre les deux algorithmes sous la forme de l'équation suivante :

$$\theta_{PX}^{[t+1]} = \theta_{EMO}^{[t+1]} + (\alpha^{[t+1]} - \alpha_0),$$

que Liu *et al.* (1998) mettent en avant pour montrer que la phase M de l'algorithme PX est à même d'exploiter par régression l'information apportée par la différence $(\alpha^{[t+1]} - \alpha_0]$ pour ajuster $\theta_{EMO}^{[t+1]}$. Liu et Wu (1999) ont repris ce même exemple sous une forme légèrement différente : 1) $y \mid \theta, \alpha, w \sim \mathcal{N}(\theta - \alpha + w, 1)$ et 2) $w \mid \theta, \alpha, \sigma^2 \sim \mathcal{N}(\alpha, \sigma^2)$ dans laquelle le décentrage porte sur la variable aléatoire manquante $w$ initialement centrée.

Des extensions de l'algorithme PX ont été également proposées par Liu et Wu (1999) et van Dyk et Meng (2001) à des fins d'inférence bayésienne sur la loi *a posteriori* $\boldsymbol{\phi} \mid \mathbf{y}$ dans le cadre de l'algorithme dit « Data augmentation » de Tanner et Wong (1987).

## 2. Application au modèle linéaire mixte

### 2.1. Rappels

#### *2.1.1. Modèle mixte*

Nous allons considérer maintenant quelques applications de l'algorithme au modèle linéaire mixte. Il y a une double justification à cela. En premier lieu, le modèle linéaire mixte offre une illustration typique du concept élargi de données manquantes par le biais des effets aléatoires qui interviennent dans ce modèle. En second lieu, ce type de modèle suscite actuellement un vif intérêt de la part des praticiens de la statistique car c'est l'outil de base pour l'analyse paramétrique des données corrélées. En effet, un modèle linéaire mixte est un modèle linéaire du type $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ dans lequel la matrice de variance covariance des observations $\mathbf{V} = \mathrm{var}(\boldsymbol{\varepsilon})$ est structurée linéairement $\mathbf{V} = \sum_m \boldsymbol{\gamma}_m \mathbf{V}_m$ en fonction de paramètres $\boldsymbol{\gamma}_m$ grâce à une décomposition de la résiduelle $\boldsymbol{\varepsilon}$ en une combinaison linéaire $\boldsymbol{\varepsilon} = \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k$ de variables aléatoires structurales $\mathbf{u}_k$ (Rao et Kleffe, 1988).

Sous la forme la plus générale, le modèle linéaire mixte s'écrit : $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, où y est le vecteur $(N \times 1)$ des observations ; $\mathbf{X}$ est la matrice $(N \times p)$ des variables explicatives (continues ou discrètes) de la partie systématique du modèle auquel correspond, le vecteur $\boldsymbol{\beta} \in \mathbb{R}^p$ des coefficients dits aussi « effets fixes » ; $\mathbf{u}$ est le vecteur $(q \times 1)$ des variables aléatoires « structurales » ou effets aléatoires de matrice d'incidence $\mathbf{Z}$ de dimension $(N \times q)$ et $\mathbf{e}$ est le vecteur $(N \times 1)$ des variables aléatoires dites résiduelles.

Ce modèle linéaire est caractérisé notamment par son espérance et sa variance qui s'écrivent : $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ où $\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$ et $\mathrm{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$.

*2.1.2. Maximum de vraisemblance*

L'estimation des paramètres de position $\boldsymbol{\beta}$ et de dispersion $\boldsymbol{\gamma} = \{\gamma_m\}$ (intervenant dans la matrice de variance covariance $\mathbf{V}$) s'effectue naturellement dans le cadre gaussien $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ par la méthode du maximum de vraisemblance (Hartley et Rao, 1967) soit, où $(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\gamma}}')' = \text{argmax}_{\beta,\gamma} L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$, où

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = -1/2\Big[N \ln(2\pi) + \ln|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big].$$

Afin de corriger le biais d'estimation de $\boldsymbol{\gamma}$ lié au maximum de vraisemblance classique (ML), Patterson et Thompson (1971) considèrent une vraisemblance de résidus $\mathbf{v} = \mathbf{S}\mathbf{y}$ où $\mathbf{S} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ qui, par construction, ne dépend pas des effets fixes $\boldsymbol{\beta}$. Par maximisation de cette fonction par rapport aux paramètres, on obtient un maximum de vraisemblance restreinte ou mieux résiduelle (REML en anglais). Harville (1977) propose de ne prendre que $N - r(\mathbf{X})$ éléments linéairement indépendants de $\mathbf{v}$ (notés $\mathbf{K}'\mathbf{y}$) qu'il appelle « contrastes d'erreur ». En définitive, on montre que moins deux fois la logvraisemblance de $\boldsymbol{\gamma}$ basée sur $\mathbf{K}'\mathbf{y}$ peut se mettre sous la forme (Foulley *et al.*, 2002),

$$-2L(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}) = C + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y},$$

où $C$ est une constante égale dans sa forme la plus simple à $[N - r(\mathbf{X})]\ln 2\pi$, $\underline{\mathbf{X}}$ correspond à toute matrice formée par $r(\mathbf{X})$ colonnes de $\mathbf{X}$ linéairement indépendantes et $\underline{\mathbf{P}} = \mathbf{V}^{-1}[\mathbf{I}_N - \mathbf{Q}]$ où $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ est le projecteur des moindres carrés généralisés.

En outre, il importe de souligner que REML peut s'interpréter et se justifier très simplement dans le cadre bayésien comme un maximum de vraisemblance marginale $p(\mathbf{y} \mid \boldsymbol{\gamma}) = \int p(\mathbf{y}, \boldsymbol{\beta} \mid \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\beta}$ après intégration des effets fixes selon un *a priori* uniforme (Harville, 1974).

## 2.2. Modèle à un facteur aléatoire

*2.2.1. EM-REML*

Nous nous placerons au départ pour simplifier dans le cadre du modèle linéaire à un facteur aléatoire $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ avec $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{u}_{(q\times 1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, $\mathbf{e}_{(N\times 1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, $\mathrm{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$ avec ici $\mathbf{G} = \sigma_1^2\mathbf{I}_q$, $\mathbf{R} = \sigma_0^2\mathbf{I}_N$ et $\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \sigma_1^2\mathbf{Z}\mathbf{Z}' + \sigma_0^2\mathbf{I}_N$.

Les données observables (ou données incomplètes dans la terminologie EM) sont constituées du vecteur $\mathbf{y}$. Le vecteur des données manquantes $\mathbf{z} = (\boldsymbol{\beta}', \mathbf{u}')'$ est choisi comme la concaténation de $\boldsymbol{\beta}$ et de $\mathbf{u}$. Ici, à l'instar de Dempster *et al.* (1977) et Searle *et al.* (1992, page 303), $\boldsymbol{\beta}$ n'est pas considéré comme un paramètre, mais comme une variable aléatoire parasite dont la variance tend vers une valeur limite infinie. Cette façon de procéder renvoie à l'interprétation bayésienne de la vraisemblance résiduelle. Ce faisant, $\boldsymbol{\beta}$ sera éliminé par

intégration d'où l'obtention de REML. Cette interprétation a également l'avantage de dépasser l'interprétation stricte de REML comme vraisemblance de contrastes d'erreur, ce qui peut s'avérer très utile dans le cas non linéaire notamment (Liao et Lipsitz, 2002).

Dans ces conditions, $\boldsymbol{\phi} = (\sigma_1^2, \sigma_0^2)'$ et $\mathbf{x} = (\mathbf{y}', \boldsymbol{\beta}', \mathbf{u}')$ si bien que la densité de $\mathbf{x}$ se factorise en $p(\mathbf{x} \mid \boldsymbol{\phi}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_0^2) p(\mathbf{u} \mid \sigma_1^2)$. Dans le cas gaussien, on obtient immédiatement :

$$\ln p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_0^2) = \ln p(\mathbf{e} \mid \sigma_0^2) = -1/2(N \ln 2\pi + N \ln \sigma_0^2 + \mathbf{e}'\mathbf{e}/\sigma_0^2), \quad (71a)$$

$$\ln p(\mathbf{u} \mid \sigma_1^2) = -1/2(q \ln 2\pi + q \ln \sigma_1^2 + \mathbf{u}'\mathbf{u}/\sigma_1^2). \quad (71b)$$

En désignant par $L_0(\sigma_0^2; \mathbf{e}) = \ln p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_0^2)$, la logvraisemblance de $\sigma_0^2$ basée sur $\mathbf{e}$, et par $L_1(\sigma_1^2; \mathbf{u}) = \ln p(\mathbf{u} \mid \sigma_1^2)$, celle de $\sigma_1^2$ basée sur $\mathbf{u}$, la logvraisemblance de $\boldsymbol{\phi}$ basée sur $\mathbf{x}$ se partitionne ainsi en deux composantes qui ne font intervenir chacune qu'un des deux paramètres :

$$L(\boldsymbol{\phi}; \mathbf{x}) = L_0(\sigma_0^2; \mathbf{e}) + L_1(\sigma_1^2; \mathbf{u}) + cste. \quad (72)$$

Cette propriété de séparabilité de la logvraisemblance va pouvoir être mise à profit à la phase E lors de l'explicitation de la fonction $Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = \mathrm{E}_c^{[t]}[L(\boldsymbol{\phi}; \mathbf{x})]$ qui, eu égard à (72), se décompose de façon analogue en :

$$Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = Q_0(\sigma_0^2; \boldsymbol{\phi}^{[t]}) + Q_1(\sigma_1^2; \boldsymbol{\phi}^{[t]}), \quad (73)$$

où

$$\begin{aligned} Q_0(\sigma_0^2; \boldsymbol{\phi}^{[t]}) &= \mathrm{E}_c^{[t]}\left[L_0(\sigma_0^2; \mathbf{e})\right] \\ &= -1/2\left[N \ln 2\pi + N \ln \sigma_0^2 + \mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})/\sigma_0^2\right], \quad (74a) \\ Q_1(\sigma_1^2; \boldsymbol{\phi}^{[t]}) &= \mathrm{E}_c^{[t]}\left[L_1(\sigma_1^2; \mathbf{u})\right] \\ &= -1/2\left[q \ln 2\pi + q \ln \sigma_1^2 + \mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u})/\sigma_1^2\right], \quad (74b) \end{aligned}$$

$\mathrm{E}_c^{[t]}(\cdot)$ désignant comme précédemment une espérance prise par rapport à la loi conditionnelle de $\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}$.

La phase M consiste en la maximisation de $Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})$ par rapport à $\boldsymbol{\phi}$, soit, compte tenu de (73), en la maximisation de $Q_0(\sigma_0^2; \boldsymbol{\phi}^{[t]})$ par rapport à $\sigma_0^2$ et en celle de $Q_1(\sigma_1^2; \boldsymbol{\phi}^{[t]})$ par rapport à $\sigma_1^2$. Les derivées premières de ces fonctions s'écrivent :

$$\frac{\partial(-2Q_0)}{\partial \sigma_0^2} = \frac{N}{\sigma_0^2} - \frac{\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})}{\sigma_0^4}, \quad (75a)$$

$$\frac{\partial(-2Q_1)}{\partial \sigma_1^2} = \frac{q}{\sigma_1^2} - \frac{\mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u})}{\sigma_1^4}. \quad (75b)$$

Leur annulation conduit immédiatement à :

$$\sigma_0^{2[t+1]} = \mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})/N, \tag{76a}$$

$$\sigma_1^{2[t+1]} = \mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u})/q, \tag{76b}.$$

Ce développement a été effectué de façon complète, étape par étape, pour des raisons pédagogiques. En fait, ces résultats extrêmement simples auraient pu être obtenus directement en se référant :

1) à une autre définition des données complètes n'incluant pas explicitement les données observées mais $\mathbf{x} = (\boldsymbol{\beta}', \mathbf{u}', \mathbf{e}')'$ (*cf.* § 1.3 «formulation de l'algorithme») ;

2) aux statistiques exhaustives $\mathbf{e}'\mathbf{e}$ de $\sigma_0^2$ et $\mathbf{u}'\mathbf{u}$ de $\sigma_1^2$, puis en égalant les espérances de celles-ci à leurs espérances conditionnelles respectives soit $\mathrm{E}(\mathbf{e}'\mathbf{e} \,|\, \sigma_0^{2[t+1]}) = \mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})$ et $\mathrm{E}(\mathbf{u}'\mathbf{u} \,|\, \sigma_1^{2[t+1]}) = \mathrm{E}_c^{(t)}(\mathbf{u}'\mathbf{u})$.

