LE MODELE LINEAIRE MIXTE

Jean-Louis Foulley
INRA- Station de Génétique quantitative & appliquée
CR de Jouy, 78352 Jouy-en-Josas Cedex
(foulley@jouy.inra.fr)

2003

Je tiens à remercier les responsables et les étudiants de la section “Biostatistiques” de l’ENSAI de Rennes et ceux du DEA de Génétique multifactorielle pour m’avoir confié et avoir suivi un enseignement sur le modèle linéaire mixte et l’estimation des composantes de la variance sans lequel ce document de synthèse n’aurait pas vu le jour.

Je suis particulièrement redevable à tous ceux avec qui j’ai travaillé sur ce sujet (Rohan Fernando, Daniel Gianola, Charles Henderson, Sotan Im, Florence Jaffrézic, Richard Quaas, Christèle Robert, Magali San Cristobal, Caroline Thaon d’Arnoldi, et David van Dyk) et qui m’ont permis, à l’occasion de nombreux échanges, de mieux comprendre les subtilités du BLUP, des équations du modèle mixte, de la méthode du maximum de vraisemblance et de l’algorithme EM.

Grand merci à Guy Lefort et à Paule Renaud pour leur enseignement irremplaçable de statistiques, à Larry Schaeffer pour m’avoir dispensé mon premier cours sur le modèle mixte et à Céline Delmas pour l’annexe sur la maximisation de la fonction de vraisemblance sous contraintes.

Ma gratitude va également à mes collègues de l’INRA, Christèle Robert, Bernard Bonaïti, et Jean-Jacques Colleau ainsi qu’à Jorge Colaço, Christian Lavergne et Gilles Celeux pour leur lecture critique de toute ou partie du manuscrit.

Chapitre I. **GENERALITES**

1. Rappels sur le modèle linéaire

11. Modèle linéaire classique

111. Ecriture du modèle

Classiquement, on écrit:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad (1.1)$$

où

$\mathbf{y}_{(N\text{x}1)}$: vecteur $(N\text{x}1)$ des variables aléatoires dépendantes (observations);

$\mathbf{X}_{(N\text{x}p)} = \left(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k, ..., \mathbf{X}_p\right)$: matrice d'incidence des variables explicatives (dites aussi variables «indépendantes» ou covariables) qui peuvent être continues (régression) ou discrètes (ANOVA);

$\boldsymbol{\beta}_{(p\text{x}1)}$: vecteur des coefficients dits de régression (covariables continues) ou «effets fixes» (covariables discrètes); a priori $\boldsymbol{\beta} \in R^p$;

$\mathbf{e}_{(N\text{x}1)}$: vecteur de variables aléatoires résiduelles

112. Hypothèses

Toutes ou partie des hypothèses suivantes peuvent être formulées ou requises selon les techniques statistiques employées:

-spécification exacte de l'espérance $\mathrm{E}(\mathbf{y})$,

et, vis-à-vis des variables aléatoires résiduelles:

-indépendance

-homoscédasticité

-normalité.

12. Estimation

121. Moindres carrés simples (OLS)

C'est une technique purement algébrique due à Legendre (1805) pour résoudre un système linéaire ayant plus d'équations que d'inconnues.

Soit $S(\boldsymbol{\beta}) = \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2$ le carré de la distance euclidienne entre les observations et la partie explicative du modèle, considérée comme une fonction de $\boldsymbol{\beta}$. La solution des moindres carrés en $\boldsymbol{\beta}$ est obtenue par minimisation de $S(\boldsymbol{\beta})$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \qquad (1.2)$$

Or

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{X} \text{ (définie non négative)}$$

La condition de convexité de la fonction $S(\boldsymbol{\beta})$ étant satisfaite, le minimum s'obtient par annulation des dérivées premières d'où le système dit des moindres carrés ou des équations normales:

$$\boxed{\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}}. \qquad (1.3)$$

Si $\mathbf{X}$ est de plein rang, $\boldsymbol{\beta}$ est «estimable», $\mathbf{X}'\mathbf{X}$ est inversible et la solution s'écrit $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$; sinon, il faut déterminer quelles sont les fonctions estimables $\mathbf{k}'\boldsymbol{\beta}$. Rappelons que $\mathbf{k}'\boldsymbol{\beta}$ est par définition une fonction estimable, si et seulement si, elle peut s'exprimer comme une combinaison linéaire de l'espérance des observations soit $\exists \mathbf{t},\ \mathbf{k}'\boldsymbol{\beta} = \mathbf{t}'\mathbf{E}(\mathbf{y}), \forall \boldsymbol{\beta}$.

L'estimation correspondante peut s'obtenir alors par l'utilisation d'une inverse généralisée:

$$\mathbf{k}'\hat{\boldsymbol{\beta}} = \mathbf{k}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}, \qquad (1.4)$$

qui assure la propriété d'invariance suivante: $\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{k}'$[1], cette formule pouvant être utilisée comme test d'estimabilité de toute combinaison linéaire $\boldsymbol{\lambda}'\boldsymbol{\beta}$ (Searle, 1971, page 185).

On gagnera le plus souvent à formuler le modèle avec une paramétrisation de plein rang, soit dès le départ, soit après une manipulation adéquate (cf annexe I-B). On utilise fréquemment la décomposition:

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}} = \mathbf{P}\mathbf{y} + (\mathbf{I} - \mathbf{P})\mathbf{y} \qquad (1.5)$$

où la matrice $\mathbf{P}$ est donnée par

$$\boxed{\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'} \qquad (1.6)$$

[1] Cette propriété découle de l'égalité $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X}$ quelle que soit l'inverse généralisée de $\mathbf{X}'\mathbf{X}$ (cf annexe I-A)

La matrice $\mathbf{P}$ est idempotente -tout comme $(\mathbf{I}-\mathbf{P})$-, et vérifie $\mathbf{PX}=\mathbf{X}$. En effet, le système (1.3) s'écrit aussi $\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{y}=\mathbf{0}$, $\forall\mathbf{y}$ d'où $\mathbf{PX}=\mathbf{X}$ et $\mathbf{P}=\mathbf{P}^2$. La matrice $\mathbf{P}$ s'interprète ainsi comme le projecteur orthogonal de $\mathbf{y}$ sur l'espace engendré par les colonnes de $\mathbf{X}$.

122. Propriétés

Sous l'hypothèse $\mathrm{E}(\mathbf{y})=\mathbf{X}\boldsymbol{\beta}$ (modèle correctement spécifié pour la partie systématique), l'estimateur des moindres carrés d'une fonction estimable est sans biais:

$$\boxed{\mathrm{E}(\mathbf{k}'\hat{\boldsymbol{\beta}})=\mathbf{k}'\boldsymbol{\beta}}. \qquad (1.7)$$

Démonstration: c'est la même que celle du critère d'estimabilité. $\mathbf{k}'\boldsymbol{\beta}$ étant une fonction estimable: $\exists\mathbf{t},\ \mathbf{k}'=\mathbf{t}'\mathbf{X}$, son estimateur des moindres carrés $\mathbf{k}'\hat{\boldsymbol{\beta}}$ peut se mettre alors sous la forme $\mathbf{k}'\hat{\boldsymbol{\beta}}=\mathbf{t}'\mathbf{P}\mathbf{y}$. Alors $\mathbf{E}(\mathbf{t}'\mathbf{P}\mathbf{y})=\mathbf{t}'\mathbf{P}\mathbf{X}\boldsymbol{\beta}$; or $\mathbf{PX}=\mathbf{X}$ d'où $\mathrm{E}(\mathbf{k}'\hat{\boldsymbol{\beta}})=\mathbf{t}'\mathbf{X}\boldsymbol{\beta}=\mathbf{k}'\boldsymbol{\beta}$, QED.

A noter que cette propriété ne nécessite aucune hypothèse sur la structure de variance covariance $\mathbf{V}$ des résidus.

Sous l'hypothèse additionnelle $\mathbf{V}=\sigma^2\mathbf{I}_N$ (indépendance et homoscédasticité), on montre que $\mathbf{k}'\hat{\boldsymbol{\beta}}$ est le meilleur estimateur linéaire sans biais (BLUE) de $\mathbf{k}'\boldsymbol{\beta}$, et que

$$\boxed{\mathrm{Var}(\mathbf{k}'\hat{\boldsymbol{\beta}})=\sigma^2\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{k}}. \qquad (1.8)$$

Comme précédemment, on part de la forme: $\mathbf{k}'\hat{\boldsymbol{\beta}}=\mathbf{t}'\mathbf{P}\mathbf{y}$ et $\mathrm{Var}\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right)=\mathbf{t}'\mathbf{P}\mathbf{P}\mathbf{t}\sigma^2$. En utilisant la propriété d'idempotence de $\mathbf{P}$ et en remplaçant $\mathbf{P}$ par son expression en (1.6), il vient $\mathrm{Var}\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right)=\mathbf{t}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{t}\sigma^2$ soit le résultat en (1.8) puisque

Enfin, sous l'hypothèse de normalité des résidus, la distribution de l'estimateur est elle aussi normale: $\boxed{\mathbf{k}'\hat{\boldsymbol{\beta}}\sim\mathcal{N}\left[\mathbf{k}'\boldsymbol{\beta},\sigma^2\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{k}\right]}$.

Soit $\mathrm{SSE}=\left\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2$, SSE peut s'écrire comme la forme quadratique suivante: $\mathrm{SSE}=\mathbf{y}'(\mathbf{I}-\mathbf{P})\mathbf{y}$, $(\mathbf{I}-\mathbf{P})$ étant idempotente; on en déduit tout d'abord un mode de calcul simple de SSE, soit

$$\mathrm{SSE}=\mathbf{y}'\mathbf{y}-\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}=\mathbf{y}'\mathbf{y}-\mathrm{R}(\boldsymbol{\beta}), \qquad (1.9)$$

où $R(\boldsymbol{\beta})$ désigne selon la notation de Searle (1971), la part de variation «expliquée» par le modèle qualifiée aussi de réduction due au modèle $\boldsymbol{\beta}$.

Par ailleurs $E(SSE) = \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta} + \mathrm{tr}(\mathbf{I}-\mathbf{P})\sigma^2$. Le premier terme s'annule puisque $\mathbf{PX} = \mathbf{X}$. De plus $(\mathbf{I}-\mathbf{P})$ étant idempotente, sa trace est égale à son rang soit $N - r(\mathbf{X})$ et $E(SSE) = \left[N - r(\mathbf{X})\right]\sigma^2$, d'où un estimateur sans biais de la variance résiduelle:

$$\hat{\sigma}^2 = SSE/\left[N - r(\mathbf{X})\right]. \qquad (1.10)$$

13. Tests d'hypothèses

Soit à tester l'hypothèse nulle $H_0 : \mathbf{k}'\boldsymbol{\beta} = \mathbf{m}$ contre son alternative contraire, $H_1 : \mathbf{k}'\boldsymbol{\beta} \neq \mathbf{m}$. où $\mathbf{k}'$ est une matrice $(r \times p)$ dont les r lignes sont linéairement indépendantes. Pour ce faire, on va se placer sous l'hypothèse forte suivante: $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Sous H_0 :

$$\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathrm{Var}(\mathbf{k}'\hat{\boldsymbol{\beta}})\right]^{-1}\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \sim \chi^2_{r(\mathbf{k})}. \qquad (1.11)$$

Or $\mathrm{Var}(\mathbf{k}'\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{k}$ (c.f. formule 1.8), la statistique définie par $Q = \left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{k}\right]^{-1}\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)$ est donc proportionnelle à un Khi-deux soit

$$Q \sim \sigma^2\chi^2_r \qquad (1.12)$$

De même,

$$SSE = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \sim \sigma^2\chi^2_{N-r(\mathbf{X})}. \qquad (1.13)$$

Comme Q et SSE sont indépendants, on peut donc former la statistique $F(H_0) = \dfrac{Q/\sigma^2 r}{SSE/\sigma^2\left[N - r(\mathbf{X})\right]}$ du rapport de deux variables Khi-deux divisée chacune par son nombre de degrés de liberté , et qui est une variable de Fisher-Snedecor. La variance inconnue σ^2 se simplifiant, on a donc:

$$\boxed{F(H_0) = \frac{Q/r}{SSE/\left[N - r(\mathbf{X})\right]} \sim F\left[r; N - r(\mathbf{X})\right]}. \qquad (1.14)$$

14. Interprétation géométrique

Considérons le sous espace vectoriel $C(\mathbf{X}) = \left\{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}; \boldsymbol{\beta} \in R^p; \mathbf{X}_{(Nxp)}\right\}$ engendré par les colonnes de $\mathbf{X}$, le principe des moindres carrés revient à chercher un vecteur de $C(\mathbf{X})$ qui minimise le carré de la norme euclidienne $\|\mathbf{y} - \boldsymbol{\mu}\|^2$. Géométriquement, il s'agit de la projection orthogonale de $\mathbf{y}$ sur $C(\mathbf{X})$. Cette projection est telle que $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ soit orthogonal à

tout vecteur colonne de $\mathbf{X}$, soit $< \mathbf{y} - \mathbf{X}\boldsymbol{\beta}, \mathbf{X}_j >= 0,\ \forall j=1,2,...,p$, $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_j, ..., \mathbf{X}_p)$ ce qui s'écrit encore

$$\begin{aligned} &\mathbf{X}_1^{'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\ &\mathbf{X}_2^{'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\ &\qquad\qquad \Leftrightarrow \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \Leftrightarrow \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \qquad (1.15) \\ &\mathbf{X}_j^{'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\ &\mathbf{X}_p^{'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \end{aligned}$$

Une illustration est fournie dans le plan Π pour $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2)$. On a donc par exemple:

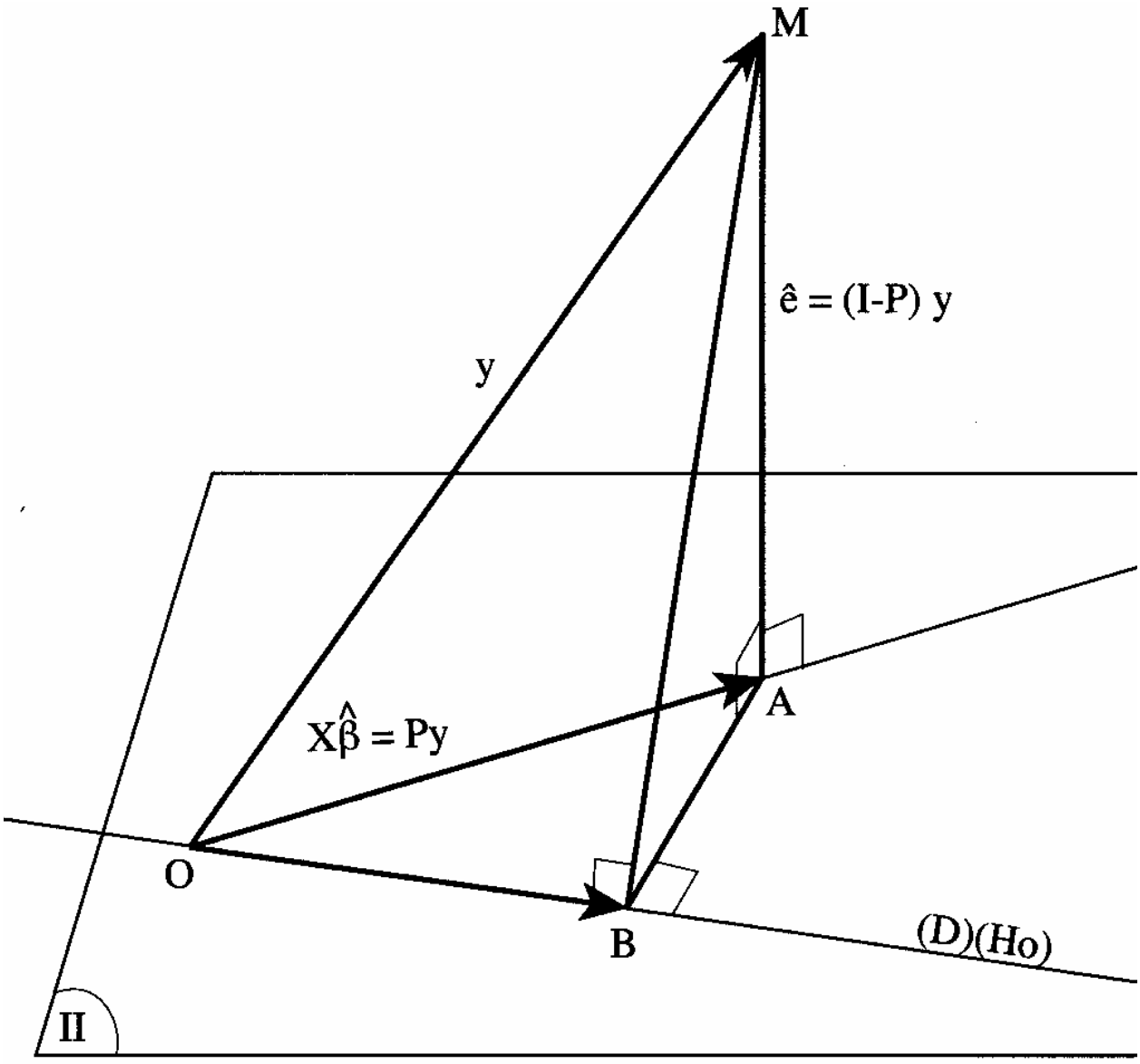
$< OM, OA >= \|OA\|^2 \Leftrightarrow \mathbf{y}'\mathbf{P}\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y}$

$\|OM\|^2 = \|OA\|^2 + \|AM\|^2 \Leftrightarrow \mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{P})\mathbf{y}$,

avec $\|OA\|^2 = \mathrm{R}(\boldsymbol{\beta}); \|AM\|^2 = SSE$, et qui traduit notamment l'orthogonalité de OA ($\mathbf{P}\mathbf{y}$) et de AM ($(\mathbf{I} - \mathbf{P})\mathbf{y}$).

De même, le test de l'hypothèse nulle $\mathrm{H}_0 : \mathbf{k}'\boldsymbol{\beta} = \mathbf{m}$ s'interprète comme la recherche d'une solution (vecteur OB) dans un sous-espace de $\mathrm{C}(\mathbf{X})$ de dimension $r < p$ telle que OB soit la projection orthogonale de OM sur ce sous-espace. Le triangle OBA est rectangle en B selon le théorème dit des «trois perpendiculaires», ce qui, formulé autrement, traduit le fait que cette solution est également la projection orthogonale sur ce sous-espace de la solution des moindres carrés du modèle complet.

Dans le cas de la partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{'}, \boldsymbol{\beta}_2^{'})'$ avec $\mathrm{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, on peut écrire: $\|OA\|^2 = \|OB\|^2 + \|BA\|^2 \Leftrightarrow \mathrm{R}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathrm{R}(\boldsymbol{\beta}_1) + \mathrm{R}(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)$ et le test de H_0 utilisera le fait que la statistique basée sur $\|BA\|^2$, $(\mathrm{R}(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)/[\mathrm{r}(\mathbf{X}) - \mathrm{r}(\mathbf{X}_1)]$ au numérateur du F) est indépendante de celle basée sur $\|AM\|^2$ ($[\mathbf{y}'\mathbf{y} - \mathrm{R}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)]/[N - \mathrm{r}(\mathbf{X})]$ au dénominateur) eu égard à l'orthogonalité des vecteurs BA et AM.



15. Généralisation

On considère le même modèle qu'en (1): $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, mais cette fois avec un vecteur de variables aléatoires résiduelles ayant une structure quelconque $\mathbf{V}$ de variance covariance $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{V})$.

$\mathbf{V}$ étant par définition une matrice définie positive, on peut lui appliquer une décomposition de Cholesky $\mathbf{V} = \mathbf{U}\mathbf{U}'$ où $\mathbf{U}$ est une matrice triangulaire inférieure de plein rang. Si l'on considère la transformation $\mathbf{y}^* = \mathbf{U}^{-1}\mathbf{y}$, le modèle correspondant à $\mathbf{y}^*$ s'écrit:

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \qquad (1.16)$$

$$\mathbf{X}^* = \mathbf{U}^{-1}\mathbf{X} \qquad (1.17)$$

$$\boldsymbol{\varepsilon}^* = \mathbf{U}^{-1}\boldsymbol{\varepsilon}. \qquad (1.18)$$

On se ramène ainsi au cas précédent d'un modèle linéaire classique avec des résidus indépendants et homoscédastiques puisque $\boldsymbol{\varepsilon}^* \sim (0, \mathbf{I}_N)$. On peut donc écrire le système sous la forme $\mathbf{X}^{*\prime}\mathbf{X}^*\hat{\boldsymbol{\beta}} = \mathbf{X}^{*\prime}\mathbf{y}$ qui équivaut avec les notations d'origine à:

$$\boxed{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}}\,. \qquad (1.19)$$

En fait, il suffit de connaître $\mathbf{V}$ à une constante près ainsi qu'on l'observe dans la forme $\mathbf{V} = \sigma^2\mathbf{H}$ où $\mathbf{H}$ est une matrice définie positive connue et σ^2 un scalaire positif inconnu.

La décomposition selon les moindres carrés (1.5) conduit à:

$$\mathbf{y}^* = \mathbf{P}^*\mathbf{y}^* + (\mathbf{I} - \mathbf{P}^*)\mathbf{y}^* \qquad (1.20)$$

où $\mathbf{P}^* = \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-}\mathbf{X}^{*\prime}$.

Revenant à $\mathbf{y}$ par la transformation inverse $\mathbf{y} = \mathbf{U}\mathbf{y}^*$, il vient

$\mathbf{y} = \mathbf{U}\mathbf{y}^* = \mathbf{U}\mathbf{P}^*\mathbf{U}^{-1}\mathbf{y} + (\mathbf{I} - \mathbf{U}\mathbf{P}^*\mathbf{U}^{-1})\mathbf{y}$ et, si l'on pose $\mathbf{Q} = \mathbf{U}\mathbf{P}^*\mathbf{U}^{-1}$, c'est-à-dire

$$\boxed{\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1}}\,, \qquad (1.21)$$

et, on a

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} = \mathbf{Q}\mathbf{y} + (\mathbf{I} - \mathbf{Q})\mathbf{y}\,. \qquad (1.22)$$

$\mathbf{Q}$ est le projecteur orthogonal de $\mathbf{y}$ sur $C(\mathbf{X})$ selon la métrique $\mathbf{V}^{-1}$ appelé aussi projecteur $\mathbf{V}^{-1}$ orthogonal; on le note quelquefois $\mathbf{Q} = \mathbf{P}_{\mathbf{X},\mathbf{V}^{-1}}$; de même $\mathbf{I} - \mathbf{Q}$ est le projecteur de $\mathbf{y}$ sur l'espace orthogonal à $C(\mathbf{X})$. $\mathbf{Q}$ est aussi idempotent et vérifie donc $\mathbf{Q}(\mathbf{I} - \mathbf{Q}) = \mathbf{0}$ (orthogonalité de $\mathbf{X}\hat{\boldsymbol{\beta}}$ et de $\hat{\boldsymbol{\varepsilon}}$).

En fait, la manipulation qui vient d'être effectuée peut s'interpréter comme un changement de base $\mathbf{Y} = \sum_i y_i\mathbf{e}_i = \sum_i y_i^*\mathbf{e}_i^*$ où les vecteurs de la nouvelle base (*) choisie orthonormale ($< \mathbf{e}_i^*, \mathbf{e}_j^* >= \delta_{ij}$), sont définis par la transformation $\mathbf{e}_i^* = \sum_k u_{ik}\mathbf{e}_k$ d'où il résulte que

$$< \mathbf{Y}, \mathbf{Y} >= \mathbf{y}^{*\prime}\mathbf{y}^* = \mathbf{y}'(\mathbf{U}^{-1})'\mathbf{U}^{-1}\mathbf{y} = \mathbf{y}'\mathbf{W}\mathbf{y}\,. \qquad (1.23)$$

La transformation $\mathbf{U}$ est choisie de façon que le jacobien $|\mathbf{J}|$ où $\mathbf{J} = \det\dfrac{\partial \mathbf{y}^{*\prime}}{\partial \mathbf{y}}$ soit, comme dans la loi normale, la racine carrée du déterminant de la précision $\mathbf{V}^{-1}$. Comme ici $\mathbf{J} = (\det \mathbf{U})^{-1} = (\det \mathbf{W})^{1/2}$, on prend $\mathbf{W} = \mathbf{V}^{-1}$.

2. Modèles linéaires mixtes

La théorie de l'analyse de variance en dispositif déséquilibré s'est développée principalement dans le cadre du modèle linéaire à effets fixes (Yates, 1934; Herr, 1986) sans qu'apparût de distinction nette entre effets fixes et aléatoires Il fallut attendre l'article d'Eisenhart (1947) pour que fussent clairement précisées les notions de modèles à effets fixes et de modèles à effets aléatoires. Il est donc nécessaire de bien définir ces concepts délicats à manipuler.

21. Définition.

211. Présentation de Rao et Kleffe

Un modèle linéaire mixte est un modèle linéaire tel qu'en (1.16) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{V})$, dans lequel la variable aléatoire $\boldsymbol{\varepsilon}$ est décomposée comme une combinaison linéaire des variables aléatoires structurales $\mathbf{u}_k;\ k = 0,1,2,...,K$ non observables (Rao and Kleffe, 1988, pages 61-63):

$$\boxed{\boldsymbol{\varepsilon} = \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k = \mathbf{Z}\mathbf{u}}, \qquad (1.24)$$

où $\mathbf{Z}_{(N \times q_+)} = \left(\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k, ..., \mathbf{Z}_K\right)$ est une concaténation de matrices connues $\mathbf{Z}_k$ de dimension $(N \times q_k)$ et $\mathbf{u}_{(q_+ x1)} = \left(\mathbf{u}_0', \mathbf{u}_1', \mathbf{u}_2', ..., \mathbf{u}_k', ..., \mathbf{u}_K'\right)'$ est le vecteur correspondant des variables structurales $\mathbf{u}_k = \{u_{kl}\};\ l$=1,2,...,$q_k$ tel que

$$\mathbf{u} \sim (\mathbf{0}, \boldsymbol{\Sigma}_u), \qquad (1.25)$$

Dans (1.25), $\boldsymbol{\Sigma}_u$ est une fonction linéaire de paramètres $\theta_m; m = 1,2,..., M$, les matrices $\mathbf{F}_m$ étant des matrices données carrées d'ordre $q_+ = \sum_{k=0}^{K} q_k$, soit

$$\boldsymbol{\Sigma}_u = \sum_{m=1}^{M} \theta_m \mathbf{F}_m . \qquad (1.26)$$

On ne posera pas de contraintes spécifiques sur θ_m et $\mathbf{F}_m$ dans le cas général hormis que ces paramètres et ces matrices doivent assurer la positivité de $\boldsymbol{\Sigma}_u$. Ce modèle étant posé, la matrice $\mathbf{V}$ de variance covariance des va observables $\mathbf{y}$ est une fonction linéaire en les paramètres $\theta_m; m = 1,2,..., M$ puisque par définition: $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}' = \sum_{m=1}^{M} \mathbf{Z}\mathbf{F}_m\mathbf{Z}'\theta_m$, soit, encore:

$$\boxed{\mathbf{V} = \sum_{m=1}^{M} \mathbf{V}_m \theta_m}. \qquad (1.27)$$

Cette propriété est une caractéristique de ce qu'on entend sous le vocable de «modèle linéaire mixte» qui est tel qu'à la fois, son espérance $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ et, sa variance $\mathbf{V} = \sum_{m=1}^{M} \mathbf{V}_m \theta_m$, sont des fonctions linéaires de paramètres.

Dans cette présentation, les effets aléatoires apparaissent en définitive comme un moyen de structurer la matrice de variance-covariance $\mathbf{V}$ des observations.

212. Exemples

Un exemple classique réside dans le modèle linéaire mixte à K facteurs aléatoires indépendants qui s'écrit:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{u}_k + \mathbf{e}\,, \tag{1.28}$$

ou encore, en incorporant la résiduelle dans la structure générale des u via $\mathbf{u}_0 = \mathbf{e}$; $\mathbf{Z}_0 = \mathbf{I}_N$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k\,, \tag{1.29}$$

et, pour lequel, $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{u}_k \sim (\mathbf{0}, \sigma_k^2 \mathbf{I}_k)$, $\mathrm{E}(\mathbf{u}_k \mathbf{u}_l^{'}) = 0, \forall k \neq l$ et, donc,

$$\boldsymbol{\Sigma}_u = \bigoplus_{k=0}^{K} \sigma_k^2 \mathbf{I}_k \text{ et } \mathbf{V} = \sum_{k=0}^{K} \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k^{'}\,.$$

Il est à noter que les mêmes propriétés de linéarité de $\mathbf{V}$ en les paramètres subsistent dans le cas où les facteurs $\mathbf{u}_k$ sont corrélés entre eux comme cela se produit dans les modèles «père» et «grand-père» des généticiens (Quaas et al, 1979) ou dans les modèles à coefficients de régression aléatoires (Laird et Ware, 1982). Ces modèles s'écrivent dans le cas le plus simple:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1 \mathbf{u}_1 + \mathbf{Z}_2 \mathbf{u}_2 + \mathbf{e} \tag{1.30}$$

où par exemple $\mathbf{u}_1 = \{u_{1i}\}$ est le vecteur des «intercepts» des individus (indicés par i) mesurés de façon répétée et $\mathbf{u}_2 = \{u_{2i}\}$, celui des pentes tels que

$$\mathrm{var}\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \begin{bmatrix} \sigma_1^2 \mathbf{I}_q & \sigma_{12} \mathbf{I}_q \\ \sigma_{12} \mathbf{I}_q & \sigma_2^2 \mathbf{I}_q \end{bmatrix} = \boldsymbol{\Sigma} \otimes \mathbf{I}_q\,, \tag{1.31}$$

avec $\boldsymbol{\Sigma} = \mathrm{var}\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ formée par les variances de l'intercept (σ_1^2), de la pente (σ_2^2) et leur covariance (σ_{12}).

Par définition du modèle, il vient:

$$\mathrm{var}\begin{pmatrix} \mathbf{e} \\ \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \sigma_0^2 \begin{bmatrix} \mathbf{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} + \sigma_1^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} + \sigma_2^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_q \end{bmatrix} + \sigma_{12} \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_q \\ \mathbf{0} & \mathbf{I}_q & \mathbf{0} \end{bmatrix} \tag{1.32},$$

qui suit bien la forme linéaire (1.27).

22. Approche marginale de modèles hiérarchiques

221. Présentation de Lindley-Smith

La référence en ce domaine est l'article de Lindley and Smith (1972) intitulé «Bayes estimates for the linear model». On considère un processus d'échantillonnage gaussien en deux étapes relatives aux données et aux paramètres de position respectivement:

a) $\mathbf{y} \mid \boldsymbol{\theta}_1, \mathbf{C}_1 \sim N(\mathbf{A}_1\boldsymbol{\theta}_1, \mathbf{C}_1)$,

b) $\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \mathbf{C}_2 \sim N(\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_2)$. (1.33)

La résultante de ces deux étapes conduit à la distribution marginale des données suivantes:

$$\mathbf{y} \mid \boldsymbol{\theta}_2, \mathbf{C}_2 \sim N(\mathbf{A}_1\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_1 + \mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{'}). \quad (1.34)$$

Pour s'en convaincre, il suffit d'écrire a) et b) sous forme de modèles linéaires soit

$$\begin{aligned} \mathbf{y} &= \mathbf{A}_1\boldsymbol{\theta}_1 + \mathbf{e};\ \mathbf{e} \sim N(\mathbf{0}, \mathbf{C}_1) \\ \boldsymbol{\theta}_1 &= \mathbf{A}_2\boldsymbol{\theta}_2 + \mathbf{u};\ \mathbf{u} \sim N(\mathbf{0}, \mathbf{C}_2) \end{aligned} \quad (1.35)$$

et, en reportant la deuxième équation dans la première, on obtient:

$$\mathbf{y} = \mathbf{A}_1\mathbf{A}_2\boldsymbol{\theta}_2 + \mathbf{A}_1\mathbf{u} + \mathbf{e}, \quad (1.36)$$

qu'on identifie bien à la structure $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$.

On se réfère quelquefois au qualificatif de «mélange» pour désigner l'approche marginale de modèles hiérarchiques. En effet, un mélange d'un nombre fini (p) de composantes ayant chacune pour loi $f(\mathbf{y} \mid \boldsymbol{\theta}_i)$ avec par exemple un vecteur de paramètres $\boldsymbol{\theta}_i = \left(\mu_i, \sigma_i^2\right)'$ comportant l'espérance μ_i et la variance σ_i^2 a une densité qui s'écrit:

$$f(\mathbf{y}) = \sum_{i=1}^{p} \pi_i f(\mathbf{y} \mid \boldsymbol{\theta}_i),$$

où π_i est la proportion de la composante i telle que $\sum_{i=1}^{p} \pi_i = 1$.