Sur la base des formules (76ab), on note dès à présent que les itérations EM qui font intervenir l'espérance de formes quadratiques définies positives, resteront donc à l'intérieur de l'espace paramétrique et c'est là une propriété importante de l'algorithme EM. Il reste maintenant à expliciter $\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})$ et $\mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u})$. Commençons donc par cette dernière forme qui est la plus simple. Par définition

$$\mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u}) = \mathrm{E}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})'\mathrm{E}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}) + \mathrm{tr}\big[\mathrm{var}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})\big]. \tag{77}$$

Or,

$$\mathrm{E}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}) = \widehat{\mathbf{u}}^{[t]} \tag{78}$$

est le BLUP [3] de $\mathbf{u}$ basé sur $\boldsymbol{\phi}^{[t]} = (\sigma_0^{2[t]}, \sigma_1^{2[t]})$. Par définition, le BLUP a pour expression $\widehat{\mathbf{u}} = \mathrm{Cov}(\mathbf{u}, \mathbf{y}')\big[\mathrm{Var}(\mathbf{y})\big]^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$ où $\widehat{\boldsymbol{\beta}}$ est l'estimateur des moindres carrés généralisés. On peut aussi l'obtenir indirectement (et avantageusement) par résolution du système des équations du modèle mixte suivant (Henderson, 1973, 1984)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}}^{[t]} \\ \widehat{\mathbf{u}}^{[t]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \tag{79}$$

où $\lambda^{[t]} = \sigma_0^{2[t]}/\sigma_1^{2[t]}$.

De même,

$$\mathrm{var}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}) = \mathrm{var}(\widehat{\mathbf{u}}^{[t]} - \mathbf{u}) = \mathbf{C}_{uu}^{[t]}\sigma_0^{2[t]}, \tag{80}$$

3. Abréviation de «Best Linear Unbiased Prediction»

où $\mathbf{C}_{uu}^{[t]}$ est le bloc relatif aux effets aléatoires dans l'inverse de la matrice des coefficients des équations d'Henderson soit

$$\mathbf{C}^{[t]} = \begin{bmatrix} \mathbf{C}_{\beta\beta}^{[t]} & \mathbf{C}_{\beta u}^{[t]} \\ \mathbf{C}_{u\beta}^{[t]} & \mathbf{C}_{uu}^{[t]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q \end{bmatrix}^{-} \tag{81}$$

En reportant (78) et (80) dans (77) puis dans (76b), il vient :

$$\sigma_1^{2[t+1]} = \left[\widehat{\mathbf{u}}^{[t]\prime}\widehat{\mathbf{u}}^{[t]} + \operatorname{tr}(\mathbf{C}_{uu}^{[t]})\sigma_0^{2[t]}\right]/q. \tag{82}$$

Le même raisonnement s'applique à l'expression de $\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})$, soit

$$\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e}) = \mathrm{E}(\mathbf{e}\,|\,\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})'\mathrm{E}(\mathbf{e}\,|\,\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}) + \operatorname{tr}\left[\operatorname{var}(\mathbf{e}\,|\,\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})\right].$$

Posons $\mathbf{T} = (\mathbf{X}, \mathbf{Z})$ et $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}')'$, les moments de la distribution conditionnelle de $\mathbf{e}$ s'écrivent :

$$\mathrm{E}(\mathbf{e}\,|\,\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}) = \mathbf{y} - \mathbf{T}\mathrm{E}(\boldsymbol{\theta}\,|\,\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}) = \mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}}^{[t]}, \tag{83a}$$

$$\operatorname{var}(\mathbf{e}\,|\,\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}) = \mathbf{T}\operatorname{var}(\boldsymbol{\theta}\,|\,\mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]})\mathbf{T}' = \mathbf{T}\mathbf{C}^{[t]}\mathbf{T}'\sigma_0^{2[t]}, \tag{83b}$$

où $\widehat{\boldsymbol{\theta}}^{[t]} = \left(\widehat{\boldsymbol{\beta}}^{[t]\prime}, \widehat{\mathbf{u}}^{[t]\prime}\right)'$ est solution du système (79) et $\mathbf{C}^{[t]}$ une inverse généralisée (81) de la matrice des coefficients.

On montre par manipulation matricielle (*cf.* annexe B) que :

$$(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}}^{[t]})'(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}}^{[t]}) = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\theta}}^{[t]\prime}\mathbf{T}'\mathbf{y} - \lambda^{[t]}\widehat{\mathbf{u}}^{[t]\prime}\widehat{\mathbf{u}}^{[t]}, \tag{84a}$$

$$\operatorname{tr}(\mathbf{C}^{[t]}\mathbf{T}'\mathbf{T}) = \operatorname{rang}(\mathbf{X}) + q - \lambda^{[t]}\operatorname{tr}(\mathbf{C}_{uu}^{[t]}). \tag{84b}$$

d'où l'on déduit l'expression de $\sigma_0^{2[t+1]}$,

$$\sigma_0^{2[t+1]} = \left\{\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\theta}}^{[t]\prime}\mathbf{T}'\mathbf{y} - \lambda^{[t]}\widehat{\mathbf{u}}^{[t]\prime}\widehat{\mathbf{u}}^{[t]} + \left[\operatorname{rang}(\mathbf{X}) + q - \lambda^{[t]}\operatorname{tr}(\mathbf{C}_{uu}^{[t]})\right]\sigma_0^{2[t]}\right\}\Big/N. \tag{85}$$

On note au passage que cette expression diffère de celle de l'algorithme d'Henderson (1973) qui s'écrit simplement $\sigma_0^{2[t+1]} = (\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\theta}}^{[t]\prime}\mathbf{T}'\mathbf{y})/[N - r(\mathbf{X})]$, alors que les formules sont identiques pour $\sigma_1^{2[t+1]}$. En fait, les formules d'Henderson peuvent s'interpréter dans le cadre EM comme une variante dérivée d'une forme ECME (Foulley et van Dyk, 2000).

Ces expressions se généralisent immédiatement au cas de plusieurs facteurs aléatoires $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2\mathbf{I}_{q_k})$ ; $(k = 1, 2, \ldots, K)$ non corrélés tels que $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sum_{k=1}^{K}\mathbf{Z}_k\mathbf{Z}_k'\sigma_k^2 + \mathbf{I}_N\sigma_0^2)$. On a alors :

$$\boxed{\sigma_k^{2[t+1]} = \left[\widehat{\mathbf{u}}^{[t]\prime}\widehat{\mathbf{u}}^{[t]} + \operatorname{tr}(\mathbf{C}_{kk}^{[t]})\sigma_0^{2[t]}\right]\Big/q_k}, \tag{86}$$

et

$$\sigma_0^{2[t+1]} = \left\{ \mathbf{y}'\mathbf{y} - \widehat{\theta}^{[t]\prime}\mathbf{T}'\mathbf{y} - \sum_{k=1}^{K} \lambda_k^{[t]} \widehat{\mathbf{u}}_k^{[t]\prime} \widehat{\mathbf{u}}_k^{[t]} + \left[ \operatorname{rang}(\mathbf{X}) + \sum_{k=1}^{K} q_k - \sum_{k=1}^{K} \lambda_k^{[t]} \operatorname{tr}(\mathbf{C}_{kk}^{[t]}) \right] \sigma_0^{2[t]} \right\} \Big/ N \tag{87}$$

Une des difficultés d'application de cet algorithme réside dans la nécessité de calculer le terme $\operatorname{tr}(\mathbf{C}_{uu}^{[t]})$ à chaque itération $[t]$. En fait on peut écrire $\mathbf{C}_{uu}$ sous la forme : $\mathbf{C}_{uu} = (\mathbf{Z}'\mathbf{S}\mathbf{Z} + \lambda \mathbf{I}_q)^{-1}$ où $\mathbf{S} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ est le projecteur classique sur l'espace orthogonal à celui engendré par les colonnes de $\mathbf{X}$. Désignons par $\delta_i(\mathbf{B})$ la ième valeur propre de la matrice $\mathbf{B}$, on sait que $\operatorname{tr}(\mathbf{C}_{uu}) = \sum_i \delta_i^{-1}(\mathbf{Z}'\mathbf{S}\mathbf{Z} + \lambda \mathbf{I}_q)$ et que la ième valeur propre de $\mathbf{Z}'\mathbf{S}\mathbf{Z} + \lambda \mathbf{I}_q$ s'obtient par une simple translation de celle correspondante de $\mathbf{Z}'\mathbf{S}\mathbf{Z}$ soit $\delta_i(\mathbf{Z}'\mathbf{S}\mathbf{Z} + \lambda \mathbf{I}_q) = \delta_i(\mathbf{Z}'\mathbf{S}\mathbf{Z}) + \lambda$, d'où $\operatorname{tr}(\mathbf{C}_{uu}) = \sum_i \left[\delta_i(\mathbf{Z}'\mathbf{S}\mathbf{Z}) + \lambda\right]^{-1}$. Le calcul des valeurs propres de $\mathbf{Z}'\mathbf{S}\mathbf{Z}$ peut donc être réalisé une fois pour toutes en ayant recours à une diagonalisation ou une tridiagonalisation (Smith et Graser, 1986).

*2.2.2. EM-ML*

Si l'on veut obtenir des estimations ML des composantes de la variance, il va falloir considérer $\boldsymbol{\beta}$ comme un paramètre et non plus comme une variable aléatoire. On définit ainsi le vecteur des paramètres par $\boldsymbol{\phi} = (\sigma_u^2, \sigma_e^2, \boldsymbol{\beta}')'$ et celui $\mathbf{z}$ des données manquantes par $\mathbf{z} = \mathbf{u}$. On décompose la densité des données complètes $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$ comme précédemment de sorte que

$$L(\boldsymbol{\phi}; \mathbf{x}) + L_0(\sigma_0^2, \boldsymbol{\beta}; \mathbf{e}) + L_1(\sigma_1^2; \mathbf{u}) + cste, \tag{88}$$

avec

$$-2L_0(\sigma_0^2; \boldsymbol{\beta}; \mathbf{e}) = N \ln 2\pi + N \ln \sigma_0^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})/\sigma_0^2,$$

et

$$-2L_1(\sigma_1^2; \mathbf{u}) = q \ln 2\pi + q \ln \sigma_1^2 + \mathbf{u}'\mathbf{u}/\sigma_1^2.$$

L'expression de $Q_1(\sigma_1^2; \boldsymbol{\phi}^{[t]})$ reste formellement inchangée si bien que, comme précédemment, $\sigma_1^{2[t+1]} = \mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u})/q$. En ce qui concerne $Q_0(\sigma_0^2, \boldsymbol{\beta}; \boldsymbol{\phi}^{[t]})$, son expression s'explicite sous la forme suivante :

$$\begin{aligned} -2Q_0(\sigma_0^2, \boldsymbol{\beta}; \boldsymbol{\phi}^{[t]}) = {} & N \ln 2\pi + N \ln \sigma_0^2 \\ & + \left[\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathrm{E}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi}^{[t]})\right]' \left[\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathrm{E}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi}^{[t]})\right] / \sigma_0^2 \\ & + \operatorname{tr}\left[\mathbf{Z} \operatorname{var}(\mathbf{u} \,|\, \mathbf{y}, \boldsymbol{\phi}^{[t]}) \mathbf{Z}'\right] / \sigma_0^2 \end{aligned} \tag{89}$$

En dérivant par rapport à $\boldsymbol{\beta}$, on obtient :

$$\partial(-2Q_0)/\partial\boldsymbol{\beta} = -2\mathbf{X}'\left[\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \boldsymbol{\phi}^{[t]})\right]/\sigma_0^2. \tag{90}$$

Par annulation, l'équation obtenue ne dépend pas de $\sigma_0^2$ et on peut résoudre en $\boldsymbol{\beta}$ :

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^{[t+1]} = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Z}\mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \boldsymbol{\phi}^{[t]}). \tag{91}$$

En fait, $\mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \boldsymbol{\phi}^{[t]})$ correspond dans ce cas à ce qu'on appelle le meilleur prédicteur linéaire (BLP selon la terminologie d'Henderson) soit $\mathrm{E}(\mathbf{u}\,|\,y, \boldsymbol{\phi}) = \mathrm{Cov}(\mathbf{u}, \mathbf{y}')\left[\mathrm{Var}(\mathbf{y})\right]^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ ou encore, dans nos notations (*cf.* paragraphe 2.2.1), $\mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \boldsymbol{\phi}) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Comme $\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$ (Henderson, 1984), $\mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \boldsymbol{\phi}^{[t]})$ peut s'obtenir simplement à partir du système suivant du type « équations du modèle mixte »

$$\mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \boldsymbol{\phi}^{[t]}) = (\mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q)^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t]}). \tag{92}$$

On peut aussi pour simplifier les calculs résoudre ces deux équations simultanément à partir des équations du modèle mixte d'Henderson soit

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}}^{[t+1]} \\ \widehat{\mathbf{u}}^{[t+1]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \tag{93}$$

où $\widehat{\mathbf{u}}^{[t+1]} = \mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \sigma_0^{2[t]}, \sigma_1^{2[t]}, \boldsymbol{\beta}^{[t+1]})$ et $\lambda^{[t]} = \sigma_0^{2[t]}/\sigma_1^{2[t]}$.
Notons que cela revient à actualiser la phase E sur la base de $\boldsymbol{\phi} = (\sigma_0^{2[t]}, \sigma_1^{2[t]}, \boldsymbol{\beta}^{[t+1]'})'$ avant d'avoir terminé la phase M ; il s'agit là d'une variante qui est décrite par Meng et Rubin (1993) à propos de l'algorithme ECM.
On termine ensuite la phase M, tout d'abord en explicitant :

$$\sigma_1^{2[t+1]} = \mathrm{E}(\mathbf{u}'\mathbf{u}\,|\,\mathbf{y}, \sigma_u^{2[t]}, \sigma_e^{2[t]}, \boldsymbol{\beta}^{[t+1]})/q \tag{94}$$

Comme on raisonne conditionnellement à $\boldsymbol{\beta} = \boldsymbol{\beta}^{[t+1]}$, l'expression de la variance de la loi conditionnelle de $\mathbf{u}$ se réduit à

$$\mathrm{var}(\mathbf{u}\,|\,\mathbf{y}, \sigma_1^{2[t]}, \sigma_0^{2[t]}, \boldsymbol{\beta}^{[t+1]}) = (\mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q)^{-1}\sigma_0^{2[t]}, \tag{95}$$

et, en reportant dans $\sigma_1^2 = \mathrm{E}_c(\mathbf{u}'\mathbf{u})/q$, on a :

$$\sigma_1^{2[t+1]} = \left[\widehat{\mathbf{u}}^{[t+1]'}\widehat{\mathbf{u}}^{[t+1]} + \mathrm{tr}(\mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q)^{-1}\sigma_0^{2[t]}\right]/q. \tag{96}$$

Par dérivation de (89) par rapport à $\sigma_0^2$, et en annulant, il vient :

$$\sigma_0^{2[t+1]} = \left\{\widehat{\mathbf{e}}^{[t+1]'}\widehat{\mathbf{e}}^{[t+1]} + \mathrm{tr}\left[\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q)^{-1}\mathbf{Z}'\right]\sigma_0^{2[t]}\right\}/N$$

où $\widehat{\mathbf{e}}^{[t+1]} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathrm{E}(\mathbf{u}\,|\,\mathbf{y}, \sigma_1^{2[t]}, \sigma_0^{2[t]}, \boldsymbol{\beta}^{[t+1]}) = \mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}}^{[t+1]}$.