Cette sommation finie se généralise au cas continu et peut alors rendre compte d'un processus de marginalisation :

$$f(\mathbf{y}) = \int \pi(\boldsymbol{\mu}) f(\mathbf{y} \mid \boldsymbol{\mu}) d\boldsymbol{\mu}. \quad (1.37)$$

tel celui réalisé en (1.34) $f(\mathbf{y}) = \int \pi(\boldsymbol{\theta}_1 \mid \mathbf{C}_1) f(\mathbf{y} \mid \boldsymbol{\theta}_1, \mathbf{C}_1) d\boldsymbol{\theta}_1$ par intégration des paramètres $\boldsymbol{\theta}_1$.

222. Exemples

-Le modèle «père» de la sélection animale

Dans les espèces animales domestiques, la sélection des reproducteurs mâles s'effectue fréquemment à partir des performances de leurs descendants obtenus par accouplement de chacun des pères avec un échantillon de femelles reproductrices. Conditionnellement au

mérite génétique u_i de chaque père, un modèle simple d'analyse réside dans le modèle linéaire suivant (Henderson, 1973 ; Thompson, 1979):

$$y_{ij} \mid u_i = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + u_i + e_{ij}$$

avec

$$\mathrm{E}(y_{ij} \mid u_i) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + u_i;\ e_{ij} \sim iid(0, \sigma_e^2)$$

où y_{ij} est la performance du $j^{ème}$ descendant du père i; $\mathbf{x}_{ij}^{'}\boldsymbol{\beta}$ représente la contribution des facteurs systématiques de milieu et e_{ij} la résiduelle.

Si l'on pose:

$$\mathbf{y}_{N\,\mathrm{x}\,1} = \{y_{ij}\},\ \mathbf{u}_{q\,\mathrm{x}\,1} = \{u_i\},\ \mathbf{X}'_{p\,\mathrm{x}\,N} = \left(\mathbf{X}_1^{'}, \mathbf{X}_2^{'}, ..., \mathbf{X}_i^{'}, ..., \mathbf{X}_q^{'}\right),$$

$$\mathbf{X}_{i(p\,\mathrm{x}\,n_i)}^{'} = \left(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{ij}, ..., \mathbf{x}_{in_i}\right),\ \mathbf{Z}_{N\,\mathrm{x}\,q} = \bigoplus_{i=1}^{q} \mathbf{1}_{n_i},$$

on peut écrire la loi conditionnelle des observations sachant les mérites des pères sous la forme matricielle suivante:

$$\mathbf{y} \mid \mathbf{u} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}), \qquad (1.38)$$

avec $\mathbf{R} = \sigma_e^2 \mathbf{I}_N$.

En fait, il est d'usage de considérer que le vecteur $\mathbf{u}$ des effets des pères est lui-même une va d'espérance $\mathbf{Qg}$ et de matrice de variance covariance $\mathbf{G}$.

$$\mathbf{u} \sim \mathrm{N}(\mathbf{Qg}, \mathbf{G}). \qquad (1.39)$$

$\mathbf{Qg}$ correspond à une structuration des pères en différentes souches ou groupes ancestraux (Thompson, 1979; Quaas and Pollak, 1980; Quaas, 1988) et $\mathbf{G} = \mathbf{A}\sigma_u^2$ représente la variabilité génétique entre ceux-ci compte tenu de la matrice de parenté $\mathbf{A}$. Marginalement c'est-à-dire après avoir éliminé par intégration la variation des $\mathbf{u}$, on a:

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{ZQg}, \mathbf{ZGZ}' + \mathbf{R}), \qquad (1.40)$$

ce qui équivaut au modèle linéaire mixte suivant:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZQg} + \mathbf{Zu}^* + \mathbf{e}, \qquad (1.41)$$

avec $\mathbf{u}^* \sim \mathrm{N}(\mathbf{0}, \mathbf{A}\sigma_u^2)$ et $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$.

-Le modèle à coefficients de régression aléatoires

On peut introduire ce modèle à partir de l'exemple simple des données de croissance faciale à 4 âges (8, 10, 12 et 14 ans) de 11 filles et 16 garçons présentées par Pothoff et Roy (1964).

Ces données ont été analysées en détail par Verbeke et Molenberghs (1997) et Foulley et al (2000).

Un modèle simple et pertinent d'analyse de ces données consiste en l'ajustement d'une droite de régression propre à chaque individu. Si i désigne l'indice du sexe (i=1,2 pour les sexes femelle et mâle), j l'indice de la période de mesure (j=1,2,3,4) avec t_j le temps correspondant et k celui de l'individu intra-sexe (k=1,2,...,11 pour i=1; k=1,2,...,16 pour i=2), ce modèle s'écrit:

$$y_{ijk} = A_{ik} + B_{ik} t_j + e_{ijk}\,, \tag{1.42}$$

où A_{ik} est l'intercept propre à l'individu ik et B_{ik} la pente.

Conditionnellement aux valeurs des coefficients de régression A_{ik}, B_{ik} des individus, le modèle (1.42) est un modèle linéaire classique du type décrit en (1.1) à va résiduelles indépendantes.

Si, maintenant, en une deuxième phase du raisonnement, on considère que les individus représentent un échantillon aléatoire des enfants de chaque sexe, les A_{ik} et les B_{ik} sont des variables aléatoires qu'on peut aisément caractériser par leurs deux premiers moments:

$$\begin{pmatrix} A_{ik} \\ B_{ik} \end{pmatrix} \sim \left[\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \right]. \tag{1.43}$$

Cela revient à décomposer l'intercept et la pente en la somme de deux parties:

$$A_{ik} = \alpha_i + a_{ik}\,, \tag{1.44a}$$

$$B_{ik} = \beta_i + b_{ik} \tag{1.44b}$$

une composante systématique α_i et β_i propre à chaque sexe et un écart aléatoire centré a_{ik} et b_{ik} propre à l'individu k du sexe i.

Ce faisant, le modèle d'origine se met sous la forme usuelle:

$$y_{ijk} = \alpha_i + \beta_i t_j + a_{ik} + b_{ik} t_j + e_{ijk}\,, \tag{1.45}$$

qui sépare la partie fixe ($\alpha_i + \beta_i t_j$) de la partie aléatoire ($a_{ik} + b_{ik} t_j$).

Si l'on pose $\mathbf{y}_{ik} = \{y_{ijk}\}, \mathbf{e}_{ik} = \{e_{ijk}\}$, $\boldsymbol{\beta}_{4x1} = (\alpha_1, \alpha_2 - \alpha_1, \beta_1, \beta_2 - \beta_1)'$ $\mathbf{u}_{ik} = (a_{ik}, b_{ik})'$ auxquels correspondent les matrices d'incidence $\mathbf{X}_{ik} = (\mathbf{1}_4, \mathbf{0}_4, \mathbf{t}, \mathbf{0}_4)$ si $i = 1$, $\mathbf{X}_{ik} = (\mathbf{1}_4, \mathbf{1}_4, \mathbf{t}, \mathbf{t})$ si $i = 2$ et $\mathbf{Z}_{ik} = (\mathbf{1}_4, \mathbf{t})$ avec $\mathbf{t}_{4x1} = \{t_j\}$, (1.45) s'écrit sous la forme matricielle typique d'un modèle linéaire mixte:

$$\mathbf{y}_{ik} = \mathbf{X}_{ik}\boldsymbol{\beta} + \mathbf{Z}_{ik}\mathbf{u}_{ik} + \mathbf{e}_{ik}\,, \tag{1.46}$$

où $\mathbf{u}_{ik} \sim (\mathbf{0}, \mathbf{G})$, $\mathbf{e}_{ik} \sim (\mathbf{0}, \mathbf{R})$ avec $\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}$ et $\mathbf{R} = \sigma_e^2 \mathbf{I}_4$.

Un modèle linéaire mixte apparaît donc comme un modèle linéaire dans lequel toute ou partie des paramètres associés à certaines unités expérimentales sont traités comme des variables aléatoires du fait de l'échantillonnage de ces unités dans une population plus large.

ANNEXE I-A

Démonstration $\mathbf{PX} = \mathbf{X}$

C'est une conséquence des deux lemmes suivants (Searle, 1982, p62-63)

Lemme 1 : Pour toute matrice réelle $\mathbf{A}_{(n\,x\,n)} = \{a_{ij}\}$; $\mathbf{A}'\mathbf{A} = \mathbf{0} \Rightarrow \mathbf{A} = \mathbf{0}$.

En effet, le jème élément diagonal du produit s'écrit $\sum_{i=1}^{n} a_{ji}^2$ et sa nullité implique $a_{ji} = 0$, $\forall i$ et cela est vrai aussi $\forall j$.

Lemme 2 : Pour toutes matrices réelles $\mathbf{R}$, $\mathbf{S}$ et $\mathbf{X}$, $\mathbf{RX}'\mathbf{X} = \mathbf{SX}'\mathbf{X} \Rightarrow \mathbf{RX}' = \mathbf{SX}'$

Cela découle de l'identité suivante : $(\mathbf{RX}'\mathbf{X} - \mathbf{SX}'\mathbf{X})(\mathbf{R} - \mathbf{S})' = (\mathbf{RX}' - \mathbf{SX}')(\mathbf{RX}' - \mathbf{SX}')'$.

Si $\mathbf{RX}'\mathbf{X} = \mathbf{SX}'\mathbf{X}$, la relation ci-dessus est nulle ; on peut donc appliquer le lemme 1 au membre de droite d'où $\mathbf{RX}' = \mathbf{SX}'$.

Par application du lemme 2 à $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$, on a $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' = \mathbf{X}'$, soit en transposant $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X}$, ie $\mathbf{PX} = \mathbf{X}$, QED.

ANNEXE I-B

Paramétrisation et codage

1) Dispositif à deux facteurs croisés A et B suivant:

A\B	(1)	(2)	(3)
(1)	μ_{11}	μ_{12}	μ_{13}
(2)	μ_{21}	μ_{22}	μ_{23}

où μ_{ij} est l'espérance des observations de la ligne i et de la colonne j.

2) Analyse par un modèle additif: $\mu_{ij} = \mu + a_i + b_j$

21) Paramétrisation en écart à une cellule de référence («incremental effects models» SAS); GLM, Mixed, Genmod

		a		b		
	μ	(1)	(2)	(1)	(2)	(3)
μ_{11}	1	1	0	1	0	0
μ_{12}	1	1	0	0	1	0
μ_{13}	1	1	0	0	0	1
μ_{21}	1	0	1	1	0	0
μ_{22}	1	0	1	0	1	0
μ_{23}	1	0	1	0	0	1

La règle pratique consiste à éliminer les colonnes de **X** relatives aux niveaux de la cellule de référence: ici a_2 et b_3. C'est la convention choisie par SAS de mise à zéro des niveaux d'indice les plus élevés. La paramétrisation relative à ce codage de **X** est la suivante : $\dot{\mu} = \mu + a_2 + b_3$; $\dot{a}_1 = a_1 - a_2$; $\dot{b}_1 = b_1 - b_3$; $\dot{b}_2 = b_2 - b_3$.

22) Paramétrisation « $\Sigma = 0$ » («deviation from the mean model» SAS); Catmod, Logistic

		a		b		
	μ	(1)	(2)	(1)	(2)	(3)
μ_{11}	1	1	0	1	0	0
μ_{12}	1	1	0	0	1	0
μ_{13}	1	1	0	-1	-1	1
μ_{21}	1	-1	1	1	0	0
μ_{22}	1	-1	1	0	1	0
μ_{23}	1	-1	1	-1	-1	1

La matrice **X** a toujours 4 colonnes pour que la paramétrisation soit de plein rang. Cette fois, on retranche la colonne a_2 de celle de a_1 et la colonne b_3 de celle de b_1 et de celle de b_2, les colonnes a_2 et b_3 étant écartées. La paramétrisation relative à ce codage de **X** est la suivante: $\ddot{\mu} = \mu + \overline{a} + \overline{b}$ où $\overline{a} = (a_1 + a_2)/2$ et $\overline{b} = (b_1 + b_2 + b_3)/3$; $\ddot{a}_1 = a_1 - \overline{a} = (a_1 - a_2)/2$; $\ddot{b}_1 = b_1 - \overline{b} = (2b_1 - b_2 - b_3)/3$; $\ddot{b}_2 = b_2 - \overline{b} = (b_1 - 2b_2 - b_3)/3$.

3) Analyse par un modèle avec interaction: $\mu_{ij} = \mu + a_i + b_j + (ab)_{ij}$

La matrice **X** a maintenant 6 colonnes, les 4 précédentes auxquelles s'ajoutent deux colonnes pour les effets d'interaction. Celles-ci s'obtiennent en multipliant la colonne a_1 par respectivement celle de b_1 et celle de b_2. Dans le cas d'une paramétrisation en écart à une cellule de référence, on obtient ainsi:

		a	b		ab	
	μ	(1)	(1)	(2)	(11)	(12)
μ_{11}	1	1	1	0	1	0
μ_{12}	1	1	0	1	0	1
μ_{13}	1	1	0	0	0	0
μ_{21}	1	0	1	0	0	0
μ_{22}	1	0	0	1	0	0
μ_{23}	1	0	0	0	0	0

La paramétrisation relative à ce codage de **X** est la suivante:

$\hat{\mu} = \mu_{23}$; $\hat{a}_1 = \mu_{13} - \mu_{23}$; $\hat{b}_1 = \mu_{21} - \mu_{23}$; $\hat{b}_2 = \mu_{22} - \mu_{23}$;

$(a\hat{b})_{11} = (\mu_{11} - \mu_{21}) - (\mu_{13} - \mu_{23})$; $(a\hat{b})_{12} = (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23})$.

De même, avec la paramétrisation « $\Sigma = 0$ », **X** s'écrit:

		a	b		ab	
	μ	(1)	(1)	(2)	(11)	(12)
μ_{11}	1	1	1	0	1	0
μ_{12}	1	1	0	1	0	1
μ_{13}	1	1	-1	-1	-1	-1
μ_{21}	1	-1	1	0	-1	0
μ_{22}	1	-1	0	1	0	-1
μ_{23}	1	-1	-1	-1	1	1

La paramétrisation relative à ce codage de X est la suivante :

$\breve{\mu} = \mu_{..}$; $\breve{a}_1 = \mu_{1.} - \mu_{..}$; $\breve{b}_1 = \mu_{.1} - \mu_{..}$; $\breve{b}_2 = \mu_{.2} - \mu_{..}$;

$(a\breve{b})_{11} = \mu_{11} - \mu_{1.} - \mu_{.1} + \mu_{..}$; $(a\breve{b})_{12} = \mu_{12} - \mu_{1.} - \mu_{.2} + \mu_{..}$,

où $\mu_{i.} = \left(\sum_{j=1}^{J} \mu_{ij}\right)/J$, $\mu_{.j} = \left(\sum_{i=1}^{I} \mu_{ij}\right)/I$ et $\mu_{..} = \left(\sum_{i=1}^{I} \mu_{i.}\right)/I = \left(\sum_{j=1}^{J} \mu_{.j}\right)/J$.

3) Justification

On se pose la question de déterminer X sous une paramétrisation de plein rang. On part d'un modèle initial $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ où $\mathbf{X}$ est une matrice $(N \times k)$ qui n'est pas de plein rang suivant les colonnes $k > \mathrm{r}(\mathbf{X}) = p$ et $\boldsymbol{\beta}$ est le vecteur correspondant $(k \times 1)$ des coefficients de régression. On introduit un vecteur $\dot{\boldsymbol{\beta}}$ $(p \times 1)$ relatif au modèle $\mathrm{E}(\mathbf{y}) = \dot{\mathbf{X}}\dot{\boldsymbol{\beta}}$ ayant une paramétrisation de plein rang et tel que $\dot{\boldsymbol{\beta}} = \mathbf{T}\boldsymbol{\beta}$, $\mathbf{T}$ étant une matrice de passage $(p \times k)$ connue et de plein rang suivant les lignes. En identifiant les deux modèles, il vient $\dot{\mathbf{X}}\mathbf{T} = \mathbf{X}$ ou encore $\dot{\mathbf{X}}\mathbf{T}\mathbf{T}' = \mathbf{X}\mathbf{T}'$. Comme $\mathbf{T}$ est de plein rang suivant les lignes, $\mathbf{T}\mathbf{T}'$ est inversible et $\dot{\mathbf{X}}$ s'écrit:

$$\boxed{\dot{\mathbf{X}} = \mathbf{X}\mathbf{T}'(\mathbf{T}\mathbf{T}')^{-1}}$$

Exemple: paramétrisation additive

1) en écart à la cellule (23)

$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$; $\mathbf{T} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$. On trouve bien par application de la formule: $\dot{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$.

2) paramétrisation « $\Sigma = 0$ »,

T s'écrit: $\mathbf{T} = \begin{bmatrix} 1 & 1/2 & 1/2 & 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & -1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2/3 & -1/3 & -1/3 \\ 0 & 0 & 0 & -1/3 & 2/3 & -1/3 \end{bmatrix}$ et on obtient $\dot{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$

Chapitre II. PREDICTION

1. Approche directe
    11. Définition
    12. Meilleur prédicteur
    13. Meilleur prédicteur linéaire
    14. BLUP
        141. formulation classique
        142. formulation de Bulmer
2. Equations du modèle mixte
    21. Approche d'Henderson
    22. Justification
    23. Variances d'erreur
    24. Interprétation bayésienne
3. Conclusion

Annexe II

A. Inverse de V

1. Approche directe

11. Définition

Le concept de prédiction apparaît rarement dans la littérature tel quel; il reste la plupart du temps indéfini ou confondu avec celui de l'estimation alors qu'il s'en distingue nettement.

Faire une prédiction, c'est substituer à une variable aléatoire W -non observable dans les conditions du problème- une variable aléatoire $\hat{W}$ qui est fonction d'une variable aléatoire Y observable[2]ie $\hat{W} = f(Y)$ et telle que la distribution de $\hat{W}$ soit aussi proche que possible de celle de W selon un critère donné. Il pourra s'agir d'une distance (ex celle de Kullback-Leibler) ou d'un critère tel que l'erreur quadratique moyenne $MSE = \mathrm{E}\left[(\hat{W} - W)^2\right]$.

12. Meilleur prédicteur

Nous utiliserons ici la terminologie d'Henderson. Il s'agit ici du meilleur prédicteur au sens de l'erreur quadratique moyenne (abréviation BP en anglais). Soit $\hat{W} = f(Y)$ le prédicteur et $\hat{w} = f(y)$ une réalisation, l'erreur quadratique moyenne se décompose en

$$\mathrm{E}\left[(\hat{W} - W)^2\right] = \mathrm{Var}(\hat{W} - W) + \left[\mathrm{E}(\hat{W}) - \mathrm{E}(W)\right]^2. \qquad (2.1)$$

On peut appliquer au premier terme de (2.1) le théorème de conditionnement-déconditionnement de la variance, soit

$$\mathrm{Var}(\hat{W} - W) = \mathrm{E}_Y\left\{\mathrm{Var}\left[(\hat{W} - W)\middle|Y = y\right]\right\} + \mathrm{Var}_Y\left\{\mathrm{E}\left[(\hat{W} - W)\middle|Y = y\right]\right\}. \qquad (2.2)$$

Or, conditionnellement à $Y = y$, $\hat{W}\middle|Y = y$ est une constante égale à $\hat{w}$; son espérance est donc $\hat{w}$ et sa variance est nulle, si bien que (2.2) se réduit alors à:

$$\mathrm{Var}(\hat{W} - W) = \mathrm{E}_Y\left[\mathrm{Var}(W\middle|Y = y)\right] + \mathrm{Var}_Y\left[\hat{w} - \mathrm{E}(W\middle|Y = y)\right]. \qquad (2.3)$$

Le premier terme de (2.3) ne dépend pas du choix du prédicteur; le second s'annule si l'on prend un prédicteur tel que

$$\boxed{\hat{w} = \mathrm{E}(W\middle|Y = y)}. \qquad (2.4)$$

Par construction, ce prédicteur vérifie

$$\mathrm{E}(\hat{W}) = \mathrm{E}_Y\left[\mathrm{E}(W\middle|Y = y)\right] = \mathrm{E}(W) \qquad (2.5)$$

Il est donc sans biais -au sens de $\mathrm{E}(\hat{W}) = \mathrm{E}(W)$ - et minimise (2.1) par construction; il vérifie également la propriété suivante:

$$MSE = \mathrm{Var}(\hat{W} - W) = \mathrm{E}_Y\left[\mathrm{Var}(W\middle|Y = y)\right]. \qquad (2.6)$$

[2] Aucune hypothèse à ce stade sur la dimension de Y

Ces résultats s'appliquent aisément au cas gaussien avec $\boldsymbol{\mu}' = \left(\mu_W, \boldsymbol{\mu}_Y'\right)$ et $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{WW} & \boldsymbol{\Sigma}_{WY} \\ \boldsymbol{\Sigma}_{YW} & \boldsymbol{\Sigma}_{YY} \end{bmatrix}$. On connaît alors la forme de la loi conditionnelle de $W \left| \mathbf{Y} = \mathbf{y} \right.$ qui est aussi normale

$$\boxed{W \left| \mathbf{Y} = \mathbf{y} \sim \mathcal{N}(\mu_{W.Y}, \boldsymbol{\Sigma}_{WW.Y}) \right.}, \qquad (2.7)$$

où, l'espérance $\mu_{W.Y} = \mu_W + \boldsymbol{\Sigma}_{WY}\boldsymbol{\Sigma}_{YY}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_Y\right)$ est linéaire en $\mathbf{y}$, et la variance $\boldsymbol{\Sigma}_{WW.Y} = \boldsymbol{\Sigma}_{WW} - \boldsymbol{\Sigma}_{WY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YW}$ ne dépend pas de la valeur particulière de $\mathbf{y}$ pour laquelle on conditionne.

13. Meilleur prédicteur linéaire

Comme précédemment, on suppose que les deux premiers moments de la loi conjointe $\left(W, \mathbf{Y}'\right)$ sont connus, mais la forme précise de celle-ci ne l'est pas. Une des possibilités est de se restreindre à une classe particulière de prédicteurs, en l'occurrence aux prédicteurs linéaires de la forme: $\hat{W} = a_0 + \mathbf{a}'\left(\mathbf{Y} - \boldsymbol{\mu}_Y\right)$. Dans ces conditions:

$$\mathrm{E}\left(\hat{W} - W\right) = a_0 - \mu_W$$

$$\mathrm{Var}\left(\hat{W} - W\right) = \mathbf{a}'\boldsymbol{\Sigma}_{YY}\mathbf{a} - 2\mathbf{a}'\boldsymbol{\Sigma}_{YW} + \boldsymbol{\Sigma}_{WW}$$

Soit $Q = \mathrm{E}\left[\left(\hat{W} - W\right)^2\right]$, la minimisation de Q par rapport aux coefficients a_0 et $\mathbf{a} = \{a_k\}$; $k = 1, 2, ..., N$ (N étant la dimension de $\mathbf{Y}$) conduit aux équations aux dérivées partielles suivantes:

$$\frac{\partial Q}{\partial a_0} = 2(a_0 - \mu_W) = 0$$

$$\frac{\partial Q}{\partial \mathbf{a}} = 2\boldsymbol{\Sigma}_{YY}\mathbf{a} - 2\boldsymbol{\Sigma}_{YW} = 0\,.$$

Il vient immédiatement: $a_0 = \mu_W$ (propriété de non biais) et $\mathbf{a} = \boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YW}$ si bien que le prédicteur s'écrit en définitive:

$$\boxed{\hat{W} = \mu_W + \boldsymbol{\Sigma}_{WY}\boldsymbol{\Sigma}_{YY}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_Y)}, \qquad (2.8)$$

et a la même forme que le meilleur prédicteur résultant du cas gaussien. Ce prédicteur vérifie

$$Q = \mathrm{Var}\left(\hat{W} - W\right) = \mathrm{Var}(W) - \mathrm{Var}(\hat{W}) \qquad (2.9)$$

avec

$$\mathrm{Var}(\hat{W}) = \mathbf{a}'\boldsymbol{\Sigma}_{YY}\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}_{YW} = \boldsymbol{\Sigma}_{WY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YW}. \qquad (2.10)$$

14. Meilleur prédicteur linéaire sans biais (BLUP)

Ce type de prédicteur est connu mondialement sous l'acronyme anglais de BLUP[3] («Best Linear Unbiased Predictor») suite aux travaux notamment de CR Henderson à l'Université Cornell et de ses élèves (Harville, Quaas, Schaeffer).

Ce prédicteur a été proposé pour répondre à la levée de l'hypothèse de moments connus de 1er ordre de la distribution de $(W, \mathbf{Y}')$. On va supposer que ceux-ci sont des fonctions linéaires d'un vecteur de paramètres inconnus $\boldsymbol{\beta} \in \mathbb{R}^p$ ie $\mu_W = \mathbf{k}'\boldsymbol{\beta}$ et $\boldsymbol{\mu}_Y = \mathbf{X}\boldsymbol{\beta}$. En fait, on exprime la variable aléatoire à prédire sous la forme: $W = \mu_W + \mathbf{m}'\mathbf{u}$. Dans ces conditions le problème se formalise dans un cadre de modèle linéaire mixte: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ où $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ et $\mathrm{Var}(\mathbf{Y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ avec $\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$ et $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$.

141. Formulation classique

On recherche un prédicteur qui soit a priori

a-linéaire,

b-sans biais au sens de $\mathrm{E}(\hat{W}) = \mathrm{E}(W)$,

c-optimum au sens de l'erreur quadratique moyenne (MSE) minimum.

Ces conditions se traduisent respectivement par:

a) $\hat{W} = \mathbf{a}'\mathbf{Y}$

b) $\mathbf{X}'\mathbf{a} - \mathbf{k} = \mathbf{0}_{p \times 1}$

c) $\mathrm{Var}(\hat{W} - W) = \mathbf{a}'\mathbf{V}\mathbf{a} + \mathbf{m}'\mathbf{G}\mathbf{m} - 2\mathbf{a}'\mathbf{C}\mathbf{m}$ minimum où $\mathbf{C} = \mathrm{Cov}(\mathbf{Y}, \mathbf{u}') = \mathbf{Z}\mathbf{G}$.

Minimiser l'expression en c) sous la contrainte b) revient à minimiser la fonction

$$Q(\mathbf{a}, \boldsymbol{\theta}) = \mathbf{a}'\mathbf{V}\mathbf{a} - 2\mathbf{a}'\mathbf{C}\mathbf{m} + 2\boldsymbol{\theta}'(\mathbf{X}'\mathbf{a} - \mathbf{k}), \qquad (2.11)$$

où $\boldsymbol{\theta}$ est un vecteur $(p \times 1)$ de multiplicateurs de Lagrange.

Les dérivées partielles par rapport à $\mathbf{a}$ et $\boldsymbol{\theta}$ s'écrivent:

$$\frac{\partial Q}{\partial \mathbf{a}} = 2\mathbf{V}\mathbf{a} - 2\mathbf{C}\mathbf{m} + 2\mathbf{X}\boldsymbol{\theta}, \qquad (2.12)$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}} = 2(\mathbf{X}'\mathbf{a} - \mathbf{k}). \qquad (2.13)$$

Par annulation, on tire: $\mathbf{a} = \mathbf{V}^{-1}(\mathbf{C}\mathbf{m} - \mathbf{X}\boldsymbol{\theta})$ et, en reportant dans (2.13), il vient:

$\mathbf{X}'\mathbf{V}^{-1}(\mathbf{C}\mathbf{m} - \mathbf{X}\boldsymbol{\theta}) = \mathbf{k}$ soit, en résolvant en $\boldsymbol{\theta}$, puis en reportant dans l'expression de $\mathbf{a}$:

[3] Le sigle a été en fait introduit par Goldberger (1962)

$\mathbf{a}' = \left(\mathbf{m}'\mathbf{C}' - \boldsymbol{\theta}'\mathbf{X}'\right)\mathbf{V}^{-1}$, d'où

$$\hat{W} = \mathbf{a}'\mathbf{Y} = \mathbf{m}'\mathbf{C}'\mathbf{V}^{-1}\mathbf{Y} + \left(\mathbf{k}' - \mathbf{m}'\mathbf{C}'\mathbf{V}^{-1}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},$$

qui, après réarrangement, se met sous la forme:

$$\boxed{\hat{W} = \mathbf{k}'\hat{\boldsymbol{\beta}} + \mathbf{m}'\mathbf{C}'\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}, \qquad (2.14)$$

où $\hat{\boldsymbol{\beta}}$ est l'estimateur GLS de $\boldsymbol{\beta}$, solution du système:

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \qquad (2.15)$$

et $\mathbf{C}'\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ s'obtient à partir du meilleur prédicteur linéaire de $\mathbf{u}$, ie $\hat{\mathbf{u}} = \mathrm{Cov}(\mathbf{u}, \mathbf{Y}')\left[\mathrm{Var}(\mathbf{Y})\right]^{-1}\left[\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right]$(cf (2.8)) dans lequel on a remplacé $\mathrm{E}(\mathbf{Y})$ par son estimateur GLS, $\mathbf{X}\hat{\boldsymbol{\beta}}$. Ce résultat est du à Goldberger (1962; page 371, eq 3.13) et Henderson (1963; page 161, equations 19 et 20).

142. Formulation de Bulmer

Bulmer (1980) s'intéresse à la meilleure prédiction d'une variable centrée telle que $\mathbf{u}$ à partir d'une variable observable $\mathbf{Y}$ d'espérance inconnue $\mathbf{X}\boldsymbol{\beta}$. Pour ce faire, il procède en deux étapes:

-corriger les observables pour les effets systématiques estimés par GLS, soit $\mathbf{Y}_c = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$

-prédire $\mathbf{u}$ par le meilleur prédicteur linéaire de $\mathbf{Y}_c$ ce qui est légitime puisque $\mathbf{u}$ et $\mathbf{Y}_c$ ont une espérance connue –en l'occurrence nulle.

Un tel prédicteur $\tilde{\mathbf{u}}$ s'écrit:

$$\tilde{\mathbf{u}} = \mathrm{Cov}(\mathbf{u}, \mathbf{Y}_c')\left[\mathrm{Var}(\mathbf{Y}_c)\right]^{-}\mathbf{Y}_c. \qquad (2.16)$$

Il est à noter que cette expression fait intervenir une inverse généralisée des observables puisque, du fait de la correction, la matrice de variance covariance correspondante n'est plus de rang N mais de rang $N - \mathrm{rang}(\mathbf{X})$. En fait $\mathbf{Y}_c$ peut s'écrire sous la forme $\mathbf{Y}_c = \mathbf{V}\underline{\mathbf{P}}\mathbf{Y}$ où $\underline{\mathbf{P}} = \mathbf{V}^{-1}\left(\mathbf{I} - \mathbf{Q}\right)$ et $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ est le projecteur défini au chapitre I (cf 1.21). Alors

$\mathrm{Var}(\mathbf{Y}_c) = \mathbf{V}\underline{\mathbf{P}}\mathbf{V}\underline{\mathbf{P}}\mathbf{V} = \mathbf{V}\underline{\mathbf{P}}\mathbf{V}$ ($\mathbf{V}$ étant une inverse généralisée de $\underline{\mathbf{P}}$),

$\mathrm{Cov}(\mathbf{u}, \mathbf{Y}_c') = \mathbf{C}'\underline{\mathbf{P}}\mathbf{V}$, $\mathbf{V}^{-1} = \left(\mathbf{V}\underline{\mathbf{P}}\mathbf{V}\right)^{-}$

et, en remplaçant dans (2.16), on obtient:

$$\tilde{\mathbf{u}} = \mathbf{C}'\underline{\mathbf{P}}\mathbf{Y}_c = \mathbf{C}'\underline{\mathbf{P}}\mathbf{V}\underline{\mathbf{P}}\mathbf{Y} = \mathbf{C}'\underline{\mathbf{P}}\mathbf{Y}, \qquad (2.17)$$

qui correspond bien au BLUP, $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ de $\mathbf{u}$, basé sur $\mathbf{Y}$ (Gianola et Goffinet, 1982).

L'expression (2.17) illustre bien la propriété d'invariance par translation du BLUP puisque $\hat{\mathbf{u}}$ est bâti sur $\left(\mathbf{I} - \mathbf{Q}\right)\mathbf{Y}$ et que $\mathbf{Q}\mathbf{X} = \mathbf{X}$.

Cette expression permet également d'établir aisément que:

$$\mathrm{Var}(\hat{\mathbf{u}}) = \mathrm{Cov}(\hat{\mathbf{u}}, \mathbf{u}') = \mathbf{C}'\underline{\mathbf{P}}\mathbf{C}, \qquad (2.18)$$

et donc, du fait de l'absence de corrélation entre $\hat{\mathbf{u}}$ et $\hat{\mathbf{u}} - \mathbf{u}$, que:

$$\mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{G} - \mathbf{C}'\underline{\mathbf{P}}\mathbf{C}. \qquad (2.19)$$

On notera à ce propos que s'il s'était agi du meilleur prédicteur linéaire-cas où $\boldsymbol{\mu}_Y$ ou $\boldsymbol{\beta}$ est connu-, on aurait eu:

$$\mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{G} - \mathbf{C}'\mathbf{V}^{-1}\mathbf{C}. \qquad (2.20)$$

On notera que, dans (2.19), $\underline{\mathbf{P}}$ remplace $\mathbf{V}^{-1}$; on retrouvera cette substitution dans l'estimation des composantes de la variance quand on passe du maximum de vraisemblance au maximum de vraisemblance restreinte.