Cette expression se simplifie à nouveau compte tenu de la relation (84a) et de ce que

$$\mathrm{tr}\left[(\mathbf{Z}'\mathbf{Z}+\lambda^{[t]}\mathbf{I}_q)^{-1}\mathbf{Z}'\mathbf{Z}\right] = q - \lambda^{[t]}\mathrm{tr}\left[(\mathbf{Z}'\mathbf{Z}+\lambda^{[t]}\mathbf{I}_q)^{-1}\right],$$

d'où

$$\boxed{\begin{aligned}\sigma_0^{2[t+1]} = \Big\{&\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\theta}}^{[t+1]\prime}\mathbf{T}'\mathbf{y} - \lambda^{[t]}\widehat{\mathbf{u}}^{[t+1]\prime}\widehat{\mathbf{u}}^{[t+1]} \\ &+ \left(q - \lambda^{[t]}\mathrm{tr}\left[(\mathbf{Z}'\mathbf{Z}+\lambda^{[t]}\mathbf{I}_q)^{-1}\right]\sigma_0^{2[t]}\right)\Big\}\Big/N\end{aligned}} \tag{97}$$

La différence entre ML et REML apparaît donc nettement au niveau de l'algorithme EM ; les calculs seront moins pénibles à réaliser avec ML puisqu'il ne faut plus disposer de l'inverse complète des équations du modèle mixte mais simplement de la partie aléatoire. Enfin, en ce qui concerne ML, d'autres variantes de type ECME ont été décrites par Liu et Rubin (1994).

*2.2.3. « Scaled» EM*

L'idée de base réside dans la standardisation des effets aléatoires. Dans le cas d'un seul facteur, cela revient à écrire le modèle sous la forme : $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma_1\mathbf{Z}\mathbf{u}^* + \mathbf{e}$ où $\mathbf{u}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$, le reste étant inchangé. Si l'on définit les données complètes par $\mathbf{x} = (\mathbf{y}', \boldsymbol{\beta}', \mathbf{u}^{*\prime})'$, on a $p(\mathbf{x} \mid \boldsymbol{\phi}) \propto p(\mathbf{e} \mid \boldsymbol{\phi})$ puisque la densité $p(\boldsymbol{\beta}, \mathbf{u}^*)$ est non informative vis-à-vis des paramètres. À la phase E, la fonction $Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})$ s'écrit donc :

$$-2Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = N\ln 2\pi + N\ln\sigma_0^2 + \mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})/\sigma_0^2. \tag{98}$$

À la phase M, il vient par dérivation :

$$\frac{\partial(-2Q)}{\partial\sigma_0^2} = \frac{N}{\sigma_0^2} - \frac{\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})}{\sigma_0^4},$$

$$\frac{\partial(-2Q)}{\partial\sigma_1} = \frac{1}{\sigma_0^2}\,\frac{\partial\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})}{\partial\sigma_1} = -\frac{2\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{Z}\mathbf{u}^*)}{\sigma_0^2}.$$

Rien ne change donc formellement pour l'actualisation de $\sigma_0^2$. Par contre en ce qui concerne $\sigma_1$, l'annulation de la dérivée conduit à l'expression suivante :

$$\boxed{\sigma_1^{[t+1]} = \frac{\mathrm{E}_c^{[t]}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{Z}\mathbf{u}^*\right]}{\mathrm{E}_c^{[t]}(\mathbf{u}^{*\prime}\mathbf{Z}'\mathbf{Z}\mathbf{u}^*)}} \tag{99}$$

dont la forme s'apparente à celle d'un coefficient de régression.

Comme précédemment, le numérateur et le dénominateur de (99) peuvent s'exprimer à partir des ingrédients des équations du modèle mixte d'Henderson soit, en ignorant l'indice $t$ pour alléger les notations :

$$\mathrm{E}_c^{[t]}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{Z}\mathbf{u}^*\right] = (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{Z}\widehat{\mathbf{u}}^* - \mathrm{tr}(\mathbf{X}'\mathbf{Z}\widetilde{\mathbf{C}}_{u\beta})\sigma_0^2, \tag{100a}$$

$$\mathrm{E}_c^{[t]}(\mathbf{u}^{*\prime}\mathbf{Z}'\mathbf{Z}\mathbf{u}^*) = \widehat{\mathbf{u}}^{*\prime}\mathbf{Z}'\mathbf{Z}\widehat{\mathbf{u}}^* + \mathrm{tr}(\mathbf{Z}'\mathbf{Z}\widetilde{\mathbf{C}}_{uu})\sigma_0^2, \qquad (100b)$$

où $\widehat{\boldsymbol{\beta}}$ et $\widehat{\mathbf{u}}^*$ sont solutions du système :

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\sigma_1 \\ \mathbf{Z}'\mathbf{X}\sigma_1 & \mathbf{Z}'\mathbf{Z}\sigma_1^2 + \mathbf{I}_q\sigma_0^2 \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \sigma_1\mathbf{Z}'\mathbf{y} \end{bmatrix},$$

$$\widetilde{\mathbf{C}} = \begin{bmatrix} \widetilde{\mathbf{C}}_{\beta\beta} & \widetilde{\mathbf{C}}_{\beta u} \\ \widetilde{\mathbf{C}}_{u\beta} & \widetilde{\mathbf{C}}_{uu} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\sigma_1 \\ \mathbf{Z}'\mathbf{X}\sigma_1 & \mathbf{Z}'\mathbf{Z}\sigma_1^2 + \mathbf{I}_q\sigma_0^2 \end{bmatrix}^- .$$

On peut aussi résoudre les équations du modèle mixte sous leur forme habituelle (*cf.* 77) puis calculer $\widehat{\mathbf{u}}^* = \widehat{\mathbf{u}}/\sigma_1$, $\widehat{\mathbf{C}}_{\beta u} = \mathbf{C}_{\beta u}/\sigma_1$ et $\widetilde{\mathbf{C}}_{uu} = \mathbf{C}_{uu}/\sigma_1^2$. Cet algorithme à effets normalisés se distingue également de l'algorithme classique de forme quadratique par ses performances (Thompson, 2002). Cette comparaison a été effectuée par Foulley et Quaas (1995) dans le cas d'un modèle d'analyse de variance équilibré à un facteur aléatoire (ici la famille de demi-frères). Alors que l'algorithme classique est très lent pour des valeurs faibles du rapport $R^2 = n/(n+\alpha)$ ($\alpha$ désignant ici le ratio $\sigma_0^2/\sigma_1^2$), par exemple $R^2 = 1/4$ ($n = 5$; $\alpha = 15$) et beaucoup plus rapide pour des valeurs élévées, par exemple $R^2 = 0,95$ ($n = 285$, $\alpha = 15$; $n = 1881$, $\alpha = 99$) la tendance est opposée en ce qui concerne l'EM normalisé (*cf.* Fig. 1).

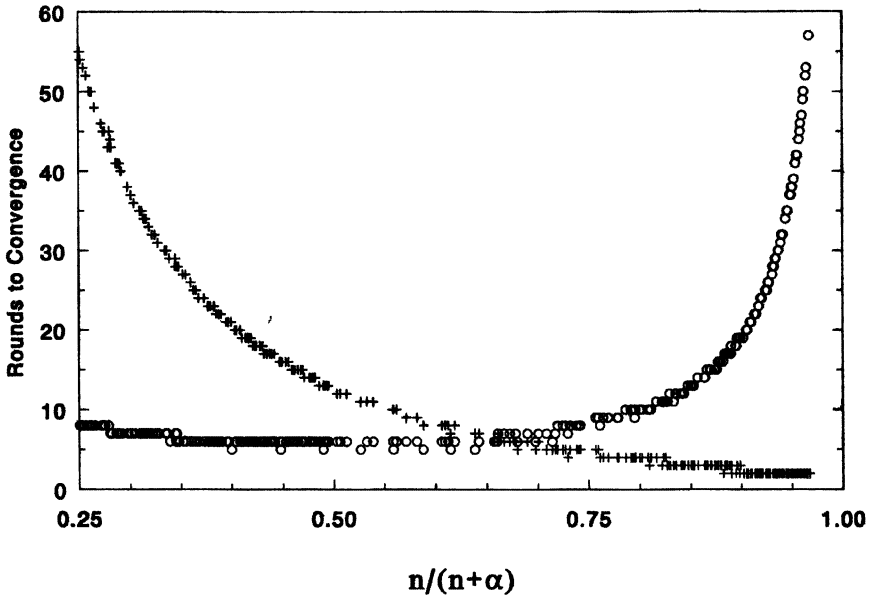


FIG 1. – Vitesse de convergence (nombre d'itérations) pour les algorithmes EM classique (croix) et «Scaled» (ronds) dans un dispositif de 100 familles de demi-frères de même taille (n) en fonction du rapport $R^2 = u/(u+\alpha)$ où $\alpha = \sigma_0^2/\sigma_1^2$ est le ratio de la variance résiduelle à la variance entre familles.

Ces auteurs ont montré également que, tout comme avec l'EM classique, les itérations restent dans l'espace paramétrique. Cette idée de la standardisation des effets aléatoires qui figure déjà dans Anderson et Aitkin (1985), a été reprise puis généralisée par Meng et van Dyk (1998) au cas où la matrice de variance covariance des effets aléatoires n'est plus diagonale : *cf.* aussi Wolfinger et Tobias (1998). Enfin, l'algorithme précédent peut être adapté facilement au cas d'une estimation ML (Foulley, 1997).

### *2.2.4. Variances hétérogènes*

Pour le modèle mixte, on fait généralement l'hypothèse d'homogénéité des composantes de variance $\mathbf{G}$ et $\mathbf{R}$, mais celle-ci n'est pas indispensable et s'avère d'ailleurs souvent démentie par les faits expérimentaux. Ainsi, dans une analyse génétique familiale, la variance entre familles ($\sigma_1^2$) tout comme la variance intra-familles ($\sigma_0^2$) dépend fréquemment des conditions de milieu dans lesquelles sont élevés les individus. Il en est de même dans une analyse longitudinale avec un modèle à coefficients aléatoires où les éléments de la matrice $\mathbf{G}$ ($g_{00}$ : variance de l'intercept aléatoire; $g_{11}$ : variance de la pente; $g_{01}$ : covariance entre la pente et l'intercept) vont différer selon certaines caractéristiques des individus (par ex. sexe, traitement, type d'activité, etc...).

Ce phénomène dit d'hétéroscédasticité peut être pris en compte dans le modèle mixte grâce à une formalisation du type suivant :

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sigma_{1,i}\mathbf{Z}_i\mathbf{u}^* + \mathbf{e}_i, \tag{101}$$

où $\mathbf{y}_i = \{y_{ij}\}$ est le vecteur $(n_i \times 1)$ des observations dans la strate $i = 1, 2, \ldots, I$; $\boldsymbol{\beta}$ est le vecteur $(p \times 1)$ des effets fixes associé à la matrice $(n_i \times p)$ de covariables $\mathbf{X}_i$. Comme dans la formulation de l'EM normalisé, la contribution des effets aléatoires est exprimée sous la forme $\sigma_{1,i}\mathbf{Z}_i\mathbf{u}^*$ où $\mathbf{u}^*$ est un vecteur d'effets aléatoires standardisés, $\mathbf{Z}_i$ la matrice $(n_i \times q)$ d'incidence correspondante et $\sigma_{1,i}$ est la racine carrée de la composante $\mathbf{u}$ de la variance dont la valeur dépend de la strate $i$ de la population. On fait par ailleurs les hypothèses classiques sur les distributions à savoir : $\mathbf{u}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ (les généticiens remplacent la matrice identité $\mathbf{I}_q$ par une matrice de parenté $\mathbf{A}$), $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma_{0,i}^2\mathbf{I}_{n_i})$ et$\mathrm{E}(\mathbf{u}^*\mathbf{e}_i')$, $\forall i$.