2. Equations du modèle mixte

21. Approche d'Henderson

C'est dans un article collectif de Biometrics publié en 1959 (Henderson et al, 1959) qu'Henderson présente le système des équations dites du modèle mixte. Celles-ci sont relatives à l'estimateur des moindres carrés généralisés, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ de $\boldsymbol{\beta}$ et au Blup, $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ de $\mathbf{u}$ qui interviennent dans un modèle linéaire mixte de la forme:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad (2.21)$$

où $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ et $\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ avec $\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$ et $\mathrm{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$.

Ce système s'écrit comme suit:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \qquad (2.22)$$

L'intérêt calculatoire de ce système est manifeste puis qu'il ne nécessite plus le calcul de l'inverse de la matrice $\mathbf{V}$ de variance-covariance des observations dont la dimension N peut être très élevée; celle-ci n'a pas dans le cas général de structure simple contrairement à $\mathbf{R}$.

22. Justification

Henderson présente ce système de façon détournée -et un peu étrange- comme le résultat de la maximisation de la densité conjointe $f(\mathbf{y},\mathbf{u})$ (ou de son logarithme) par rapport à $\boldsymbol{\beta}$ et $\mathbf{u}$ sous l'hypothèse de multinormalité, soit:

$$\tilde{\boldsymbol{\beta}},\tilde{\mathbf{u}} = \arg\max_{\boldsymbol{\beta},\mathbf{u}} \ln f(\mathbf{y},\mathbf{u}) . \qquad (2.23)$$

On résumera ici les étapes de cette maximisation. On part de la décomposition de la densité conjointe en le produit suivant: $f(\mathbf{y},\mathbf{u}) = f(\mathbf{y}|\mathbf{u})f(\mathbf{u})$ où $\mathbf{y}|\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u},\mathbf{R})$ et $\mathbf{u} \sim \mathcal{N}(\mathbf{0},\mathbf{G})$. De là, on tire:

$$-2\ln f(\mathbf{y}|\mathbf{u}) = N\ln 2\pi + \ln|\mathbf{R}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})$$

$$-2\ln f(\mathbf{u}) = q\ln 2\pi + \ln|\mathbf{G}| + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u} .$$

La minimisation de la somme de ces deux termes considérée comme fonction de $\boldsymbol{\beta}$ et $\mathbf{u}$ se fait par l'écriture et l'annulation des dérivées partielles. Soit $l(\boldsymbol{\beta},\mathbf{u};y) = \ln f(\mathbf{y},\mathbf{u})$, on a:

$$\frac{\partial[-2l(\boldsymbol{\beta},\mathbf{u};y]}{\partial\boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) = \mathbf{0}$$

$$\frac{\partial[-2l(\boldsymbol{\beta},\mathbf{u};y]}{\partial\mathbf{u}} = -2\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) + 2\mathbf{G}^{-1}\mathbf{u} = \mathbf{0}$$

d'où découle immédiatement le système (2.22).

Cependant, $l(\boldsymbol{\beta},\mathbf{u};y)$ n'est pas le logarithme d'une vraisemblance classique de données observées et, de plus, $\mathbf{u}$ n'est pas un paramètre si bien qu'a priori toute cette manipulation paraît tout à fait illégitime sinon infondée; heureusement, comme on le verra à la fin, la maximisation de cette fonction trouve sa pleine justification non plus dans un cadre classique mais dans la théorie bayésienne.

Au préalable, et ce fut la démarche d'Henderson, on va montrer qu'on peut identifier les solutions $\tilde{\boldsymbol{\beta}}$ et $\tilde{\mathbf{u}}$ du système (2.22) à celle des moindres carrés généralisés d'une part, et au Blup, d'autre part.

La première équation s'écrit aussi: $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{Z}\tilde{\mathbf{u}})$. De même la deuxième, $(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})\tilde{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})$ et, en reportant l'expression de $\tilde{\mathbf{u}}$ de celle-ci dans celle-là, on obtient:

$$\mathbf{X}'\mathbf{W}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{W}\mathbf{y} , \qquad (2.24)$$

où

$$\mathbf{W} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}. \qquad (2.25)$$

On peut montrer (cf. annexe II) que:

$$\mathbf{W} = \left(\mathbf{Z}'\mathbf{G}\mathbf{Z} + \mathbf{R}\right)^{-1} = \mathbf{V}^{-1}, \qquad (2.26)$$

et donc $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

De la deuxième équation, on tire:

$$\tilde{\mathbf{u}} = \left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

et, en utilisant les résultats de l'annexe A, on montre alors que:

$$\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}$$

ce qui établit l'identité entre la solution $\tilde{\mathbf{u}}$ en $\mathbf{u}$ et le BLUP $\hat{\mathbf{u}}$.

23. Variances d'erreur

Il s'agit des variances d'échantillonnage des effets fixes et des variances d'erreur de prédiction des effets aléatoires. Leurs expressions ont été établies par Henderson (1975) dans un article de Biometrics.

Soit $\mathbf{C}$ une inverse de la matrice[4] des coefficients et qu'on peut partitionner comme suit:

$$\begin{bmatrix} \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \\ \mathbf{C}_{u\beta} & \mathbf{C}_{uu} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \qquad (2.27)$$

On montre que:

$$\mathrm{Var}(\mathbf{k}'\hat{\boldsymbol{\beta}}) = \mathbf{k}'\mathbf{C}_{\beta\beta}\mathbf{k}, \qquad (2.28a)$$

$$\mathrm{Cov}(\mathbf{k}'\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}') = \mathbf{0}, \qquad (2.28b)$$

$$\mathrm{Cov}\left[\mathbf{k}'\hat{\boldsymbol{\beta}}, (\hat{\mathbf{u}} - \mathbf{u})'\right] = \mathbf{k}'\mathbf{C}_{\beta u}, \qquad (2.28c)$$

$$\mathrm{Var}(\hat{\mathbf{u}}) = \mathbf{G} - \mathbf{C}_{uu}. \qquad (2.28d)$$

$$\mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{uu}, \qquad (2.28e)$$

La formule en a) découle directement de l'expression de l'inverse d'une matrice partitionnée en blocs. En effet $\mathbf{C}_{\beta\beta} = \left[\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\right]^{-1}$ soit, compte tenu de la propriété (2.26) se réduit à $\mathbf{C}_{\beta\beta} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$ QED.

[4] Celle-ci est supposée ici de plein rang pour simplifier la présentation, mais les résultats s'appliquent aussi à une inverse généralisée.

La propriété b) découle de l'orthogonalité des projecteurs $\mathbf{Q}$ et $\mathbf{I}-\mathbf{Q}$ (c.f. 1.21) comme suit. $\mathbf{k}'\boldsymbol{\beta}$ étant une fonction estimable peut s'exprimer comme une combinaison linéaire de l'espérance des observations soit $\mathbf{k}'\boldsymbol{\beta}=\boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta}$ d'où $\mathbf{k}'\hat{\boldsymbol{\beta}}=\boldsymbol{\lambda}'\mathbf{X}\hat{\boldsymbol{\beta}}=\boldsymbol{\lambda}'\mathbf{Q}\mathbf{y}$; par ailleurs comme on l'a montré précédemment (2.21) $\hat{\mathbf{u}}=\mathbf{C}'\underline{\mathbf{P}}\mathbf{y}$ d'où $\mathrm{Cov}(\mathbf{k}'\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}')=\boldsymbol{\lambda}'\mathbf{Q}\mathbf{V}\underline{\mathbf{P}}\mathbf{C}$; or $\mathbf{Q}\mathbf{V}\underline{\mathbf{P}}=\mathbf{Q}(\mathbf{I}-\mathbf{Q})=\mathbf{0}$.

Compte tenu de b), la relation c) est équivalente à $\mathrm{Cov}\left(\mathbf{k}'\hat{\boldsymbol{\beta}},\mathbf{u}'\right)=-\mathbf{k}'\mathbf{C}_{\beta u}$. Or $\mathbf{k}'\hat{\boldsymbol{\beta}}$ peut se mettre sous la forme: $\mathbf{k}'\hat{\boldsymbol{\beta}}=\mathbf{k}'\begin{pmatrix}\mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u}\end{pmatrix}\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{pmatrix}\mathbf{R}^{-1}\mathbf{y}$ de telle sorte que $\mathrm{Cov}\left(\mathbf{k}'\hat{\boldsymbol{\beta}},\mathbf{u}'\right)=\mathbf{k}'\begin{pmatrix}\mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u}\end{pmatrix}\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\end{pmatrix}\mathbf{G}$. Or, par définition de l'inverse:

$\begin{pmatrix}\mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u}\end{pmatrix}\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{pmatrix}=\mathbf{0}$, d'où $\mathrm{Cov}\left(\mathbf{k}'\hat{\boldsymbol{\beta}},\mathbf{u}'\right)=-\mathbf{k}'\mathbf{C}_{\beta u}\mathbf{G}^{-1}\mathbf{G}$,QED.

Pour établir la relation d), on peut utiliser la propriété du BLUP selon laquelle $\mathrm{Var}(\hat{\mathbf{u}})=\mathrm{Cov}(\hat{\mathbf{u}},\mathbf{u}')$ (c.f. 2.18); puis, on procédera selon la même méthode que précédemment en écrivant: $\hat{\mathbf{u}}=\begin{pmatrix}\mathbf{C}_{u\beta} & \mathbf{C}_{uu}\end{pmatrix}\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{pmatrix}\mathbf{R}^{-1}\mathbf{y}$ si bien que $\mathrm{Cov}(\hat{\mathbf{u}},\mathbf{u}')=\begin{pmatrix}\mathbf{C}_{u\beta} & \mathbf{C}_{uu}\end{pmatrix}\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\end{pmatrix}\mathbf{G}$ ou encore $\left(\mathbf{I}-\mathbf{C}_{uu}\mathbf{G}^{-1}\right)\mathbf{G}$, QED.

La relation e) sur la variance des erreurs de prédiction $\hat{\mathbf{u}}-\mathbf{u}$ découle de la relation précédente et du corollaire de (2.18) à savoir. $\mathrm{Var}(\mathbf{u})=\mathrm{Var}(\hat{\mathbf{u}})+\mathrm{Var}(\mathbf{u}-\hat{\mathbf{u}})$.

Enfin, s'agissant d'une combinaison linéaire $w=\mathbf{k}'\boldsymbol{\beta}+\mathbf{m}'\mathbf{u}$ quelconque d'effets fixes et d'effets aléatoires, sa variance d'erreur de prédiction s'obtient par :

$$\mathrm{Var}(\hat{w}-w)=\mathbf{k}'\mathbf{C}_{\beta\beta}\mathbf{k}+\mathbf{m}'\mathbf{C}_{uu}\mathbf{m}+2\mathbf{k}'\mathbf{C}_{\beta u}\mathbf{m}. \qquad (2.29)$$

24. Interprétation bayésienne

Les liens qui unissent le BLUP et la statistique bayésienne ont été soulignés depuis longtemps (Dempfle, 1977; Lefort,1980; Gianola et Fernando, 1986, Searle et al, 1992). Les fondements de l'analyse bayésienne du modèle linéaire ont été donnés par Lindley et Smith (1972) et c'est cette présentation que nous utiliserons ici comme au chapitre I (cf 2.2 «approche marginale de modèles hiérarchiques»). Rappelons brièvement qu'on considère ici un modèle gaussien avec échantillonnage en deux étapes suivantes:

1) $\mathbf{y}\,|\,\boldsymbol{\theta}\sim\mathcal{N}(\mathbf{T}\boldsymbol{\theta},\mathbf{R})$, (2.30a)

2) $\boldsymbol{\theta}\,|\,\boldsymbol{\alpha}\sim\mathcal{N}(\mathbf{W}\boldsymbol{\alpha},\boldsymbol{\Omega})$. (2.30b)

D'après le théorème de Bayes, $f(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto f(\mathbf{y}\,|\,\boldsymbol{\theta})f(\boldsymbol{\theta})$, et comme les densités a priori $f(\boldsymbol{\theta})$ et conditionnelle des observations $f(\mathbf{y}\,|\,\boldsymbol{\theta})$ sont conjuguées,

$$f(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto \exp(-Q/2), \tag{2.31}$$

avec

$$Q = (\mathbf{y}-\mathbf{T}\boldsymbol{\theta})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{T}\boldsymbol{\theta}) + (\boldsymbol{\theta}-\mathbf{W}\boldsymbol{\alpha})'\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}-\mathbf{W}\boldsymbol{\alpha})$$

On montre que Q peut se mettre sous la forme alternative suivante:

$$\begin{aligned} Q = {} & (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T}+\boldsymbol{\Omega}^{-1})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} \\ & -\hat{\boldsymbol{\theta}}'(\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}+\boldsymbol{\Omega}^{-1}\mathbf{W}\boldsymbol{\alpha}) + \boldsymbol{\alpha}'\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{W}\boldsymbol{\alpha} \end{aligned}, \tag{2.32}$$

où $\hat{\boldsymbol{\theta}}$ est solution du système

$$\boxed{(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T}+\boldsymbol{\Omega}^{-1})\hat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{R}^{-1}\mathbf{y}+\boldsymbol{\Omega}^{-1}\mathbf{W}\boldsymbol{\alpha}}. \tag{2.33}$$

Seul le premier terme de (2.32) concoure à l'expression du noyau de la densité a posteriori qui est donc

$$\boldsymbol{\theta}\,|\,\mathbf{y} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \mathbf{C}). \tag{2.34}$$

où

$$\mathbf{C} = (\mathbf{T}'\mathbf{R}^{-1}\mathbf{T}+\boldsymbol{\Omega}^{-1})^{-1}$$

Posons $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}')'$, $\mathbf{T} = (\mathbf{X}, \mathbf{Z})$, $f(\boldsymbol{\theta}) = f(\boldsymbol{\beta})f(\mathbf{u})$ avec $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{B})$ et $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$. On va différencier le statut de $\boldsymbol{\beta}$ et de $\mathbf{u}$ en postulant une information a priori uniforme sur $\boldsymbol{\beta}$ qu'on peut assimiler à un cas limite de la spécification précédente pour $\mathbf{B} \to \infty$; dans ce cas,

$\boldsymbol{\Omega}^{-1} \to \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{pmatrix}$, $\boldsymbol{\Omega}^{-1}\mathbf{W}\boldsymbol{\alpha} \to \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}_0 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$ si bien que: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}')'$ est solution du système des équations du modèle mixte.

De plus, on a compte tenu de (2.34):

$$\boldsymbol{\beta}\,|\,\mathbf{y}, \mathbf{G}, \mathbf{R} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{C}_{\beta\beta}), \tag{2.35}$$

$$\mathbf{u}\,|\,\mathbf{y}, \mathbf{G}, \mathbf{R} \sim \mathcal{N}(\hat{\mathbf{u}}, \mathbf{C}_{uu}) \ . \tag{2.36}$$

propriétés qui conduisent à une interprétation plus riche des solutions des équations du modèle mixte; en particulier l'estimateur GLS $\hat{\boldsymbol{\beta}}$ des effets fixes s'interprète dans cette formulation comme l'espérance a posteriori de $\boldsymbol{\beta}$ sachant $\mathbf{G}$ et $\mathbf{R}$ avec une information a priori uniforme sur $\boldsymbol{\beta}$. Cela permet au passage d'illustrer la différence de présentation des

propriétés de l'inférence sur les effets fixes en statistique bayésienne $\boldsymbol{\beta} \mid \mathbf{y} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \mathbf{C}_{\beta\beta}\right)$ par rapport à la présentation classique $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \mathbf{C}_{\beta\beta}\right)$. De la même façon, le BLUP de $\mathbf{u}$ s'interprète comme l'espérance de la distribution a posteriori de $\mathbf{u}$ sachant $\mathbf{G}$ et $\mathbf{R}$ et la variance de cette distribution équivaut alors à la variance des erreurs de prédiction sous l'hypothèse de normalité:

$$\mathrm{Var}(\mathbf{u}|\mathbf{y}, \mathbf{G}, \mathbf{R}) = \mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) \qquad (2.37)$$

Le lien apparaît maintenant clairement avec la justification première donnée par Henderson de ses équations. En effet, considérons la densité conjointe de $(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta})$. Sous l'hypothèse d'une distribution uniforme de $\boldsymbol{\beta}$, $\mathrm{f}(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta})$ est proportionnelle à $\mathrm{f}(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta})$:

$$\mathrm{f}(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta}) = \mathrm{f}(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta})\mathrm{f}(\boldsymbol{\beta}) \propto \mathrm{f}(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta}), \qquad (2.38)$$

qui est la densité qu'Henderson maximisait par rapport à $\boldsymbol{\beta}$ et $\mathbf{u}$. D'un autre côté par application du thèorème de Bayes, on trouve que $\mathrm{f}(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta})$ est proportionnelle à $\mathrm{f}(\boldsymbol{\beta}, \mathbf{u}|\mathbf{y})$

$$\mathrm{f}(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta}) = \mathrm{f}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta})\mathrm{f}(\mathbf{u}, \boldsymbol{\beta}) \propto \mathrm{f}(\boldsymbol{\beta}, \mathbf{u}|\mathbf{y}), \qquad (2.39)$$

Donc maximiser le logarithme de (2.38) par rapport à $\boldsymbol{\beta}$ et $\mathbf{u}$ équivaut à chercher le mode de $\mathrm{f}(\boldsymbol{\beta}, \mathbf{u}|\mathbf{y})$ en (2.39). Or, sous l'hypothèse de normalité, l'espérance et le mode de la densité a posteriori de $(\boldsymbol{\beta}, \mathbf{u})$ sont confondus et égaux à la solution GLS $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ et au BLUP $\hat{\mathbf{u}}$ de $\mathbf{u}$ basé sur les observations $\mathbf{y}$.

3. Conclusion

Au terme de ce chapitre, nous avons défini un cadre conceptuel rigoureux pour aborder le problème de la prédiction. Celle-ci se décline suivant différents vocables selon les hypothèses faites sur la distribution conjointe de la variable à prédire (W) et de la variable prédictrice (Y).

L'espérance de la distribution conditionnelle joue un rôle clé dans la prédiction avec le critère d'erreur quadratique moyenne. Dans ce cadre, la théorie permet des développements simples quand on se place sous l'hypothèse de normalité ou dans le cadre de prédicteurs linéaires.

Dans le cas où les moments de premier ordre ne sont pas connus et peuvent se formaliser dans le cadre d'un modèle linéaire mixte, on aboutit dans la classe des prédicteurs linéaires sans biais à un prédicteur aux propriétés remarquables et qui, à la suite de Goldberger et d'Henderson, est dénommé BLUP. Son intérêt est d'autant plus grand qu'on

peut l'obtenir simplement à partir d'un système d'équations dites du modèle mixte d'Henderson qui est proche de celui des moindres carrés et dont la justification apparaît tout naturellement dans un cadre bayésien.

Cela explique pourquoi le BLUP et les équations du modèle mixte ont une portée qui dépasse largement le cadre des applications de la génétique et la sélection animale pour lesquelles ces outils ont été conçus au départ par Henderson, puis développés avec grand succès par ses élèves. Ils sont aussi au cœur des méthodes d'estimation des composantes de la variance et des algorithmes correspondants tel l'algorithme EM. Les équations d'Henderson constituent donc un outil incontournable dans la traitement général -qu'il soit classique ou bayésien- des modèles mixtes linéaires et même non linéaires.

ANNEXE II-A

INVERSE DE V

Considérons la partition suivante en blocs d'une matrice carrée non singulière et de son inverse :

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} \qquad \text{(II-A.1)}$$

où $\mathbf{A}_{11}$ et $\mathbf{A}_{22}$ sont deux matrices carrées supposées non singulières.

La démonstration proposée repose sur les deux résultats suivants:

$$\mathbf{A}^{11} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}, \qquad \text{(II-A.2)}$$

$$\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22} = -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}. \qquad \text{(II-A.3)}$$

Idem pour $\mathbf{A}^{22}$ et $\mathbf{A}^{21}$.

Posons maintenant:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{R}^{-1} & \mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}. \qquad \text{(II-A.4)}$$

En appliquant la 1ère partie de (II-A.2) à $\mathbf{A}^{22}$, il vient:

$$\mathbf{A}^{22} = \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} = \left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\right)^{-1} = \mathbf{G}$$

De même pour $\mathbf{A}^{11}$, on a:

$$\mathbf{A}^{11} = \left[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\right]^{-1} = \mathbf{W}^{-1} \qquad \text{(II-A.5)}$$

puis la 2ème partie de (II-A.2) donne:

$$\mathbf{A}^{11} = \mathbf{R} + \mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{R} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}' = \mathbf{V}, \text{ QED.} \qquad \text{(II-A.6)}$$

L'application de (II-A.3) à la matrice définie en (2-I.4) conduit à:

$$\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22} = -\mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\mathbf{G} = -\mathbf{Z}\mathbf{G}$$

$$\mathbf{A}^{12} = -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} = -\mathbf{V}\mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}$$

soit à l'égalité:

$$\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} = \left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}, \qquad \text{(II-A.7)}$$

établissant l'équivalence entre le BLUP $\hat{\mathbf{u}}$ et la solution $\tilde{\mathbf{u}}$ des équations d'Henderson.

Chapitre III. METHODES DU MAXIMUM DE VRAISEMBLANCE

Introduction

1. Méthode dite ML
    - 11. Fonction de vraisemblance
    - 12. Maximisation
        - 121. Dérivées premières
        - 122. Cas général
        - 123. Cas du modèle mixte
    - 13. Variantes
        - 131. Vraisemblance profilée
        - 132. Forme de Hartley-Rao
    - 14. Aspects calculatoires
        - 141. Algorithme d'Henderson
        - 142. Calcul de –2L
    - 15. Tests d'hypothèses
        - 151. Loi asymptotique
        - 152. Statistiques de Wald
        - 153. Statistique du rapport de vraisemblance
        - 154. Statistique du score
        - 155. Discussion
2. Méthode dite REML
    - 21. Exemple simple
        - 211. Estimateur
        - 212. Correction du biais
    - 22. Cas général
        - 221. Concept de vraisemblance marginale
        - 222. Application au modèle linéaire mixte gaussien
        - 223. Interprétation bayésienne
    - 23. Aspects calculatoires
        - 231. Algorithmes «type-Henderson» et d'Harville
        - 232. Calcul de –2RL
    - 24. Vraisemblance résiduelle et tests
        - 241. Approximation de Kenward et Roger
        - 242. Approche de Welham et Thompson
        - 243. Tests des effets aléatoires

Discussion-Conclusion

Introduction

Le maximum de vraisemblance est une méthode générale d'estimation due à Fisher (1922,1925) qui possède des propriétés statistiques intéressantes surtout dans les conditions asymptotiques (Cox et Hinkley, 1974). Dans le cas de la variance, cette méthode a été utilisée par Crump (1947) dans des situations simples (modèle à une voie, dispositifs équilibrés). Mais ce sont Hartley et Rao (1967) qui, les premiers, en donnèrent un formalisme général dans le cadre du modèle linéaire mixte gaussien (cf la revue historique de Searle, 1989). L'article de Hartley et Rao marque la rupture avec les estimateurs quadratiques. Ceux-ci s'inspiraient de l'analyse de variance qui fut la technique reine imprégnant fortement tout le secteur de l'estimation des composantes de la variance depuis les travaux originaux de Fisher sur le coefficient de corrélation intra classe jusqu'aux méthodes d'Henderson (1953) dites I, II et III basées sur les idées de Yates (1934).

Avec Rao (1971ab), le choix des formes quadratiques quitta l'univers de l'ANOVA et des moindres carrés pour se rationaliser autour de propriétés d'optimalité. En fait, cette classe d'estimateurs quadratiques sans biais et localement de norme minimum (dits MINQUE) (LaMotte, 1973) n'apparaît plus aujourd'hui que comme une transition entre la période d'Henderson et celle du maximum de vraisemblance puisque le MINQUE aboutit naturellement sous sa forme itérée à un estimateur du maximum de vraisemblance.

On distingue à cet égard deux approches. La première, dite en abrégé ML, utilise le concept classique de fonction de vraisemblance de l'ensemble des paramètres (position et dispersion). L'autre méthode, dite REML, fut introduite par Anderson et Bancroft (1952) et Thompson (1962) dans l'analyse de dispositifs équilibrés puis généralisée à un modèle mixte gaussien quelconque par Patterson et Thompson (1971). Cette méthode considère la vraisemblance d'une fonction des observations, libre des effets fixes –«contrastes d'erreur» dans la terminologie d'Harville (1977)- d'où son appellation de vraisemblance restreinte ou résiduelle (acronyme anglais REML). Cette vraisemblance résiduelle possède par ailleurs une interprétation bayésienne (Harville, 1974) en terme de vraisemblance marginalisée par intégration des effets fixes selon une distribution uniforme.

Les techniques du maximum de vraisemblance ont suscité beaucoup d'intérêt en biostatistiques depuis le début de la décennie 80. La raison principale de l'essor de ces méthodes en est la faisabilité numérique grâce au développement simultané des ordinateurs, d'algorithmes performants (algorithmes dits EM «Expectation Maximisation» ou AI

«Average Information» par exemple) et de logiciels faciles d'accès et d'utilisation (SAS, ASREML, Splus). L'objet de cet article est de faire le point sur ces deux techniques d'inférence dans une optique à la fois pédagogique et opérationnelle.

1. Méthode dite ML

11. Fonction de vraisemblance

Nous nous placerons tout d'abord dans le cadre du modèle linéaire gaussien exprimé sous sa forme la plus générale:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_{\gamma}) \qquad (3.1)$$

où $\mathbf{X}$ est la matrice $(N \text{ x } p)$ des p variables explicatives relatives aux N éléments du vecteur $\mathbf{y}$ des observations et $\boldsymbol{\beta}$, le vecteur $(p \text{ x } 1)$ des coefficients de ces variables ou effets fixes. $\mathbf{V}_{\gamma}$ est la matrice $(N \text{ x } N)$ de variance-covariance des observations (notée en abrégé $\mathbf{V}$) supposée symétrique, définie-positive, dépendant d'un vecteur $\boldsymbol{\gamma} \in \Gamma$ de paramètres et dont la structure caractéristique est, dans le cas des modèles linéaires mixtes, $\mathbf{V} = \sum_{k=0}^{K} \mathbf{V}_k \gamma_k$ où les $\mathbf{V}_k$ sont des matrices réelles connues.

La densité des observations $\mathbf{y}$ s'écrit:

$$\mathrm{p}_Y(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma}) = (2\pi)^{-N/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right], \qquad (3.2),$$

d'où le logarithme de la vraisemblance $\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) = \ln \mathrm{p}_Y(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})$ (dite logvraisemblance) considéré ici comme une fonction des paramètres $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$ (Edwards, 1972):

$$\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \qquad (3.3a)$$

ou, sous sa forme « –2L »

$$\boxed{-2\,\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) = N\ln(2\pi) + \ln|\mathbf{V}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \qquad (3.3b)$$

12. Maximisation

121. Dérivées premières

Rappelons que la recherche des points $\boldsymbol{\alpha} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ qui maximisent $\mathrm{L}(\boldsymbol{\alpha};\mathbf{y})$ (ou minimisent $-2\,\mathrm{L}(\boldsymbol{\alpha};\mathbf{y})$) soit

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in A = \mathbb{R}^p \text{ x } \Gamma} \mathrm{L}(\boldsymbol{\alpha};\mathbf{y}) \qquad (3.4)$$

se fait habituellement en annulant les dérivées premières:

$$\frac{\partial \mathrm{L}(\boldsymbol{\alpha};\mathbf{y})}{\partial \boldsymbol{\alpha}} = \mathbf{0} \tag{3.5}$$

Une telle démarche ne doit pas être abordée sans prudence. Il importe, en effet, de bien vérifier 1) que les points ainsi obtenus appartiennent à l'espace paramétrique, et 2) que les dérivées secondes en ces points $\frac{\partial^2 \mathrm{L}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} < \mathbf{0}$ forment une matrice définie-négative. Si la condition $\boldsymbol{\beta} \in \mathbb{R}^p$ ne pose aucune difficulté, par contre l'espace paramétrique Γ de $\boldsymbol{\gamma}$ doit être soigneusement précisé en fonction du modèle adopté. La restriction minimale découle de la condition $\mathbf{V} > 0$ (définie-positive) mais, dans la plupart des cas, la définition de l'espace paramétrique Γ imposera des restrictions supplémentaires. Par exemple, dans un modèle linéaire mixte unidimensionnel à K facteurs aléatoires indépendants plus une résiduelle tel que $\mathbf{V} = \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{Z}_k^{'} \sigma_k^2$, on aura $\Gamma = \left\{ \sigma_0^2 > 0;\ \sigma_k^2 \geq 0,\ \forall k = 1,...,K \right\}$.

La propriété de négativité de la matrice des dérivées secondes aux points annulant les dérivées premières conditionnent l'existence d'un maximum mais qui n'est pas nécessairement global. Il est peut être difficile -du moins fastidieux- de répertorier tous les maxima locaux et d'évaluer la vraisemblance en ces points ainsi qu'en bordure de l'espace paramétrique. Cela nécessite alors le recours à des techniques de maximisation sous contraintes (cf annexe I). Les choses se simplifient beaucoup lorsque L est une fonction concave du paramètre (ou d'un transformé bijectif) puisque alors les conditions de premier ordre garantissent l'existence d'un maximum global.

Les dérivées premières s'écrivent:

$$\frac{\partial(-2\,\mathrm{L})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{3.6}$$

$$\frac{\partial(-2\,\mathrm{L})}{\partial \gamma_k} = \frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.7}$$

Or, d'après des résultats généraux (cf par exemple Searle, 1982, pages 335-337 ; Harville (1997, pages 305-308)

$$\frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \gamma_k} \right) \tag{3.8}$$

$$\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k} = -\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \gamma_k} \mathbf{V}^{-1}. \tag{3.9}$$

d'où

$$\frac{\partial(-2\,\mathrm{L})}{\partial\gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\gamma_k}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\gamma_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,. \tag{3.10}$$

L'annulation des dérivées premières en (3.6) et (3.7) conduit au système suivant:

$$\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}\,, \tag{3.11a}$$

$$\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\gamma_k}\right)_{\mathbf{V}=\hat{\mathbf{V}}} - (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}\left.\frac{\partial\mathbf{V}}{\partial\gamma_k}\right|_{\mathbf{V}=\hat{\mathbf{V}}}\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = 0\,. \tag{3.11b}$$

où $\hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{V}}$ solutions de ce système (quand elles existent) désignent les estimations du maximum de vraisemblance (ML).

Quelques simplifications sont possibles. Tout d'abord, on élimine $\hat{\boldsymbol{\beta}}$ de (11b) en reportant son expression $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ de (3.11a) dans (3.11b) et en remarquant que: $\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{V}}^{-1}\left(\mathbf{I}-\hat{\mathbf{Q}}\right)\mathbf{y} = \underline{\hat{\mathbf{P}}}\mathbf{y}$ où $\underline{\hat{\mathbf{P}}}$ représente la notation abrégée de la valeur de la matrice

$$\mathbf{P} = \mathbf{V}^{-1}\left(\mathbf{I}-\mathbf{Q}\right) = \mathbf{V}^{-1} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\,, \tag{3.12}$$

(Searle,1979) évaluée au point $\mathbf{V} = \hat{\mathbf{V}}$, $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ représentant le projecteur des moindres carrés généralisés.

122. Cas général

Le système en (3.11ab) ainsi obtenu n'est pas soluble plus avant et l'on a recours à un algorithme du second ordre tel que l'algorithme de Newton-Raphson ou celui des scores de Fisher qui implique le calcul respectivement du hessien $\ddot{\mathbf{L}}(\boldsymbol{\alpha};\mathbf{y}) = \partial^2\,\mathrm{L}(\boldsymbol{\alpha};\mathbf{y})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'$ et de la matrice d'information $\mathbf{J}(\boldsymbol{\alpha}) = \mathrm{E}_{Y|\boldsymbol{\alpha}}\left[-\ddot{\mathbf{L}}(\boldsymbol{\alpha};\mathbf{y})\right]$ (cf annexe II), soit , pour cette dernière

$$\boxed{\mathbf{J}(\boldsymbol{\alpha}) = \begin{bmatrix}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}/2\end{bmatrix}}\,. \tag{3.13}$$

où

$$\left(\mathbf{F}\right)_{kl} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\gamma_k}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\gamma_l}\right). \tag{3.14}$$

En ce qui concerne $\boldsymbol{\gamma}$, on résout itérativement le système suivant:

$$\mathbf{J}(\boldsymbol{\gamma}^{[n]})\boldsymbol{\Delta}^{[n+1]} = \dot{\mathbf{L}}(\boldsymbol{\gamma}^{[n]}) \tag{3.15}$$

où

$$\boldsymbol{\Delta}^{[n+1]} = \boldsymbol{\gamma}^{[n+1]} - \boldsymbol{\gamma}^{[n]} \;;\; \dot{\mathbf{L}}(\boldsymbol{\gamma}) = \partial \mathrm{L}(\boldsymbol{\alpha};\mathbf{y}) / \partial \boldsymbol{\gamma} \,,$$

$$\dot{\mathbf{L}}(\boldsymbol{\gamma}^{[n]}) = \left\{ -\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k} \right) + \frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\mathbf{y} \right\}\Bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{[n]}}, \tag{3.16}$$

$$\mathbf{J}(\boldsymbol{\gamma}^{[n]}) = 1/2\mathbf{F}\left(\boldsymbol{\gamma}^{[n]}\right). \tag{3.17}$$

L'estimation $\hat{\boldsymbol{\gamma}}$ étant obtenue, on en déduit $\hat{\boldsymbol{\beta}}$ par résolution de (3.11a) qui est alors linéaire en $\boldsymbol{\beta}$. Si $\mathbf{V}$ est connu, l'estimateur des moindres carrés généralisés (dite GLS en anglais) est solution du système $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. On retrouve ici pour le ML de $\boldsymbol{\beta}$ une forme similaire dans laquelle $\mathbf{V}$ est remplacé par son estimation ML, $\hat{\mathbf{V}}$.