Quand la stratification est simple (un seul facteur par exemple), le modèle (101) peut être abordé tel que. En fait, dès l'instant où plusieurs facteurs se trouvent mis en cause dans l'hétéroscédasticité, il devient souhaitable de modéliser l'influence de ceux-ci sur les composantes de variance $(\sigma_{0,i}^2, \sigma_{1,i}^2)$. Une des façons les plus simples de procéder est d'avoir recours à un modèle structural de type linéaire généralisé impliquant la fonction de lien logarithmique (Leonard, 1975; Aitkin, 1987; Nair et Pregibon, 1988; Foulley *et al.*, 1992; San Cristobal *et al.*, 2002). Comme l'a bien montré Robert (1996) dans l'étude des mélanges, il peut être intéressant pour des raisons numériques, de substituer, à une paramétrisation des deux variances, une paramétrisation impliquant l'une d'entre elles, la plus facile à estimer (ici $\sigma_{0,i}^2$), et le rapport de l'autre à celle-ci (ici on prend le rapport des écarts types $\tau_i = \sigma_{1,i}/\sigma_{0,i}$).

On écrit alors, à l'instar de Foulley (1997),

$$\ln \sigma^2_{0,i} = \mathbf{p}'_i \boldsymbol{\delta}, \tag{102a}$$

$$\ln \tau_i = \mathbf{h}'_i \boldsymbol{\lambda}, \tag{102b}$$

où $\boldsymbol{\delta}$ est le vecteur $(r \times 1)$ des coefficients réels des $r$ variables explicatives $\mathbf{p}_i$ influençant le logarithme de la variance résiduelle relative à la strate $i$; idem pour le vecteur $\boldsymbol{\lambda}(s \times 1)$ des coefficients des variables explicatives $\mathbf{h}_i$ du logarithme du ratio $\tau_i$ des écarts types.

Si l'on pose $\boldsymbol{\phi} = (\boldsymbol{\delta}', \boldsymbol{\lambda}')'$ et $\mathbf{x} = (\boldsymbol{\beta}', \mathbf{u}^{*\prime})'$, la phase E conduit comme précédemment à :

$$-2Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = N \ln 2\pi + \sum_{i=1}^{I} n_i \ln \sigma^2_{0,i} + \sum_{i=1}^{I} \mathrm{E}_c^{[t]}(\mathbf{e}'_i \mathbf{e}_i)/\sigma^2_{0,i}, \tag{103}$$

où

$$\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \tau_i \sigma_{0,i} \mathbf{Z}_i \mathbf{u}^*.$$

En l'absence d'expression explicite des maxima, on a recours à une version « gradient-EM » de l'algorithme *via*, par exemple, la formule de Newton-Raphson (*cf.* 61)

$$-\ddot{Q}(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{[t,k]}} (\boldsymbol{\phi}^{[t,k+1]} - \boldsymbol{\phi}^{[t,k]}) = \dot{Q}(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{[t,k]}}.$$

Ayant calculé les dérivées partielles première et seconde par rapport aux paramètres (*cf.* annexe C), le système des équations à résoudre peut se mettre sous la forme itérative suivante :

$$\begin{bmatrix} \mathbf{P}'\mathbf{W}_{\delta\delta}\mathbf{P} & \mathbf{P}'\mathbf{W}_{\delta\lambda}\mathbf{H} \\ \mathbf{H}'\mathbf{W}_{\lambda\delta}\mathbf{P} & \mathbf{H}'\mathbf{W}_{\lambda\lambda}\mathbf{H} \end{bmatrix}_{\boldsymbol{\phi}=\boldsymbol{\phi}^{[t,k]}} \begin{bmatrix} \Delta\boldsymbol{\delta} \\ \Delta\boldsymbol{\lambda} \end{bmatrix}^{[t,k+1]} = \begin{bmatrix} \mathbf{P}'\mathbf{v}_\delta \\ \mathbf{H}'\mathbf{v}_\lambda \end{bmatrix}_{\boldsymbol{\phi}=\boldsymbol{\phi}^{[t,k]}}, \tag{104}$$

où

$$\Delta\boldsymbol{\delta}^{[t,k+1]} = \boldsymbol{\delta}^{[t,k+1]} - \boldsymbol{\delta}^{[t,k]}, \Delta\boldsymbol{\lambda}^{[t,k+1]} = \boldsymbol{\lambda}^{[t,k+1]} - \boldsymbol{\lambda}^{[t,k]},$$
$$\mathbf{P}'_{(R\times 1)} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_i, \ldots, \mathbf{p}_I),\ \mathbf{H}'_{(ss\times I)} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_i, \ldots, \mathbf{h}_I).$$

Les éléments de $\mathbf{v}_\delta$, $\mathbf{v}_\lambda$ s'écrivent, en ignorant les indices $[t,k]$ pour alléger les notations :

$$\mathbf{v}_{\delta(I\times 1)} = \{v_{\delta,i} = 1/2(\sigma^{-2}_{0,i} \mathrm{E}_c[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i] - n_i)\}, \tag{105a}$$

$$\mathbf{v}_{\lambda(I\times 1)} = \{v_{\lambda,i} = \tau_i \sigma^{-1}_{0,i} \mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}'_i\mathbf{e}_i)\}. \tag{105b}$$

Les matrices de pondération $\mathbf{W}_{\delta\delta}$, $\mathbf{W}_{\delta\lambda} = \mathbf{W}_{\lambda\delta}$ et $\mathbf{W}_{\lambda\lambda}$ sont des matrices diagonales $(I \times I)$ dont les éléments s'explicitent en

$$w_{\delta\delta,ii} = 1/2\sigma^{-2}_{0,i}\{\mathrm{E}_c[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})] - \tau_i\sigma_{0,i}\mathrm{E}_c[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*]/2\}, \tag{106a}$$

$$w_{\delta\lambda,ii} = 1/2\tau_i\sigma^{-1}_{0,i}\mathrm{E}_C[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*], \tag{106b}$$

$$w_{\lambda\lambda,ii} = \tau_i\{2\tau_i \mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i\mathbf{u}^*) - \sigma_{0,i}^{-1}\mathrm{E}_c[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*]\}. \quad (106c)$$

Tous les éléments décrits en (105ab) et (106abc) peuvent s'obtenir aisément à partir des ingrédients des équations du modèle mixte d'Henderson soit, en posant $S_{i,\varepsilon\varepsilon} = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'(\mathbf{y}_i - \mathbf{X}_i, \boldsymbol{\beta})$, $S_{i,\varepsilon u} = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*$ et $S_{i,uu} = \mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i\mathbf{u}^*$,

$$\widehat{S}_{i,\varepsilon\varepsilon} = \mathrm{E}_c(S_{i,\varepsilon\varepsilon}) = (\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})'(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) + \mathrm{tr}(\mathbf{X}_i'\mathbf{X}_i\mathbf{C}_{\beta\beta}), \quad (107a)$$

$$\widehat{S}_{i,\varepsilon u} = \mathrm{E}_c(S_{i,\varepsilon u}) = (\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})'\mathbf{Z}_i\widehat{\mathbf{u}}^* + \mathrm{tr}(\mathbf{Z}_i'\mathbf{X}_i\mathbf{C}_{\beta u}), \quad (107b)$$

$$\widehat{S}_{i,uu} = \mathrm{E}_c(S_{i,uu}) = \widehat{\mathbf{u}}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i\widehat{\mathbf{u}}^* + \mathrm{tr}(\mathbf{Z}_i'\mathbf{Z}_i\mathbf{C}_{uu}). \quad (107c)$$

Ici les équations du modèle mixte s'écrivent $\left(\sum_{i=1}^I \sigma_{0,i}^{-2}\mathbf{T}_i'\mathbf{T}_i + \boldsymbol{\Sigma}^-\right)\widehat{\boldsymbol{\theta}} = \sum_{i=1}^I \sigma_{0,i}^{-2}\mathbf{T}_i'\mathbf{y}_i$ avec $\mathbf{T}_i = (\mathbf{X}_i, \tau_i\sigma_{0,i}\mathbf{Z}_i)$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}^{*\prime})'$, $\boldsymbol{\Sigma}^- = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1} \end{pmatrix}$ et les termes $\mathbf{C}_{\beta\beta}$, $\mathbf{C}_{\beta u}$ et $\mathbf{C}_{uu}$ sont les blocs ainsi indicés dans l'inverse de la matrice des coefficients.

On peut également développer une version des scores de Fisher de cet algorithme en exploitant le fait que $\mathrm{E}_{\mathbf{y}}[\mathrm{E}(S \mid \mathbf{y}, \boldsymbol{\phi})] = \mathrm{E}(S)$ dont l'expression est particulièrement simple dans les cas abordés ici, soit

$$\begin{aligned} \overline{w}_{\delta\delta,ii} &= 1/2[n_i + \tau_i^2\,\mathrm{tr}(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i)/2], \\ \overline{w}_{\delta\lambda,ii} &= \tau_i^2\,\mathrm{tr}(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i)/2, \\ \overline{w}_{\lambda\lambda,ii} &= \tau_i^2\,\mathrm{tr}(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i). \end{aligned}$$

Dans le cas d'un seul facteur aléatoire discret (matrice $\mathbf{Z}_i$ formée de 0 et de 1), la matrice $\mathbf{Z}_i'\mathbf{Z}_i$ est diagonale et, $\mathbf{A}$ ayant des éléments diagonaux unité, $\mathrm{tr}(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i) = n_i$ si bien que tous ces poids se simplifient en $\overline{w}_{\delta\delta,ii} = 1/2n_i(1 + \tau_i^2/2)$ ; $\overline{w}_{\delta\lambda,ii} = n_i\tau_i^2/2$ et $\overline{w}_{\lambda\lambda,ii} = n_i\tau_i^2$.

Une tâche importante va consister à choisir les covariables $\mathbf{P}$ et $\mathbf{H}$ des modèles (107ab) des logvariances *via* par exemple un test du rapport de vraisemblance. Les comparaisons mises en œuvre à cet égard doivent se faire à structure d'espérance $\mathbf{X}\boldsymbol{\beta}$ fixée ; celle-ci en retour sera sélectionnée à structure de variance covariance fixée, ou mieux à partir d'un procédé robuste tel que par exemple celui de Liang et Zeger (1986) en situation de données répétées.

D'autres sous-modèles des variances peuvent être envisagés et testés. En effet, il importe de garder présent à l'esprit la difficulté d'estimer les variances avec précision, notamment les composantes $u$ si l'on ne dispose pas d'un dispositif adéquat et d'un échantillon suffisamment grand, d'où l'intérêt voire la nécessité de modèles parcimonieux. On peut citer à cet égard un modèle à ratio $\tau_i = \sigma_{1,i}/\sigma_{0,i}$ constant (Foulley, 1997), voire un modèle à composante $u$ constante, ces deux modèles étant des variantes d'un modèle plus général de la forme $\sigma_{1,i}/\sigma_{0,i}^b = \mathrm{cste}$ (Foulley *et al.*, 1998).

Par exemple le modèle $\ln\sigma_{1,i}^2 = \mathbf{p}_i'\boldsymbol{\delta}$ et $\sigma_{1,i} = \mathrm{cste}$ conduit au système (Foulley *et al.*, 1992) :

$$(\mathbf{P}'\mathbf{W}_{\delta\delta}\mathbf{P})\Delta\boldsymbol{\delta} = \mathbf{P}'\mathbf{v}_\delta, \quad (108)$$

où

$$\mathbf{v}_{\delta(I\times 1)} = \{v_{\delta,i} = 1/2[\sigma_{0,i}^{-2}\mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i) - n_i]\}, \tag{109a}$$

$$w_{\delta\delta,ii} = 1/2\sigma_{0,i}^{-2}\mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i), \tag{109b}$$

et

$$\overline{w}_{\delta\delta,ii} = 1/2n_i. \tag{109c}$$

Diverses applications de ces modèles mixtes hétéroscédastiques à la génétique animale sont décrits dans Robert *et al.* (1997), Robert *et al.* (1999) ainsi que dans San Cristobal *et al.* (2002).

## 2.3. Modèle à plusieurs facteurs corrélés

### *2.3.1. EM standard (EM0)*

Le cas de plusieurs facteurs aléatoires non corrélés ne pose pas de difficulté particulière et découle d'une généralisation immédiate du cas d'un seul facteur (*cf.* § 221). Le modèle considéré ici s'écrit :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

où le vecteur $\mathbf{u}$ des effets aléatoires et la matrice d'incidence $\mathbf{Z}$ sont les concaténations respectivement des vecteurs $\mathbf{u}_k$ et des matrices d'incidence $\mathbf{Z}_k$ relatifs aux $K$ facteurs élémentaires $k = 1, 2, \ldots, K$ :

$$\mathbf{u}_{(q_+\times 1)} = (\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_k', \ldots, \mathbf{u}_K')'; \mathbf{Z}_{(N\times q_+)} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k, \ldots, \mathbf{Z}_K).$$

Comme à l'accoutumée, ce modèle est tel que $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ où $\mathrm{Var}(\mathbf{u}) = \mathbf{G}$, $\mathrm{Var}(\mathbf{e}) = \mathbf{R}$ et $\mathrm{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$.

On se restreint ici à la classe des modèles dont les $\mathbf{u}_k$ présentent le même nombre d'éléments $q_k = q, \forall k$ et dont la matrice de variance covariance $\mathbf{G}$ s'écrit, par exemple pour $K = 2$ :

$\mathbf{G} = \mathrm{Var}\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \begin{pmatrix}\sigma_{11}\mathbf{I}_q & \sigma_{12}\mathbf{I}_q\\ \sigma_{12}\mathbf{I}_q & \sigma_{22}\mathbf{I}_q\end{pmatrix} = \mathbf{G}_0 \otimes \mathbf{I}_q$ où $\mathbf{G}_0 = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}$ et, de façon générale, $\mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$ avec $\mathbf{G}_0 = \{\sigma_{kl}\}$ pour $k, l = 1, 2, \ldots, K$ et $\mathbf{A}_q = \mathbf{I}_q$ si les unités expérimentales ($i = 1, 2, \ldots, q$; individus, familles) supports des $q$ éléments de chacun des vecteurs $\mathbf{u}_k$ sont indépendantes.

Pour chacune d'entre elles, le modèle s'écrit :

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i, \tag{110}$$

où $\mathbf{y}_i = \{y_{ij}\}$; $j = 1, , \ldots, n_i$ est le vecteur des $n_i$ observations $y_{ij}$ faites sur l'unité expérimentale $i$.

Ici $\mathbf{u}_{i(K\times 1)} = (u_{i1}, u_{i2}, \ldots, u_{ik}, \ldots, u_{iK})'$ et $\mathbf{Z}_{i(n_i\times K)} = [\mathbf{Z}_{i1}, \mathbf{Z}_{i2}, \ldots, \mathbf{Z}_{ik}, \ldots, \mathbf{Z}_{iK}]$ si bien que $\mathbf{u}_{i(K\times 1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_0)$ et $\mathbf{e}_i = \{e_{ij}\} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$ avec $\mathbf{R} = \oplus_{i=1}^{q}\mathbf{R}_i$.

Dans le cas le plus simple de résidus homogènes et indépendants, $\mathbf{R}_i = \sigma_0^2\mathbf{I}_{n_i}$, mais d'autres structures sont envisageables telle que, par exemple, pour

des données longitudinales, une structure autorégressive ou de processus temporel continu stationnaire de type exponentiel : $\mathbf{R}_i = \sigma_0^2 \mathbf{H}_i$ avec $h_{i,jj'} = f(\rho, |t_{ij} - t_{ij'}|)$.

Si l'on pose $\mathbf{g}_0 = \text{vech}(\mathbf{G}_0)$ [5], $\mathbf{r}$ le vecteur des paramètres intervenant dans $\mathbf{R}$ par exemple $\mathbf{r} = \sigma_0^2$ ou $\mathbf{r} = (\sigma_0^2, \rho)'$, $\boldsymbol{\phi} = (\mathbf{g}_0', \mathbf{r}')'$ et $\mathbf{x} = (\boldsymbol{\beta}', \mathbf{u}', \mathbf{e}')'$, on a, comme dans le cas d'un seul facteur aléatoire,

$$\mathrm{L}(\boldsymbol{\phi}; \mathbf{x}) = \mathrm{L}_0(\mathbf{r}; \mathbf{e}) + \mathrm{L}_1(\mathbf{g}_0; \mathbf{u}) + \text{cste}, \tag{111}$$

et

$$\mathrm{Q}(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = \mathrm{Q}_0(\mathbf{r}; \boldsymbol{\phi}^{[t]}) + \mathrm{Q}_1(\mathbf{g}_0; \boldsymbol{\phi}^{[t]}) + \text{cste}. \tag{112}$$

Dans (112),

$$-2\mathrm{Q}_0(\mathbf{r}; \boldsymbol{\phi}^{[t]}) = N \ln 2\pi + \ln|\mathbf{R}| + \mathrm{tr}[\mathbf{R}^{-1} \mathrm{E}_c^{[t]}(\mathbf{e}\mathbf{e}')], \tag{113a}$$

$$2\mathrm{Q}_1(\mathbf{g}_0; \boldsymbol{\phi}^{[t]}) = qK \ln 2\pi + \ln|\mathbf{G}| + \mathrm{tr}[\mathbf{G}^{-1} \mathrm{E}_c^{[t]}(\mathbf{u}\mathbf{u}')],$$

soit, compte tenu du fait que $\mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$

$$-2\mathrm{Q}_1(\mathbf{g}_0; \boldsymbol{\phi}^{[t]}) = qK \ln 2\pi + K \ln|\mathbf{A}| + q \ln|\mathbf{G}_0| + \mathrm{tr}(\mathbf{G}_0^{-1} \Omega^{[t]}), \tag{113b}$$

avec

$$\Omega^{[t]}_{(K \times K)} = \{\varpi_{kl}^{[t]} = \mathrm{E}(\mathbf{u}_k' \mathbf{A}^{-1} \mathbf{u}_l \mid \mathbf{y}, \boldsymbol{\gamma}^{[t]})\}. \tag{114}$$

À la phase M, on maximise (113a) et (113b) par rapport respectivement à $\mathbf{r}$ et $\mathbf{g}$. Par application d'un lemme d'Anderson (1984; page 62, 3.2.2) cela conduit à :

$$\sigma_0^{2[t+1]} = \mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})/N = \left[\sum\nolimits_{i=1}^{N} \mathrm{E}_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i)\right]/N, \tag{115a}$$

$$\mathbf{G}_0^{[t+1]} = \Omega^{[t]}/q. \tag{115b}$$

On s'est limité ici au cas simple où $\mathbf{R} = \sigma_0^2 \mathbf{I}_N$, mais on peut aussi traiter des structures plus complexes comme par exemple celle de l'autorégression (Foulley *et al.*, 2000).

Comme dans le cas monofactoriel, l'espérance des formes quadratriques et bilinéaires intervenant en (108ab) peut s'obtenir à partir des équations du modèle mixte d'Henderson soit ici, à titre d'exemple pour $K = 2$ :

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_1 + \sigma_0^2\sigma^{11}\mathbf{A}^{-1} & \mathbf{Z}_1'\mathbf{Z}_2 + \sigma_0^2\sigma^{12}\mathbf{A}^{-1} \\ \mathbf{Z}_2'\mathbf{X} & \mathbf{Z}_2'\mathbf{Z}_1 + \sigma_0^2\sigma^{21}\mathbf{A}^{-1} & \mathbf{Z}_2'\mathbf{Z}_2 + \sigma_0^2\sigma^{22}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}}_1 \\ \widehat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \\ \mathbf{Z}_2'\mathbf{y} \end{bmatrix}, \tag{116}$$

où $\begin{pmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{21} & \sigma^{22} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1}$.

5. vech($\mathbf{X}$) est la notation de vectorisation d'une matrice, homoloque de vec($\mathbf{X}$), mais qui s'applique à une matrice symétrique, seuls les éléments distinctifs étant pris en compte (Searle, 1982).

*2.3.2. PX-EM*

Pour mettre en œuvre cet algorithme, on introduit des paramètres de travail sous la forme ici d'une matrice $\boldsymbol{\alpha} = \{\alpha_{kl}\}$ carrée $(K \times K)$ réelle inversible telle qu'au modèle d'origine en (110) (dit modèle O) se substitue le nouveau modèle (dit modèle X),

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\alpha}\tilde{\mathbf{u}}_i + \mathbf{e}_i, \tag{117}$$

ou encore, avec une écriture par facteur, $\mathbf{u}_k = \sum_{l=1}^{K} \alpha_{kl}\mathbf{u}_l$.

Par définition, les $\tilde{\mathbf{u}}_i$ sont tels que $\tilde{\mathbf{u}}_{i(K\times 1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_{0^*})$ où $\mathbf{G}_{0^*} = \boldsymbol{\alpha}^{-1}\mathbf{G}_0(\boldsymbol{\alpha}^{-1})'$, la loi des $\tilde{\mathbf{u}}_i$ apparaissant en quelque sorte comme une extension paramétrique de celle des $\mathbf{u}_i$. En particulier, pour la valeur de référence $\boldsymbol{\alpha}_0 = \mathbf{I}_K$, la loi de $\tilde{\mathbf{u}}_i(\boldsymbol{\alpha}_0)$ se réduit à celle de $\mathbf{u}_i$.

Posons $\boldsymbol{\varphi} = [\boldsymbol{\phi}'_*, (\mathrm{vec}\boldsymbol{\alpha})']'$ avec $\boldsymbol{\phi}_* = (\mathbf{g}'_{0^*}, \mathbf{r}'_*)'$ et $\boldsymbol{\varphi}^{[t,0]} = [(\boldsymbol{\phi}_*^{[t]} = \boldsymbol{\phi}^{[t]})', (\mathrm{vec}\,\boldsymbol{\alpha} = \mathrm{vec}\,\boldsymbol{\alpha}_0)']'$. L'étape E consiste en l'explicitation de $\mathrm{Q}(\boldsymbol{\varphi}; \boldsymbol{\varphi}^{[t,0]})$. Le fait de travailler conditionnellement aux paramètres de la loi des $\tilde{\mathbf{u}}_i(\boldsymbol{\alpha}_0)$ offre l'avantage de ne rien changer à l'étape E de l'EM standard (EMO).

La maximisation de $\mathrm{Q}_1(\mathbf{g}_{0^*}; \boldsymbol{\varphi}^{[t,0]})$ par rapport à $\mathbf{g}_{0^*}$ revient à celle de $\mathbf{g}_0$ sous EMO soit

$$\mathbf{G}_{0^*}^{[t+1]} = \Omega^{[t]}/q, \tag{118}$$

où $\Omega^{[t]}$ est le même qu'en (114).

Ensuite, on maximise $\mathrm{Q}_0(\boldsymbol{\alpha}, \mathbf{r}_*; \boldsymbol{\varphi}^{[t,0]})$ dont les dérivées partielles par rapport aux éléments de $\mathrm{vec}(\boldsymbol{\alpha})$ s'écrivent :

$$\frac{\partial \mathrm{Q}_0}{\partial \alpha_{kl}} = \sigma_0^{-2} \sum_{i=1}^{I} \mathrm{E}\left[\mathbf{u}'_i \frac{\partial \boldsymbol{\alpha}}{\partial \alpha_{kl}} \mathbf{Z}'_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\alpha}\tilde{\mathbf{u}}_i) \mid \mathbf{y}, \boldsymbol{\phi}_*^{[t]} = \boldsymbol{\phi}^{[t]}, \boldsymbol{\alpha} = \boldsymbol{\alpha}_0\right],$$

où ici $\mathbf{R}_i = \sigma_0^2 \mathbf{I}_{n_i}$.

La résolution de ces $K^2$ équations ne fait pas intervenir $\sigma_0^2$ et se réduit, à une itération donnée, à celle du système linéaire $\mathbf{F}\mathrm{vec}(\boldsymbol{\alpha}) = \mathbf{h}$, soit encore

$$\sum_{m=1}^{K}\sum_{n=1}^{K} f_{kl,mn}^{[t]} \alpha_{mn}^{[t+1]} = h_{kl}^{[t]}; k, l = 1, 2, \ldots, K \tag{119}$$

où

$$f_{kl,mn}^{[t]} = \mathrm{tr}\left[\mathbf{Z}'_k\mathbf{Z}_m \,\mathrm{E}(\mathbf{u}_n\mathbf{u}'_l) \mid \mathbf{y}, \boldsymbol{\varphi}^{[t,0]}\right], \tag{120a}$$

$$h_{kl}^{[t]} = \mathrm{tr}\left\{\mathbf{Z}'_k \,\mathrm{E}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\mathbf{u}'_l] \mid \mathbf{y}, \boldsymbol{\varphi}^{[t,0]}\right\}. \tag{120b}$$

Soit $\mathbf{T}_{kl} = \mathbf{Z}'_k\mathbf{Z}_l$, and $\mathbf{v}_{k_{(q\times 1)}} = \mathbf{Z}'_k(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'$ et $\mathrm{E}_c(.)$ désignant l'espérance conditionnelle sachant $\mathbf{y}, \boldsymbol{\varphi}^{[t,0]}$, le membre de gauche qui est symétrique s'exprime par (exemple de $K = 2$) :

| | 11 | 12 | 21 | 22 |
|---|---|---|---|---|
| 11 | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{T}_{11}\mathbf{u}_1)$ | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{T}_{11}\mathbf{u}_2)$ | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{T}_{12}\mathbf{u}_1)$ | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{T}_{12}\mathbf{u}_2)$ |
| 12 | | $\mathrm{E}_c(\mathbf{u}_2'\mathbf{T}_{11}\mathbf{u}_2)$ | $\mathrm{E}_c(\mathbf{u}_2'\mathbf{T}_{12}\mathbf{u}_1)$ | $\mathrm{E}_c(\mathbf{u}_2'\mathbf{T}_{12}\mathbf{u}_2)$ |
| 21 | | | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{T}_{22}\mathbf{u}_1)$ | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{T}_{22}\mathbf{u}_2)$ |
| 22 | | | | $\mathrm{E}_c(\mathbf{u}_2'\mathbf{T}_{22}\mathbf{u}_2)$ |

et celui de droite :

| | 11 | 12 | 21 | 22 |
|---|---|---|---|---|
| | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{v}_1)$ | $\mathrm{E}_c(\mathbf{u}_2'\mathbf{v}_1)$ | $\mathrm{E}_c(\mathbf{u}_1'\mathbf{v}_2)$ | $\mathrm{E}_c(\mathbf{u}_2'\mathbf{v}_2)$ |