123. Cas du modèle mixte.

Alors $\mathbf{V} = \sum_{l=0}^{K} \mathbf{V}_l \gamma_l$, $\partial \mathbf{V} / \partial \gamma_k = \mathbf{V}_k$ où $\mathbf{V}_k$ est une matrice $(N \,\mathrm{x}\, N)$ connue, par exemple: $\mathbf{V}_k = \mathbf{Z}_k \mathbf{Z}_k^{'}$ (cf 1.29) et l'équation (3.11b) devient

$$\mathrm{tr}\left(\hat{\mathbf{V}}^{-1}\mathbf{V}_k\right) - \mathbf{y}'\hat{\mathbf{P}}\mathbf{V}_k\hat{\mathbf{P}}\mathbf{y} = 0 \,. \tag{3.18}$$

Du fait de la linéarité de $\mathbf{V}$, on peut expliciter le terme de gauche de (3.18) en

$$\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{V}_k\right) = \sum_{l=0}^{K} \mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{V}_k\mathbf{V}^{-1}\mathbf{V}_l\right)\gamma_l \,.$$

Le système en (3.18) s'écrit alors

$$\boxed{\sum_{l=0}^{K} \mathrm{tr}\left(\hat{\mathbf{V}}^{-1}\mathbf{V}_k\hat{\mathbf{V}}^{-1}\mathbf{V}_l\right)\hat{\gamma}_l = \mathbf{y}'\hat{\underline{\mathbf{P}}}\mathbf{V}_k\hat{\underline{\mathbf{P}}}\mathbf{y}}, (k = 0,1,...,K) \tag{3.19a}$$

soit encore, sous forme matricielle :

$$\boxed{\hat{\mathbf{F}}\hat{\boldsymbol{\gamma}} = \hat{\mathbf{g}}}, \tag{3.19b}$$

où $\mathbf{F}$ est une matrice $\left(K+1\right)\mathrm{x}\left(K+1\right)$ symétrique et $\mathbf{g}$ un vecteur $\left(K+1\right)$ définis par

$$\mathbf{F} = \left\{f_{kl}\right\} = \left\{\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{V}_k\mathbf{V}^{-1}\mathbf{V}_l\right)\right\}, \tag{3.20a}$$

$$\mathbf{g} = \left\{g_k\right\} = \left\{\mathbf{y}'\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{y}\right\}, \tag{3.20b}$$

$\hat{\mathbf{F}}$ et $\hat{\mathbf{g}}$ correspondant à $\mathbf{F}$ et $\mathbf{g}$ évalués au point $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$.

Le système en (3.19ab) est un système non linéaire qui, en général, n'a pas de solution analytique; on le résout numériquement par un algorithme itératif ayant la forme d'un système linéaire en $\boldsymbol{\gamma}$:

$$\boxed{\mathbf{F}\left(\boldsymbol{\gamma}^{[n]}\right)\boldsymbol{\gamma}^{[n+1]} = \mathbf{g}\left(\boldsymbol{\gamma}^{[n]}\right)}, \tag{3.21}$$

où $\boldsymbol{\gamma}^{[n]}$ est la valeur courante du paramètre à l'itération n à partir de laquelle on évalue la matrice des coefficients $\mathbf{F}$ et le second membre $\mathbf{g}$; puis on résout le sytème ainsi obtenu en $\boldsymbol{\gamma}$ de façon à obtenir la valeur du paramètre à l'itération suivante.

On montre aisément que le système (3.21) équivaut à celui des équations des scores de Fisher (3.15) au coefficient ½ près.

Lorsque $\mathbf{V}_k = \mathbf{Z}_k \mathbf{Z}_k'$, le calcul des éléments de $\mathbf{F}$ et de $\mathbf{g}$ en (3.20ab) et (3.21) peut être à son tour grandement simplifié en tirant avantage du fait que la trace du produit d'une matrice et de sa transposée est égale à la somme des carrés des élements de la matrice, ie $\text{tr}(\mathbf{AA}') = \sum_{ij} a_{ij}^2$.

Ainsi, $f_{kl} = \sum_{ij} \left(\mathbf{Z}_k' \mathbf{V}^{-1} \mathbf{Z}_l \right)_{ij}^2$ et $g_k = \sum_i \left(\mathbf{Z}_k' \underline{\mathbf{P}} \mathbf{y} \right)_i^2$.

13. Variantes

131. Vraisemblance profilée

L'idée à la base de la vraisemblance profilée est de maximiser la vraisemblance par étapes successives. On va d'abord maximiser $\text{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$ par rapport à $\boldsymbol{\beta}$, puis la fonction ainsi obtenue $\text{L}_P(\boldsymbol{\gamma}; \mathbf{y}) = \text{L}\left(\hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}; \mathbf{y} \right)$ (du seul paramètre $\boldsymbol{\gamma}$) dite vraisemblance profilée (Cox et Reid, 1987) ou concentrée (Harville et Callanan, 1990) par rapport à $\boldsymbol{\gamma}$. En bref

$$\begin{aligned} \text{Max}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \, \text{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) &= \text{Max}_{\boldsymbol{\gamma}} \left[\text{Max}_{\boldsymbol{\beta}} \, \text{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) \right] \\ &= \text{Max}_{\boldsymbol{\gamma}} \, \text{L}\left(\hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}; \mathbf{y} \right) \\ &= \text{Max}_{\boldsymbol{\gamma}} \, \text{L}_P(\boldsymbol{\gamma}; \mathbf{y}) \end{aligned} \quad , \qquad (3.22)$$

où $\hat{\boldsymbol{\beta}}_\gamma = \left(\mathbf{X}' \mathbf{V}_\gamma^{-1} \mathbf{X} \right)^- \mathbf{X}' \mathbf{V}_\gamma^{-1} \mathbf{y}$ est solution GLS de $\boldsymbol{\beta}$.

Compte tenu de (3.7b), il vient immédiatement :

$$-2\,\text{L}_P(\boldsymbol{\gamma}; \mathbf{y}) = N \ln(2\pi) + \ln\left| \mathbf{V}_\gamma \right| + \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\gamma \right)' \mathbf{V}_\gamma^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\gamma \right)$$

ou encore, en ignorant l'indiçage par $\boldsymbol{\gamma}$ dans $\mathbf{V}$:

$$\boxed{-2\,\text{L}_P(\boldsymbol{\gamma}; \mathbf{y}) = N \ln(2\pi) + \ln\left| \mathbf{V} \right| + \mathbf{y}' \underline{\mathbf{P}} \mathbf{y}} \, . \qquad (3.23)$$

Sachant que $\dfrac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\underline{\mathbf{P}} \dfrac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}}$ (cf annexe II), on en déduit facilement l'expression du gradient :

$$\frac{\partial \left[-2\,\text{L}_P(\boldsymbol{\gamma}; \mathbf{y}) \right]}{\partial \gamma_k} = \text{tr}\left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \gamma_k} \right) - \mathbf{y}' \underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}} \mathbf{y} \qquad (3.24)$$

qui coïncide bien (au coefficient ½ près) avec (3.11b).

Deux remarques méritent d'être faites à ce stade: 1) la vraisemblance profilée permet de réduire la dimensionnalité du problème en «concentrant» la fonction de logvraisemblance sur le paramètre d'intérêt après avoir éliminé le paramètre parasite; 2) toutefois, la fonction ainsi obtenue n'est pas à proprement parler –en dépit de son appellation- une fonction de logvraisemblance même si, à l'occasion, elle conserve certaines de ses propriétés (Berger et al, 1999).

132. Formulation de Hartley-Rao

Hartley et Rao (1967) se placent dans le cadre du modèle linéaire mixte gaussien usuel décrit en (1.28) et (1.29).

Au lieu de paramétrer $\mathbf{V}$ en terme de variances $\boldsymbol{\sigma}^2 = \left\{\sigma_k^2\right\}$, Hartley et Rao isolent la variance résiduelle σ_0^2 et introduisent le vecteur $\boldsymbol{\eta}_{K\,\mathrm{x}\,1} = \left\{\eta_k = \sigma_k^2 / \sigma_0^2\right\}$ des rapports de variance. Pour ce faire, ils posent $\mathbf{V} = \mathbf{H}\sigma_0^2$ où $\mathbf{H} = \mathbf{I}_N + \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{Z}_k^{'} \eta_k$ est fonction du seul vecteur $\boldsymbol{\eta}$. Comme $\left|\mathbf{V}\right| = \left|\mathbf{H}\right| \sigma_0^{2N}$, la logvraisemblance s'écrit :

$$\begin{aligned} -2\,\mathrm{L}\left(\boldsymbol{\beta}, \sigma_0^2, \boldsymbol{\eta}; \mathbf{y}\right) = N \ln\left(2\pi\right) + \ln\left|\mathbf{H}\right| + N \ln \sigma_0^2 \\ + \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \mathbf{H}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) / \sigma_0^2 \end{aligned} . \tag{3.25}$$

On calcule ensuite les dérivées partielles de $-2\,\mathrm{L}\left(\boldsymbol{\beta}, \sigma_0^2, \boldsymbol{\eta}; \mathbf{y}\right)$ par rapport aux paramètres soit :

$$\frac{\partial(-2\,\mathrm{L})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) / \sigma_0^2 , \tag{3.26a}$$

$$\frac{\partial(-2\,\mathrm{L})}{\partial \sigma_0^2} = \frac{N}{\sigma_0^2} - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma_0^4} , \tag{3.26b}$$

$$\frac{\partial(-2\,\mathrm{L})}{\partial \eta_k} = \mathrm{tr}\left(\mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial \eta_k} \right) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial \eta_k} \mathbf{H}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) / \sigma_0^2 . \tag{3.26c}$$

Par annulation de ces dérivées, on obtient immédiatement :

$$\mathbf{X}'\hat{\mathbf{H}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\hat{\mathbf{H}}^{-1}\mathbf{y} , \tag{3.27a}$$

$$\hat{\sigma}_0^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{H}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) / N , \tag{3.27b}$$

$$\mathrm{tr}\left(\hat{\mathbf{H}}^{-1}\mathbf{H}_k\right) - (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{H}}^{-1}\mathbf{H}_k\hat{\mathbf{H}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) / \hat{\sigma}_0^2 = 0 , \tag{3.27c}$$

où $\mathbf{H}_k = \partial \mathbf{H} / \partial \eta_k = \mathbf{Z}_k \mathbf{Z}_k^{'}$.

On retrouve en (3.27a) le même résultat que celui obtenu avec l'estimateur GLS dont l'expression ne dépend pas explicitement de la variance résiduelle. La formulation de Hartley-

Rao permet l'obtention directe d'un estimateur ML de cette variance dont Henderson (1973) a donné un algorithme de calcul très simple faisant intervenir les éléments des équations du modèle mixte. Comme précédemment (cf (26a)), on peut remplacer (34c) par une équation plus accessible. Sachant que, $\mathbf{H} = \mathbf{I}_N + \sum_{l=1}^{K} \mathbf{H}_l \eta_l$, on a:

$\mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{H}_k\right) = \sum_{l=1}^{K} \mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}\mathbf{H}_l\right)\eta_l + \mathrm{tr}\left(\mathbf{H}^{-2}\mathbf{H}_k\right)$, d'où le système linéaire itératif suivant :

$$\begin{aligned} &\sum_{l=1}^{K} \mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}\mathbf{H}_l\right)\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^{[n]}} \eta_l^{[n+1]} = \\ &\left[(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}_0^2 - \mathrm{tr}\left(\mathbf{H}^{-2}\mathbf{H}_k\right)\right]\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^{[n]}} \end{aligned} \qquad (3.28)$$

pour $k = 1, 2, ..., K$.

La même remarque qu'en (3.21) s'applique ici quant à la simplification des calculs des éléments des traces intervenant en (3.28).

14. Aspects calculatoires

141. Algorithme d'Henderson

Henderson (1973) se place également dans le cadre du modèle linéaire mixte précédent et considère la dérivée de $-2\,\mathrm{L}_P$ par rapport à σ_k^2 (cf (3.18)) qui s'écrit :

$$\partial\left(-2\,\mathrm{L}_P\right)/\partial\sigma_k^2 = \mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_k\mathbf{Z}_k'\right) - \mathbf{y}'\underline{\mathbf{P}}\mathbf{Z}_k\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{y}\,.$$

Or, le meilleur prédicteur linéaire sans biais (acronyme BLUP en anglais) $\hat{\mathbf{u}}_k$ de $\mathbf{u}_k$ s'écrit par définition: $\hat{\mathbf{u}}_k = \mathrm{Cov}\left(\mathbf{u}_k, \mathbf{y}'\right)\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)$, soit $\hat{\mathbf{u}}_k = \sigma_k^2\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{y}$ d'où une façon d'exprimer la forme quadratique $\mathbf{y}'\underline{\mathbf{P}}\mathbf{Z}_k\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{y}$ sous la forme équivalente: $\hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k/\sigma_k^4$.

De même, Henderson montre que: $\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_k\mathbf{Z}_k'\right) = \dfrac{q_k}{\sigma_k^2} - \dfrac{\mathrm{tr}\left(\underline{\mathbf{C}}_{kk}\right)\sigma_0^2}{\sigma_k^4}$ où, en posant $\mathbf{Z} = \left(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k, ..., \mathbf{Z}_K\right)$, $\underline{\mathbf{C}}_{kk} = \left[\left(\mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1}\right)^{-1}\right]_{kk}$ est le bloc relatif au facteur k de taille $(q_k \,\mathrm{x}\, q_k)$ dans l'inverse de la partie relative aux effets aléatoires (ici $\mathbf{G} = \bigoplus_{k=1}^{K}\sigma_k^2\mathbf{I}_{q_k}$) de la matrice des coefficients des équations dites du modèle mixte. L'annulation de la dérivée conduit à :

$$q_k\hat{\sigma}_k^2 = \hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k + \mathrm{tr}\left(\hat{\underline{\mathbf{C}}}_{kk}\right)\hat{\sigma}_0^2 \qquad (3.29)$$

Pour la variance résiduelle σ_0^2, le raisonnement s'appuie sur la vraisemblance profilée $-2\,\mathrm{L}_p(\boldsymbol{\eta};\mathbf{y}) = -2\,\mathrm{L}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}), \hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta};\mathbf{y}\right]$ relative à la formulation d'Hartley-Rao, soit

$$-2\,\mathrm{L}_p(\boldsymbol{\eta};\mathbf{y}) = N(\ln 2\pi + 1) + \ln|\mathbf{H}| + N\ln\hat{\sigma}_0^2(\boldsymbol{\eta}).$$

où

$$\hat{\sigma}_0^2(\boldsymbol{\eta}) = \left[\mathbf{y} - \hat{\boldsymbol{\beta}}(\boldsymbol{\eta})\right]'\mathbf{H}^{-1}\left[\mathbf{y} - \hat{\boldsymbol{\beta}}(\boldsymbol{\eta})\right]/N$$

avec $\hat{\boldsymbol{\beta}}(\boldsymbol{\eta})$ solution de $\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}) = \mathbf{X}'\mathbf{H}^{-1}\mathbf{y}$.