Les calculs correspondants peuvent être effectués en utilisant les équations du modèle mixte décrites précédemment en (116), c'est-à-dire (en ignorant les indices supérieurs)

$$f_{kl,mn} = \mathrm{tr}\left[\mathbf{Z}_k'\mathbf{Z}_m(\widehat{\mathbf{u}}_n\widehat{\mathbf{u}}_l' + \sigma_0^2\mathbf{C}_{u_nu_l})\right], \tag{121a}$$

$$h_{kl} = \widehat{\mathbf{u}}_l'\mathbf{Z}_k'\mathbf{y} - \mathrm{tr}\left[\mathbf{Z}_k'\mathbf{X}(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{u}}_l' + \sigma_0^2\mathbf{C}_{\beta u_l})\right], \tag{121b}$$

où $\mathbf{Z}_k'\mathbf{Z}_m$ est le bloc relatif aux effets $\mathbf{u}_k$ et $\mathbf{u}_m$ dans la matrice des coefficients ; $\mathbf{Z}_k'\mathbf{X}$ est le bloc correspondant à $\mathbf{u}_k$ et $\boldsymbol{\beta}$ ; $\mathbf{C}_{u_ku_m}$ et $\mathbf{C}_{u_k\beta} = \mathbf{C}_{\beta u_k}'$ sont les blocs homologues dans l'inverse de la matrice des coefficients ; $\mathbf{Z}_k'\mathbf{y}$ est le sous vecteur du second membre relatif à $\mathbf{u}_k$ ; $\widehat{\boldsymbol{\beta}}$ et $\widehat{\mathbf{u}}_k$ sont les solutions de $\boldsymbol{\beta}$ et $\mathbf{u}_k$.

La matrice des coefficients $\boldsymbol{\alpha}^{[t+1]}$ étant obtenue, on revient à $\mathbf{G}_0$ par

$$\mathbf{G}_0^{[t+1]} = \boldsymbol{\alpha}^{[t+1]}\mathbf{G}_{0^*}^{[t+1]}(\boldsymbol{\alpha}^{[t+1]})'. \tag{122}$$

Enfin, quant à $\sigma_0^2$, la maximisation de $\mathrm{Q}_0(\boldsymbol{\alpha}, \sigma_0^2 \mid \boldsymbol{\varphi}^{[t,0]})$ conduit à :

$$\sigma_0^{2[t+1]} = \mathrm{E}(\mathbf{e}'\mathbf{e} \mid \mathbf{y}, \boldsymbol{\varphi}^{[t,0]})/N, \tag{123}$$

la résiduelle $\mathbf{e}$ étant ajustée en fonction de $\boldsymbol{\alpha}^{[t+1]}$ *via* $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\alpha}^{[t+1]}\widetilde{\mathbf{u}}_i$. Un procédé rapide consiste en une maximisation conditionnelle basée sur $\boldsymbol{\alpha} = \mathbf{I}_K$ ce qui redonne la formule classique de l'EM0.

Quoiqu'il en soit, le nombre d'itérations nécessaires à la convergence à une précision donnée s'avère considérablement réduit par rapport à la version standard EM0 de l'algorithme.

Le nombre d'itérations est réduit d'un facteur de l'ordre de 3 à 4 comme le montre la figure 2 relative à la variance de l'intercept dans l'analyse de données de croissance (Foulley et van Dyk, 2000).

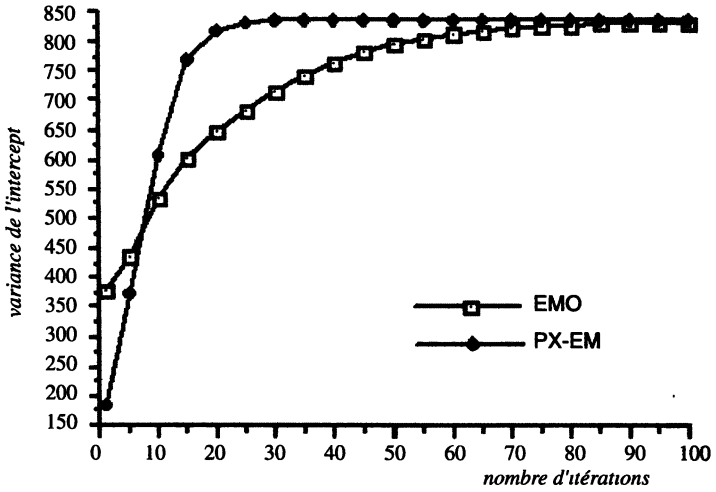


FIG 2. – Séquences typiques d'itérations EMO et PX-EM.

Par ailleurs, on a pu observer que cette version PX permet d'obtenir des estimations REML d'une matrice de variance y compris en bordure de l'espace paramétrique alors que les autres algorithmes ne convergent pas (Delmas *et al.*, 2002).

On peut également combiner la modélisation à plusieurs facteurs aléatoires corrélés et celle de variances hétérogènes; un exemple en est fourni par les modèles à coefficients aléatoires hétéroscédastiques (Robert-Granié *et al.*, 2002). D'un point de vue algorithmique, l'algorithme EM permet très bien de réaliser cette synthèse sur la base des techniques présentées précédemment (Foulley et Quaas, 1995).

## Conclusion

L'algorithme EM trouve dans le calcul des estimateurs du maximum de vraisemblance des composantes de la variance du modèle linéaire mixte un terrain d'application privilégié. Il permet d'obtenir aussi bien des estimations ML que REML avec dans les deux cas des expressions très simples. Un des avantages de l'algorithme – et non des moindres – est qu'il assure le maintien des valeurs dans l'espace des paramètres. Sa flexibilité est telle qu'on l'adapte facilement à des situations plus complexes telles que celles par exemple de variances hétérogènes décrites par des modèles loglinéaires structuraux. On peut également améliorer très significativement ses performances par standardisation des effets aléatoires, et plus généralement, grâce à la technique d'extension paramétrique qui apparaît très prometteuse y compris dans ses prolongements stochastiques. À cet égard, dans le cadre d'un modèle très proche de (110),

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i; \mathbf{u}_i \sim \mathcal{N}(\boldsymbol{\xi}, \mathbf{G}_0),$$

van Dyk et Meng, (2002) proposent cette fois une transformation affine des effets aléatoires $\tilde{\mathbf{u}}_i = \boldsymbol{\alpha}^{-1}\mathbf{u}_i + \boldsymbol{\eta}$ qu'ils introduisent dans un algorithme d'augmentation de données en considérant des *a priori* gaussiens sur $\text{vec}(\alpha)$ et $\boldsymbol{\eta}$. Cet algorithme comparé à la procédure standard sur quelques exemples s'avère très performant pourvu que la matrice de transformation $\boldsymbol{\alpha}$ soit complète et non pas triangulaire comme cela avait été déjà remarqué par van Dyk (2000) et Foulley et van Dyk (2000).

Enfin, il faut être pleinement conscient que le champ d'application de l'algorithme est beaucoup plus vaste que celui abordé ici. Maintes modélisations font appel à des structures cachées qui peuvent donner lieu à une inférence ML *via* l'algorithme EM. Un domaine particulièrement propice à cette approche réside dans les modèles de Markov cachés. Ceux-ci sont par exemple utilisés dans l'analyse des séquences biologiques comme celles de l'ADN. Dans ces modèles, la succession des états cachés représente l'hétérogénéité de la séquence. Les paramètres sont trop nombreux et le calcul de la vraisemblance trop complexe pour faire l'objet d'une maximisation directe. Diverses approches sont possibles pour contourner ces difficultés, mais l'algorithme EM s'avère encore la méthode à la fois la plus simple à mettre en œuvre et la plus efficace (Nicolas *et al.*, 2002).

## Remerciements

Je tiens à remercier tous ceux avec qui j'ai travaillé sur ce sujet (Daniel Gianola, Sotan Im, Florence Jaffrézic, Richard Quaas, Christèle Robert, Magali San Cristobal, Caroline Thaon d'Arnoldi, Florence Jaffrézic et David van Dyk) et qui m'ont permis, à l'occasion de nombreux échanges, de mieux comprendre les subtilités multiples de l'algorithme EM.

Je suis particulièrement reconnaissant à Gilles Celeux (INRIA, Grenoble) pour avoir relu attentivement une première version du manuscrit et y avoir apporté ses commentaires et ses suggestions d'expert.

Ma gratitude va également à mes collègues Bernard Bonaïti, Jean-Jacques Colleau, aux lecteurs et au rédacteur de la revue, pour leur lecture critique du manuscrit ainsi qu'à François Rodolphe (INRA-MIG, Jouy-en-Josas) pour m'avoir indiqué tout le profit qu'il pouvait tirer de l'algorithme EM dans l'analyse des séquences.



## RÉFÉRENCES

ANDERSON T.W. (1984), *An introduction to multivariate analysis*, J Wiley and Sons, New York.

ANDERSON D.A., AITKIN M. (1985), Variance components models with binary response : interviewer probability, *Journal of the Royal Statistical Society* B, 47,203-210.

AITKIN M. (1987), Modelling variance heterogeneity in normal regression using GLIM, *Applied Statistics*, 36, 332-339.

BOOTH J.G., HOBERT J.P. (1999), Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society* B, 61,265-285.

BOYLES R.A. (1983), On the convergence of the EM algorithm, *Journal of the Royal Statistical Society* B, 45,47-50.

CELEUX G., DIEBOLT J. (1985), The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quaterly*, 2, 73-82.

CELEUX G., DIEBOLT J. (1992), A Stochastic Approximation Type EM Algorithm for the Mixture Problem, *Stochastics and Stochastics Reports*, 41, 119-134.

CELEUX G., GOVAERT G. (1992), A classification algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14, 315-322.

CELEUX G., CHAUVEAU D., DIEBOLT J. (1996), Some stochastic versions of the EM algorithm. *Journal of Statistical Computation and Simulation*, 55, 287-314.

DELMAS C., FOULLEY J.L., ROBERT-GRANIÉ C. (2002), Further insights into tests of variance components and model selection, Proceedings of the 7$^{th}$ World Congress of Genetics applied to Livestock Production, Montpellier, France, 19-23 August 2002.

DELYON B., LAVIELLE M., MOULINES E. (1999), Convergence of a stochastic approximation version of the EM algorithm, *Annals of Statistics*, 27, 94-128.

DEMPSTER A., LAIRD N., RUBIN R. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* B, 39,1-38.

EFRON B. (1977), Discussion on maximum likelihood from incomplete data via the EM algorithm (by Dempster A., Laird N., Rubin R.), *Journal of the Royal Statistical Society* B, 39,1-38.

FISHER R.A. (1925), Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.

FOULLEY J.L. (1993), A simple argument showing how to derive restricted maximum likelihood, *Journal of Dairy Science*, 76, 2320-2324.

FOULLEY J.L. (1997), ECM approaches to heteroskedastic mixed models with constant variance ratios. *Genetics Selection Evolution*, 29, 297-318.

FOULLEY J.L., IM S., GIANOLA D., HOESCHELE I. (1987), Empirical Bayes estimation of parameters for n polygenic binary traits, *Genetics Selection Evolution*, 19, 127-224.

FOULLEY J.L., SAN CRISTOBAL M., GIANOLA D., IM S. (1992), Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models, *Computational Statistics and Data Analysis*, 13, 291-305.

FOULLEY J.L., QUAAS R.L. (1995), Heterogeneous variances in Gaussian linear mixed models, *Genetics Selection Evolution*, 27, 211-228.

FOULLEY J.L., QUAAS R.L., THAON d'ARNOLDI C. (1998), A link function approach to heterogeneous variance components, *Genetics Selection Evolution*, 30, 27-43.

FOULLEY J.L., van DYK D.A. (2000), The PX EM algorithm for fast fitting of Henderson's mixed model, *Genetics Selection Evolution*, 32, 143-163.

FOULLEY J.L., JAFFREZIC F., ROBERT-GRANIÉ C. (2000), EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis, *Genetics Selection Evolution*, 32, 129-141.

FOULLEY J.L., DELMAS C., ROBERT-GRANIÉ C. (2002), Méthodes du maximum de vraisemblance en modèle linéaire mixte, *Journal de la Société Française de Statistique*, 143, 1-2, 5-52.

GRIMAUD A., HUET S., MONOD H., JENCZEWSKI E., EBER F. (2002), Mélange de modèles mixtes : application à l'analyse des appariements de chromosomes chez des haploïdes de colza, *Journal de la Société Française de Statistique*, 143, 1-2, 147-153.

HARTLEY H.O., RAO J.N.K. (1967), Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika*, 54, 93-108.

HARVILLE D.A. (1974), Bayesian inference for variance components using only error contrasts, *Biometrika*, 61, 383-385.

HARVILLE D.A. (1977), Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, 72, 320-340.

HENDERSON C.R. (1973), Sire evaluation and genetic trends, In : *Proceedings of the animal breeding and genetics symposium in honor of Dr J Lush.* American Society Animal Science-American Dairy Science Association, 10-41, Champaign, IL.

HENDERSON C.R. (1984), *Applications of linear models in animal breeding*, University of Guelph, Guelph, 1984.

KUHN E., LAVIELLE M. (2002), Coupling a stochastic approximation version of EM with a MCMC procedure, Rapport technique, Université Paris Sud, 15pages.

LAIRD N.M. (1982), The computation of estimates of variance components using the EM algorithm, *Journal of Statistical Computation and Simulation*, 14, 295-303.

LAIRD N.M., WARE J.H. (1982), Random effects models for longitudinal data, *Biometrics*, 38 963-974.

LAIRD N.M., LANGE N., STRAM D. (1987), Maximum likelihood computations with repeated measures : Application of the EM algorithm. *Journal of the American Statistical Association*, 82, 97-105.

LANGE K. (1995), A gradient algorithm locally equivalent to the EM algorithm, *Journal of the Royal Statistical Society* B, 57, 425-437.

LEONARD T. (1975), A Bayesian approach to the linear model with unequal variances, *Technometrics*, 17, 95-102.

LEONARD T., HSU JSJ. (1999), *Bayesian methods, an analysis for statisticians and interdisciplinary researchers*, Cambridge University Press, Cambridge, UK.

LIANG K.Y., ZEGER S.L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika*, 73, 13-22.

LIAO J.G., LIPSITZ S.R. (2002) A type of restricted maximum likelihood estimator of variance components in generalized linear mixed models, *Biometrika*, 89, 401-409.

LINDLEY D.V., SMITH A.F.M. (1972), Bayes Estimates for the Linear Model, *Journal of the Royal Statistical Society* B, 34, 1-41.

LINDSTRÖM M.J., BATES D.M. (1988), Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data, *Journal of the American Statistical Association*, 83, 1014-1022.

LIU C., RUBIN D.B. (1994), The ECME algorithm : a simple extension of the EM and ECM with faster monotone convergence, *Biometrika*, 81, 633-648.

LIU C., RUBIN D.B., WU Y.N. (1998), Parameter expansion to accelerate EM : the PX-EM algorithm, *Biometrika*, 85, 755-770.

LIU J.S., WU Y.N. (1999), Parameter expansion scheme for data augmentation, *Journal of the American Statistical Association*, 94, 1264-1274.

LOUIS T.A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society* B, 44, 226-233.

MCLACHLAN G.J., BASHFORD K.E. (1988) *Mixture models : inferences and applications to clustering*, Marcel Dekker, New York.

MCLACHLAN G.J., KRISHNAN T. (1997), *The EM algorithm and extensions*, John Wiley & Sons, New York.

MCLACHLAN G.J., PEEL D. (2000), *Finite mixture models*, John Wiley & Sons, New York.

MENG X.L. (2000) Missing data : dial M for ? ? ?, *Journal of the American Statistical Association*, 95, 1325-1330.

MENG X.L., RUBIN D.B. (1991), Using EM to obtain asymptotic variance-covariance matrices : the SEM algorithm, *Journal of the American Statistical Association*, 86, 899-909.

MENG X.L., RUBIN D.B. (1993), Maximum likelihood estimation via the ECM algorithm : a general framework, *Biometrika*, 80, 267-278.

MENG X.L., van DYK D.A. (1997), The EM algorithm-an Old Folk-song Sung to a Fast New Tune, *Journal of the Royal Statistical Society* B 59, 511-567.

MENG X.L., van DYK D.A. (1998), Fast EM-type implementations for mixed effects models, *Journal of the Royal Statistical Society* B 60, 559-578.

NAIR V.N., PREGIBON D. (1988), Analyzing dispersion effects from replicated factorial experiments, *Technometrics*, 30, 247-257.

NICOLAS P., BIZE L., MURI F., HOEBEKE M., RODOLPHE F., EHRLICH S., PRUM B., BESSIÈRES P. (2002), Mining bacillus subtilis chromosome heterogeneities using hidden Markov models, *Nucleic Acid Research*, 30, 1418-1426.

NIELSEN S.F. (2000), The stochastic EM algorithm : estimation and asymptotic results, *Bernoulli*, 6, 457-489.

PATTERSON H.D., THOMPSON R. (1971), Recovery of inter-block information when block sizes are unequal, *Biometrika*, 58, 545-554.

RAO C.R. (1973), *Linear Statistical Inference and its Applications*, 2nd edition. Wiley, New-York.

RAO C.R., KLEFFE J. (1988), *Estimation of variance components and applications*, North Holland series in statistics and probability, Elsevier, Amsterdam.

ROBERT-GRANIÉ C., DUCROCQ V., FOULLEY J.L. (1997), Heterogeneity of variance for type traits in the Montbéliarde cattle. *Genetics Selection Evolution*, 29, 545-570.

ROBERT-GRANIÉ C., BONAITI B., BOICHARD D., BARBAT A. (1999), Accounting for variance heterogeneity in French dairy cattle genetic evaluation, *Livestock Production Science*, 60, 343-357.

ROBERT-GRANIÉ C., HEUDE B., FOULLEY J.L. (2002), Modelling the growth curve of Maine Anjou beef cattle using heteroskedastic random regression models. *Genetics Selection Evolution*, 34, 423-445.

ROBERT C.P. (1996), Mixtures of distributions : inference and estimation, In *Markov Chain Monte Carlo in Practice* (Gilks W.R., Ricardson S., Spiegelhalter D.J., editors), Chapman & Hall, London, 441-464.

ROBERT C.P., CASELLA G. (1999), *Monte Carlo Statistical Methods*, Springer, Berlin.

SAN CRISTOBAL M., ROBERT-GRANIÉ C., FOULLEY J.L. (2002), Hétéroscédasticité et modèles linéaires mixtes : théorie et applications en génétique quantitative, *Journal de la Société Française de Statistiques*, 143, 155-165.

SEARLE S.R. (1992), *Matrix algebra useful for statistics*, J. Wiley and Sons, New-York.

SEARLE S.R., CASELLA G., MC CULLOCH C.E. (1992), *Variance components*, J. Wiley and Sons, New-York.

SMITH S.P., GRASER H.U. (1986), Estimating variance components in a class of mixed models by restricted maximum likelihood, *Journal of Dairy Science*, 69, 1156-1165.

TANNER M.A. (1996), *Tools for Statistical Inference : Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer, New York.

TANNER M.A., WONG W.H. (1987), The calculation of posterior distributions by Data Augmentation (with discussion), *Journal of the American Statistical Association*, 82, 528-550.

TITTERINGTON D.M. (1984), Recursive parameter estimation using incomplete data, *Journal of the Royal Statistical Society* B, 46, 257-267.

TITTERINGTON D.M., SMITH A.F.M., MAKOV U.E. (1985), *Statistical Analysis of Finite Mixture*, John Wiley & Sons, New York.

THOMPSON R. (2002), A review of genetic parameter estimation, Proceedings of the $7^{th}$ World Congress of Genetics applied to Livestock Production, Montpellier, France, 19-23 August 2002.

van DYK D.A. (2000), Fitting mixed-effects models using efficient EM-type algorithms, *Journal of Computational and Graphical Statistics*, 9, 78-98.

van DYK D.A, MENG X.L. (2001), The art of data augmentation, *Journal of Computational and Graphical Statistics* 10, 1-50.

WEI G.C.G., TANNER M.A.(1990), A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, 85, 699-704.

WEIR B.S. (1996), *Genetic data analysis II*, Sinauer associates, Sunderland, Massachussets.

WOLFINGER R.D., TOBIAS R.D. (1998), Joint estimation of location, dispersion, and random effects in robust design, *Technometrics*, 40, 62-71.

WU C.F.J. (1983), On the convergence properties of the EM algorithm. *Annals of Statistics*, 11, 95-103.

WU R., MA C-X., LITTLE R.C., CASELLA G. (2002), A statistical model for the genetic origin of allometric scaling laws in biology, *Journal of Theoretical Biology*, 219, 121-135.

# ANNEXE A

## Score et hessien : résultats de base

### 1. Dérivée première

Par définition de la dérivée logarithmique, il vient

$$\frac{\partial \ln g(\mathbf{y} \mid \boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} = \frac{\partial g(\mathbf{y} \mid \boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} \frac{1}{g(\mathbf{y} \mid \boldsymbol{\Phi})}. \qquad \textbf{(A.1)}$$

Or la densité marginale correspond à

$$g(\mathbf{y} \mid \boldsymbol{\phi}) = \int f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi}) \mathrm{d}\mathbf{z},$$

d'où sa dérivée

$$\frac{\partial g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \int \frac{\partial f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \, \mathrm{d}\mathbf{z} \tag{A.2}$$

Le terme sous le signe somme peut de nouveau être développé comme une dérivée logarithmique en

$$\frac{\partial f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \, f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi}), \tag{A.3}$$

en explicitant aussi la densité conjointe en fonction des densités marginale et conditionnelle,

$$f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi}) = g(\mathbf{y} \mid \boldsymbol{\phi}) h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}). \tag{A.4}$$

En reportant l'expression de (A.4) dans (A.3) puis celle-ci dans (A.2) et (A.1), il vient :

$$\frac{\partial \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{1}{g(\mathbf{y} \mid \boldsymbol{\phi})} \int \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \, g(\mathbf{y} \mid \boldsymbol{\phi}) h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\mathbf{z},$$

soit après simplification,

$$\frac{\partial \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \int \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\mathbf{z}, \tag{A.5}$$

ou encore,

$$\boxed{\frac{\partial \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \mathrm{E}_c \left[ \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \right]}, \tag{A.6}$$

l'espérance notée $\mathrm{E}_c(\cdot)$ étant prise par rapport à la densité de $\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}$.

### 2. Dérivée seconde

Dérivons à nouveau l'expression précédente (A.5), il vient :

$$\begin{aligned} \frac{\partial^2 \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi} \, \partial \boldsymbol{\phi}'} = & \int \frac{\partial^2 \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi} \, \partial \boldsymbol{\phi}'} \, h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\mathbf{z} \\ & + \int \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \, \frac{\partial \ln h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} \, h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\mathbf{z} \end{aligned} \tag{A.7}$$

Or, par définition de $h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi})$,

$$\frac{\partial \ln h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} - \frac{\partial \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}}.$$

En reportant dans (A.7), on obtient

$$\frac{\partial^2 \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} = \mathrm{E}_C \left[ \frac{\partial^2 \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}\, \partial \boldsymbol{\phi}'} \right] + \mathrm{E}_C \left[ \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \right] \left[ \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} \right] - \frac{\partial \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \int \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} \, h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\mathbf{z}$$

et, eu égard à (A.5 et 6), on en déduit que :

$$\boxed{\frac{\partial^2 \ln g(\mathbf{y} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} = \mathrm{E}_C \left[ \frac{\partial^2 \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}\, \partial \boldsymbol{\phi}'} \right] + \mathrm{Var}_C \left[ \frac{\partial \ln f(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \right]}. \quad \text{(A.8)}$$

# ANNEXE B

## Eléments de l'expression de la variance résiduelle en EM

**1. Démonstration de** $\boxed{(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}}) = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y} - \lambda \widehat{\mathbf{u}}'\widehat{\mathbf{u}}}$, (B.1)

Partons des équations du modèle mixte sous leur forme condensée

$$(\mathbf{T}'\mathbf{T} + \boldsymbol{\Lambda})\widehat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{y}, \quad \text{(B.2)}$$

où $\mathbf{T}(\mathbf{X}, \mathbf{Z})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}')'$ et $\boldsymbol{\Lambda} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda \mathbf{I}_q \end{pmatrix}$

En multipliant le système (B.2) à gauche par $\widehat{\boldsymbol{\theta}}'$, il vient :

$$\widehat{\boldsymbol{\theta}}'(\mathbf{T}'\mathbf{T} + \boldsymbol{\Lambda})\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y}$$

En introduisant cette égalité dans $(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}}) = \mathbf{y}'\mathbf{y} - 2\widehat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y} + \widehat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{T}\widehat{\boldsymbol{\theta}}$, on obtient :

$$(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{T}\widehat{\boldsymbol{\theta}}) = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y} - \widehat{\boldsymbol{\theta}}'\boldsymbol{\Lambda}\widehat{\boldsymbol{\theta}}$$

et cela, adjoint au fait que $\widehat{\boldsymbol{\theta}}'\boldsymbol{\Lambda}\widehat{\boldsymbol{\theta}} = \lambda \widehat{\mathbf{u}}'\widehat{\mathbf{u}}$, établit la démonstration de (B.1).

**2. Démonstration de** $\boxed{\mathrm{tr}(\mathbf{C}\mathbf{T}'\mathbf{T}) = \mathrm{rang}(\mathbf{X}) + q - \lambda \ \mathrm{tr}(\mathbf{C}_{uu})}$ (B.3)

La matrice $\mathbf{C}$ vérifie par définition la relation suivante :

$$\mathbf{C}(\mathbf{T}'\mathbf{T} + \boldsymbol{\Lambda}) = \mathbf{I}_{p+q} \quad \text{(B.4)}$$

On suppose pour simplifier l'écriture que $\mathbf{X}_{(N \times p)}$ est de plein rang.

Dans ces conditions,

$$\mathbf{C}\mathbf{T}'\mathbf{T} = \mathbf{I}_{p+q} - C\boldsymbol{\Lambda}$$

et, posant $\mathbf{C} = \begin{bmatrix} \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \\ \mathbf{C}_{u\beta} & \mathbf{C}_{uu} \end{bmatrix}$,

$\operatorname{tr}(\mathbf{CT}'\mathbf{T}) = p + q - \lambda \operatorname{tr}(\mathbf{C}_{uu})$, QED

# ANNEXE C

## Variances hétérogènes : dérivées intervenant à la phase M

La fonction $Q$ à maximiser présente la forme suivante :

$$Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = -1/2 \left[ N \ln 2\pi + \sum_{i=1}^{I} n_i \ln \sigma_{0,i}^2 + \sum_{i=1}^{I} \mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i)/\sigma_{0,i}^2 \right], \quad \text{(C.1)}$$

avec

$$\ln \sigma_{1,i}^2 = \mathbf{p}_i'\boldsymbol{\delta}, \quad \text{(C.2)}$$

$$\ln \tau_i = \mathbf{h}_i'\boldsymbol{\lambda}, \quad \text{(C.3)}$$

et,

$$\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \tau_i \sigma_{0,i} \mathbf{Z}_i \mathbf{u}^*. \quad \text{(C.4)}$$

### 1. Dérivée première par rapport à $\delta$

L'application des dérivées de fonctions en chaîne conduit à :

$$\frac{\partial Q}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{I} \frac{\partial Q}{\partial \ln \sigma_{0,i}^2} \frac{\partial \ln \sigma_{0,i}^2}{\partial \boldsymbol{\delta}}$$

Or

$$\frac{\partial Q}{\partial \ln \sigma_{0,i}^2} = \sigma_{0,i}^2 \frac{\partial Q}{\partial \sigma_{0,i}^2}$$

$$\frac{\partial \ln \sigma_{0,i}^2}{\partial \boldsymbol{\delta}} = \mathbf{p}_i$$

soit

$$\frac{\partial Q}{\partial \sigma_{0,i}^2} = -\frac{1}{2} \left[ \frac{n_i}{\sigma_{0,i}^2} - \frac{\mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i)}{\sigma_{0,i}^4} + \frac{1}{\sigma_{0,i}^2} \frac{\partial \mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i)}{\partial \sigma_{0,i}^2} \right],$$

$$\frac{\partial \mathrm{E}_c(\mathbf{e}'\mathbf{e}_i)}{\partial \sigma_{0,i}^2} = \frac{\partial \sigma_{0,i}}{\partial \sigma_{0,i}^2} \frac{\partial \mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i)}{\partial \sigma_{0,i}} = \frac{1}{2\sigma_{0,i}} 2\mathrm{E}_c \left[ \left( \frac{\partial \mathbf{e}_i'}{\partial \sigma_{0,i}} \right) \mathbf{e}_i \right]$$

et,

$$\frac{\partial \mathbf{e}_i}{\partial \sigma_{0,i}} = \tau_i \mathbf{Z}_i \mathbf{u}^*$$

D'où

$$\frac{\partial Q}{\partial \sigma_{0,i}^2} = -\frac{1}{2} \left[ \frac{n_i}{\sigma_{0,i}^2} - \frac{\mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i)}{\sigma_{0,i}^4} - \tau_i \frac{\mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{e}_i)}{\sigma_{0,i}^3} \right].$$

Soit $v_{\delta,i} = \dfrac{\partial Q}{\partial \ln \sigma^2_{0,i}}$ tel que $\dfrac{\partial Q}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{I} v_{\delta,i}\mathbf{p}_i = \mathbf{P}'\mathbf{v}_\delta$, le terme $v_{\delta,i}$ s'exprime par :

$$\boxed{v_{\delta,i} = \frac{1}{2}\left[\frac{\mathrm{E}_c(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i}{\sigma^2_{0,i}} - n_i\right]}. \tag{C.5}$$

### 2. Dérivée première par rapport à $\boldsymbol{\lambda}$

En suivant la même démarche que précédemment, on **a** :

$$\frac{\partial Q}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{I} \frac{1}{\sigma^2_{0,i}} \frac{\partial \mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i)}{\partial \tau_i} \frac{\partial \tau_i}{\partial \ln \tau_i} \frac{\partial \ln \tau_i}{\partial \boldsymbol{\lambda}},$$

avec

$$\frac{\partial \mathrm{E}_c(\mathbf{e}_i'\mathbf{e}_i)}{\partial \tau_i} = 2\mathrm{E}_c\left[\left(\frac{\partial \mathbf{e}_i'}{\partial \tau_i}\right)\mathbf{e}_i\right],$$

$$\frac{\partial \mathbf{e}_i}{\partial \tau_i} = -\sigma_{0,i}\mathbf{Z}_i\mathbf{u}^*$$

$$\frac{\partial \tau_i}{\partial \ln \tau_i} = \tau_i,$$

et

$$\frac{\partial \ln \tau_i}{\partial \boldsymbol{\lambda}} = \mathbf{h}_i$$

D'où

$$\frac{\partial Q}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{I} v_{\lambda,i}\mathbf{h}_i = \mathbf{H}'\mathbf{v}_\lambda,$$

avec

$$\boxed{v_{\lambda,i} = \frac{\tau_i}{\sigma_{0,i}}\,\mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{e}_i)}. \tag{C.6}$$

### 3. Dérivée seconde par rapport à $\delta$

Posons

$$-\frac{\partial^2 Q}{\partial \boldsymbol{\delta}\,\partial \boldsymbol{\delta}'} = \sum_{i=1}^{I} w_{\delta\delta,ii}\mathbf{p}_i\mathbf{p}_i' = \mathbf{P}'\mathbf{W}_{\delta\delta}\mathbf{P},$$

où

$$w_{\delta\delta,ii} = -\frac{\partial v_{\delta,i}}{\partial \ln \sigma^2_{0,i}} = -\sigma^2_{0,i}\frac{\partial v_{\delta,i}}{\partial \sigma^2_{0,i}}.$$

Or

$$\frac{\partial v_{\delta,i}}{\partial \sigma^2_{0,i}} = -\frac{1}{2\sigma^4_{0,i}}\,\mathrm{E}_c\left[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i\right] + \frac{1}{2\sigma^2_{0,i}}\frac{\partial \mathrm{E}_c\left[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i\right]}{\partial \sigma^2_{0,i}},$$

$$\frac{\partial\left[\mathrm{E}_c(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i\right]}{\partial\sigma_{0,i}^2}=\frac{1}{2\sigma_{0,i}}\,\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\frac{\partial\mathbf{e}_i}{\partial\sigma_{0,i}}\right]=-\frac{\tau_i}{2\sigma_{0,i}}\,\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*\right]$$

et

$$\frac{\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i\right]}{\sigma_{0,i}^2}=\frac{1}{\sigma_{0,i}^2}\,\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right]-\frac{\tau_i}{\sigma_{0,i}}\,\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*\right],$$

d'où

$$\boxed{w_{\delta\delta,ii}=\frac{1}{2\sigma_{0,i}^2}\left\{\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right]-\frac{\tau_i\sigma_{0,i}}{2}\,\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*\right]\right\}}. \tag{C.7}$$

**4. Dérivée seconde par rapport à $\boldsymbol{\lambda}$**

De la même façon,

$$-\frac{\partial^2 Q}{\partial\boldsymbol{\lambda}\,\partial\boldsymbol{\lambda}'}=\sum_{i=1}^{I}w_{\lambda\lambda,ii}\mathbf{h}_i\mathbf{h}_i'=\mathbf{H}'\mathbf{W}_{\lambda\lambda}\mathbf{H}$$

où

$$w_{\lambda\lambda,ii}=-\frac{\partial v_{\lambda,i}}{\partial\ln\tau_i}=-\tau_i\,\frac{\partial v_{\lambda,i}}{\partial\tau_i},$$

$$\begin{aligned}\frac{\partial v_{\lambda,i}}{\partial\tau_i}&=\frac{\mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{e}_i)}{\sigma_{0,i}}+\frac{\tau_i}{\sigma_{0,i}}\,\mathrm{E}_c\left[\mathbf{u}^{*\prime}\mathbf{Z}_i'\left(\frac{\partial\mathbf{e}_i}{\partial\tau_i}\right)\right]\\&=\frac{1}{\sigma_{0,i}}\left\{\mathrm{E}_c\left[\mathbf{u}^{*\prime}\mathbf{Z}_i'(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})-\tau_i\sigma_{0,i}\mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i\mathbf{u}^*)\right]\right\}-\tau_i\mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i\mathbf{u}^*)\end{aligned}$$

d'où

$$\boxed{w_{\lambda\lambda,ii}=\tau_i\left\{2\tau_i\mathrm{E}_c(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i\mathbf{u}^*)-\frac{1}{\sigma_{0,i}}\,\mathrm{E}_c\left[\mathbf{u}^{*\prime}\mathbf{Z}_i'(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right]\right\}}. \tag{C.8}$$

**5. Dérivée seconde croisée $\boldsymbol{\delta}-\boldsymbol{\lambda}$**

Soit $-\dfrac{\partial^2 Q}{\partial\boldsymbol{\delta}\,\partial\boldsymbol{\lambda}'}=\displaystyle\sum_{i=1}^{I}w_{\delta\lambda,ii}\mathbf{p}_i\mathbf{h}_i'=\mathbf{P}'\mathbf{W}_{\delta\lambda}\mathbf{H},$

où

$$w_{\delta\lambda,ii}=-\frac{\partial v_{\delta,i}}{\partial\ln\tau_i}=-\tau_i\,\frac{\partial v_{\delta,i}}{\partial\tau_i},$$

$$\frac{\partial v_{\delta,i}}{\partial\tau_i}=\frac{1}{2\sigma_{0,i}^2}\,\mathrm{E}_c\left[(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\left(\frac{\partial\mathbf{e}_i}{\partial\tau_i}\right)\right].$$

Comme $\dfrac{\partial \mathbf{e}_i}{\partial \tau_i} = -\sigma_{0,i}\mathbf{Z}_i\mathbf{u}^*$,

$$w_{\delta\lambda,ii} = -\tau_i \times \frac{1}{2\sigma_{0,i}^2} \times -\sigma_0 \mathrm{E}_c\big[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*\big],$$

c'est-à-dire

$$\boxed{w_{\delta\lambda,ii} = \frac{\tau_i}{2\sigma_{0,i}} \mathrm{E}_c\big[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*\big]}. \qquad \text{(C.9)}$$

On vérifie aisément la propriété de symétrie des dérivées, soit

$$-\frac{\partial^2 Q}{\partial\boldsymbol{\lambda}\,\partial\boldsymbol{\delta}'} = \sum_{i=1}^{I} w_{\delta\lambda,ii}\mathbf{h}_i\mathbf{p}_i' = \mathbf{H}'\mathbf{W}_{\lambda\delta}\mathbf{P} = (\mathbf{P}'\mathbf{W}_{\delta\lambda}\mathbf{H})' \text{ avec } \mathbf{W}_{\delta\lambda} = \mathbf{W}_{\lambda\delta}$$

### 6. Espérances des dérivées secondes

Soit à expliciter : $\widetilde{w}_{\delta\delta,ii} = \mathrm{E}(w_{\delta\delta,ii})$, $\widetilde{w}_{\alpha\delta,ii} = \mathrm{E}(w_{\alpha\delta,ii})$ et $\widetilde{w}_{\alpha\alpha,ii} = \mathrm{E}(w_{\alpha\alpha,ii})$.

Par définition

$$\mathrm{E}_y\left\{\mathrm{E}\big[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i\big] \,\big|\, \mathbf{y}, \boldsymbol{\phi}\right\} = \mathrm{E}\big[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i\big].$$

Comme $\mathbf{u}^*$ et $\mathbf{e}_i$ ne sont pas corrélés,

$$\mathrm{E}\big[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{e}_i\big] = \mathrm{E}(\mathbf{e}_i'\mathbf{e}_i) = n_i\sigma_{0,i}^2.$$

De même,

$$\mathrm{E}\big[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{Z}_i\mathbf{u}^*\big] = \tau_i\sigma_{0,i}\mathrm{E}(\mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i u^*) = \tau_i\sigma_{0,i}\ \mathrm{tr}(\mathbf{Z}_i'\mathbf{Z}_i\mathbf{A}).$$

Dans ces conditions, $\widetilde{w}_{\delta\delta,ii}$ se réduit à

$$\widetilde{w}_{\delta\delta,ii} = 1/2\big[n_i + \tau_i^2\mathrm{tr}(\mathbf{Z}_i'\mathbf{Z}_i\mathbf{A})/2\big].$$

De même : $\widetilde{w}_{\delta\lambda,ii} = 1/2\tau_i^2\mathrm{tr}(\mathbf{Z}_i'\mathbf{Z}_i\mathbf{A})$ et $\widetilde{w}_{\lambda\lambda,ii} = \tau_i^2\mathrm{tr}(\mathbf{Z}_i'\mathbf{Z}_i\mathbf{A})$.