Sachant que le BLUP $\hat{\mathbf{e}}$ de $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}$ s'écrit $\hat{\mathbf{e}} = \mathbf{R}\underline{\mathbf{P}}\mathbf{y}$, (ici $\mathbf{R} = \mathbf{I}_N\sigma_0^2$), on en déduit une forme équivalente à cette dernière expression:

$$\hat{\sigma}_0^2(\boldsymbol{\eta}) = \left[\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\boldsymbol{\eta})\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'(\boldsymbol{\eta})\mathbf{Z}'\mathbf{y}\right]/N \qquad (3.30)$$

Henderson propose alors d'utiliser les expressions (3.29) et (3.30) comme bases d'un algorithme itératif de calcul des estimateurs ML de σ_k^2, soit :

$$\boxed{\sigma_k^{2[t+1]} = \left\{\hat{\mathbf{u}}_k'\left(\boldsymbol{\eta}^{[t]}\right)\hat{\mathbf{u}}_k\left(\boldsymbol{\eta}^{[t]}\right) + \mathrm{tr}\left[\underline{\mathbf{C}}_{kk}\left(\boldsymbol{\eta}^{[t]}\right)\right]\sigma_0^{2[t]}\right\}/q_k}, \qquad (3.31a)$$

$$\boxed{\sigma_0^{2[t+1]} = \left[\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\left(\boldsymbol{\eta}^{[t]}\right)\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'\left(\boldsymbol{\eta}^{[t]}\right)\mathbf{Z}'\mathbf{y}\right]/N} \qquad (3.31b)$$

où $\boldsymbol{\eta}^{[t]} = \left\{\sigma_k^{2[t]}/\sigma_0^{2[t]}\right\}$ est le vecteur $(K\,\mathrm{x}\,1)$ des rapports de variance des K facteurs aléatoires à la variance résiduelle à l'itération t. Ainsi, dès 1973, Henderson anticipait un algorithme de type EM permettant de calculer simplement les estimateurs ML des composantes de variance.
Une variante de cet algorithme qui mérite attention a été formulée par Harville (1977). L'idée est de récrire (3.29) sous la forme suivante: $q_k\hat{\sigma}_k^2 = \hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k + \mathrm{tr}\left(\hat{\underline{\mathbf{C}}}_{kk}\right)\hat{\sigma}_k^2/\hat{\eta}_k$ et de factoriser $\hat{\sigma}_k^2$ à gauche d'où la formule:

$$\boxed{\sigma_k^{2[t+1]} = \left[\hat{\mathbf{u}}_k'\left(\boldsymbol{\eta}^{[t]}\right)\hat{\mathbf{u}}_k\left(\boldsymbol{\eta}^{[t]}\right)\right]/\left\{q_k - \mathrm{tr}\left[\underline{\mathbf{C}}_{kk}\left(\boldsymbol{\eta}^{[t]}\right)\right]/\eta_k^{[t]}\right\}} \qquad (3.31c)$$

qui est combinée pour la variance résiduelle avec (3.31b). Outre la simplicité de leur forme, ces deux algorithmes garantissent la localisation des valeurs dans l'espace paramétrique. Enfin, dans de nombreux exemples, l'algorithme d'Harville s'est avéré nettement plus rapide que celui d'Henderson.

142. Calcul de $-2\,\mathrm{L}_p$

Reprenons l'expression (3.23) de la logvraisemblance profilée (multipliée par moins deux)

$$-2\,\mathrm{L}_P = N\ln 2\pi + \ln|\mathbf{V}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \qquad (3.32)$$

On a montré, d'une part, que $\underline{\mathbf{P}}\mathbf{y} = \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, et d'autre part, que dans le cadre d'un modèle linéaire mixte tel que $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, $\underline{\mathbf{P}}\mathbf{y} = \mathbf{R}^{-1}\hat{\mathbf{e}}$, d'où il découle que :

$$\mathbf{y}'\underline{\mathbf{P}}\mathbf{y} = \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}, \tag{3.33}$$

où $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}'\right)'$est solution des équations du modèle mixte d'Henderson.

Par ailleurs, si l'on utilise les règles du calcul du déterminant de matrices partitionnées (cf annexe III), on montre que :

$$\left|\mathbf{V}\right| = \left|\mathbf{R}\right|\left|\mathbf{G}\right|\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|. \tag{3.34}$$

On en déduit le résultat général suivant, applicable à tout modèle linéaire gaussien de type $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZGZ}' + \mathbf{R})$:

$$\boxed{\begin{aligned} -2\,\mathrm{L}_P = N \ln 2\pi + \ln\left|\mathbf{R}\right| + \ln\left|\mathbf{G}\right| + \ln\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right| \\ + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y} \end{aligned}}. \tag{3.35}$$

Cette formule permet de simplifier grandement le calcul de la logvraisemblance profilée et donc aussi du maximum L_m de la logvraisemblance

$$-2\,\mathrm{L}_m = -2\,\mathrm{L}_P(\mathbf{G} = \hat{\mathbf{G}}_{ML}, \mathbf{R} = \hat{\mathbf{R}}_{ML})$$

grâce au recours aux éléments des équations du modèle mixte d'Henderson. Par ailleurs, cette formule va encore se simplifier dans maintes situations par la prise en compte des structures particulières de $\mathbf{R}$ et de $\mathbf{G}$.

1421. $\mathbf{V} = \mathbf{H}\sigma_0^2$

C'est la formulation d'Hartley-Rao, mais elle s'applique également à des modèles plus complexes qui ne supposent pas nécessairement $\mathbf{R} = \mathbf{I}_N\sigma_0^2$ comme par exemple les modèles à structure d'erreurs autorégressives (Foulley, Jaffrézic et Robert-Granié, 2000). Dans ce cas:

$$\mathbf{y}'\mathbf{R}^{-1}\hat{\mathbf{e}} = \left[\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\mathbf{H}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right] / \hat{\sigma}_0^2 = N\hat{\sigma}_0^2 / \hat{\sigma}_0^2 = N$$

et,

$$-2\,\mathrm{L}_P = N\left(\ln 2\pi + 1\right) + \ln\left|\mathbf{R}\right| + \ln\left|\mathbf{G}\right| + \ln\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|. \tag{3.36}$$

En (3.36), $\mathrm{L}_P = \mathrm{L}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}), \hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta}\right]$, $\mathbf{R} = \mathbf{R}\left[\hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta}\right]$, de même pour $\mathbf{G} = \mathbf{G}\left[\hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta}\right]$, $\boldsymbol{\eta}$ étant le vecteur des paramètres dont dépend $\mathbf{H}$.

1422. $\mathbf{R} = \mathbf{I}_N\sigma_0^2$

Alors $\ln|\mathbf{R}| = N\ln\sigma_0^2$ et $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} = \left(\mathbf{Z}'\mathbf{Z}+\sigma_0^2\mathbf{G}^{-1}\right)/\sigma_0^2$, d'où

$$-2\,\mathrm{L}_P = N\left(\ln 2\pi + 1\right) + \left(N-q\right)\ln\hat{\sigma}_0^2 + \ln|\mathbf{G}| + \ln\left|\mathbf{Z}'\mathbf{Z}+\hat{\sigma}_0^2\mathbf{G}^{-1}\right|, \qquad (3.37)$$

où q représente le nombre de colonnes de $\mathbf{Z}$.

1423. $\mathbf{G} = \bigoplus_{k=1}^{K}\mathbf{G}_k$ et $\mathbf{G}_k = \mathbf{A}_k\sigma_k^2$

C'est la situation relative à K facteurs aléatoires indépendants, chacun ayant une matrice de variance-covariance de la forme $\mathbf{A}_k\sigma_k^2$ où $\mathbf{A}_k$ est une matrice définie-positive connue (par ex $\mathbf{A}$ matrice des relations de parenté entre pères ou entre individus, ou à l'extrême $\mathbf{A}_k = \mathbf{I}_{q_k}$, matrice identité).

$$\begin{aligned} -2\,\mathrm{L}_P = {} & N\left(\ln 2\pi + 1\right) + \left(N - \textstyle\sum_{k=1}^{K} q_k\right)\ln\hat{\sigma}_0^2 + \textstyle\sum_{k=1}^{K} q_k \ln\sigma_k^2 + \\ & \textstyle\sum_{k=1}^{K}\ln|\mathbf{A}_k| + \ln\left|\mathbf{Z}'\mathbf{Z} + \bigoplus_{k=1}^{K}\mathbf{A}_k^{-1}\left(\hat{\sigma}_0^2/\sigma_k^2\right)\right| \end{aligned} . \qquad (3.38)$$

Cet inventaire n'a aucune prétention à l'exhaustivité. Il faudrait également envisager les les modèles multidimensionnels. Dans tous les cas, la formule générale (3.36) peut être appliquée.

15. Tests d'hypothèses

151. Loi asymptotique

Soit $\hat{\boldsymbol{\alpha}}_N$, l'estimateur ML de $\boldsymbol{\alpha}\in\mathrm{A}$ basé sur les observations $\mathbf{y}_N$ d'un échantillon de taille N. Sous des conditions de régularité précisées par ailleurs dans les ouvrages spécialisés (espace paramétrique A compact; logvraisemblance continue et continûment dérivable à l'ordre deux; existence de la matrice d'information et de son inverse), la suite $\sqrt{N}\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}\right)$ converge en loi vers une distribution normale centrée, de matrice de variance-covariance $\mathrm{Lim}\, N\left[\mathbf{J}_N\left(\boldsymbol{\alpha}\right)\right]^{-1}$ quand $N\to\infty$ (Sweeting, 1980 ; Mardia et Marshall, 1984) soit, en bref:

$$\sqrt{N}\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}\right) \underset{\mathcal{L}}{\to} \mathcal{N}\left(0, \mathrm{Lim}\; N\left[\mathbf{J}_N\left(\boldsymbol{\alpha}\right)\right]^{-1}\right), \qquad (3.39)$$

où $\mathbf{J}_N\left(\boldsymbol{\alpha}\right) = \mathrm{E}\left[-\partial^2\,\mathrm{L}\left(\boldsymbol{\alpha};\mathbf{y}_N\right)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'\right]$ est la matrice d'information de Fisher relative à $\boldsymbol{\alpha}$.

Comme $\mathrm{Lim}\, N\left[\mathbf{J}_N\left(\boldsymbol{\alpha}\right)\right]^{-1}$ s'estime de façon convergente par $N\left[\mathbf{J}_N\left(\hat{\boldsymbol{\alpha}}\right)\right]^{-1}$, on peut alors former le pivot asymptotique suivant (Leonard and Hsu, 1999, page 33-35):

$$\hat{\mathbf{J}}_N^{T/2}\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}\right) \underset{\mathcal{L}}{\to} \mathcal{N}\left(\mathbf{0},\mathbf{I}\right), \qquad (3.40)$$

où $\hat{\mathbf{J}}_N^{T/2}$ est la notation condensée relative à la décomposition de Cholesky suivante $\mathbf{J}_N(\hat{\boldsymbol{\alpha}}) = \hat{\mathbf{J}}_N = \hat{\mathbf{J}}_N^{1/2}\hat{\mathbf{J}}_N^{T/2}$.

La propriété en (3.39) se généralise à une fonction $\mathbf{g}(\boldsymbol{\alpha})$ continûment dérivable (de $\mathbb{R}^p$ dans $\mathbb{R}^q$)

$$\sqrt{N}\left[\mathbf{g}(\hat{\boldsymbol{\alpha}}_N) - \mathbf{g}(\boldsymbol{\alpha})\right] \underset{\mathcal{L}}{\rightarrow} \mathcal{N}\left(0, \operatorname{Lim} N \frac{\partial \mathbf{g}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'}\left[\mathbf{J}_N(\boldsymbol{\alpha})\right]^{-1}\frac{\partial \mathbf{g}'(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right). \qquad (3.41)$$

152. Statistique de Wald

On va considérer le test de l'hypothèse nulle: $\mathrm{H}_0 : \mathbf{k}'\boldsymbol{\beta} = \mathbf{m}$ contre son alternative contraire: $\mathrm{H}_1 : \mathbf{k}'\boldsymbol{\beta} \neq \mathbf{m}$ où $\mathbf{k}'$ est une matrice $(r \times p)$ avec $r < p$ dont les r lignes sont linéairement indépendantes et $\mathbf{m}$ un vecteur $(r \times 1)$ de constantes, souvent nulles mais pas nécessairement.

Nous avons vu précédemment (cf (3.13)) qu'asymptotiquement les lois de $\hat{\boldsymbol{\beta}}$ et de $\hat{\boldsymbol{\gamma}}$ (estimateurs ML) étaient indépendantes sachant que :

$$\mathbf{J}_N(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{J}_\beta = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_\gamma = \mathbf{F}/2 \end{bmatrix}.$$

Dans ces conditions, on peut appliquer les résultats (3.40) et (3.41) à $\mathbf{k}'\hat{\boldsymbol{\beta}}$, soit, sous l'hypothèse nulle,

$$\sqrt{N}\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \underset{\mathcal{L}}{\rightarrow} \mathcal{N}\left(0, \operatorname{Lim} N\, \mathbf{k}'\mathbf{J}_\beta^{-1}\mathbf{k}\right), \qquad (3.42)$$

et, en posant $\hat{\mathbf{J}}_\beta = \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}$

$$\left[\left(\mathbf{k}'\hat{\mathbf{J}}_\beta^{-1}\mathbf{k}\right)^{-1}\right]^{T/2}\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \underset{\mathcal{L}}{\rightarrow} \mathcal{N}\left(\mathbf{0}, \mathbf{I}_r\right), \qquad (3.43)$$

d'où, l'on déduit le Khi-deux asymptotique à r degrés de liberté:

$$\boxed{\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{k}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{k}\right]^{-1}\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \underset{\mathcal{L}}{\rightarrow} \chi_r^2}, \qquad (3.44)$$

qui est la statistique de Wald relative au test étudié. On obtient donc formellement la même chose que dans le cas où $\mathbf{V}$ est connu, à la nuance près qu'il s'agit ici d'une distribution asymptotique. C'est pourquoi, l'on voit souvent cette propriété présentée sous la forme classique suivante :

$$\mathbf{k}'\hat{\boldsymbol{\beta}} \approx \mathcal{N}\left[\mathbf{k}'\boldsymbol{\beta},\ \mathbf{k}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{k}\right] \tag{3.45}$$

A proprement parler, la distribution asymptotique de $\mathbf{k}'\hat{\boldsymbol{\beta}}$ est dégénérée et cette notation est donc un abus de langage qu'il faut interpréter avec prudence comme un raccourci opérationnel, en gardant à l'esprit le cheminement rigoureux qui y conduit.

De nombreux logiciels proposent une option de Fisher-Snedecor pour ce test des effets fixes par analogie avec le cas où $\mathbf{V}$ est connu à σ_0^2 près. En effet si $\mathbf{V} = \mathbf{H}\sigma_0^2$ et $\mathbf{H}$ est connu, en désignant par W la statistique $\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{k}'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{k}\right]^{-1}\left(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)$, on sait que, sous H_0, $\left[W\left(\hat{\sigma}_0^2\right)\right]/r \sim \mathrm{F}\left[r, N - \mathrm{r}(\mathbf{X})\right]$. Ici, on forme $\hat{W}/r$ où $\hat{W}$ est la statistique (51) qu'on compare à un $F(r,d)$ avec un nombre de degrés de liberté d qui est calculé selon une méthode approchée (Satterthwaite par exemple). Mais ce procédé n'a pas de justification théorique.

153. Statistique du rapport de vraisemblance

Une alternative au test de Wald réside dans celui du rapport de vraisemblance de Neyman-Pearson qu'on peut formuler ainsi (Mood et al, 1974, page 419 ; Cox et Hinkley, 1974, page 322, formule 50):

$\mathrm{H}_0 : \left\{\boldsymbol{\beta} \in \mathrm{B}_0 \subset \mathbb{R}^p\right\} \mathrm{x} \left\{\boldsymbol{\gamma} \in \Gamma\right\}$ contre $\mathrm{H}_1 : \left\{\boldsymbol{\beta} \in \left(\mathrm{B} \setminus \mathrm{B}_0\right)\right\} \mathrm{x} \left\{\boldsymbol{\gamma} \in \Gamma\right\}$. Par exemple, dans le cas précédent, B correspond à $\mathbb{R}^p$ et B_0 est un sous-espace réel de dimension $p - r$ correspondant à $\mathbb{R}^p$ contraint par les r relations $\mathbf{k}'\boldsymbol{\beta} = \mathbf{m}$.

Si l'on considère le maximum de la logvraisemblance $\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = \log \mathrm{p}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma})$ selon les deux modalités H_0 et $\mathrm{H}_0 \cup \mathrm{H}_1$, et que l'on note respectivement:

$$\mathbf{L}_R = \mathrm{Max}_{\boldsymbol{\beta} \in \mathrm{B}_0, \boldsymbol{\gamma} \in \Gamma}\, \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$$

$$\mathbf{L}_C = \mathrm{Max}_{\boldsymbol{\beta} \in \mathrm{B}, \boldsymbol{\gamma} \in \Gamma}\, \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}),$$

on sait que la statistique $\lambda = -2\mathbf{L}_R + 2\mathbf{L}_C$ suit asymptotiquement, sous H_0, une loi de Khi-deux dont le nombre de degrés de liberté est la différence de dimensions de B et de B_0, (Cox and Hinkley, 1974, page 322) soit

$$\boxed{\lambda = -2\mathbf{L}_R + 2\mathbf{L}_C\big|_{\mathrm{H}_0} \underset{\mathcal{L}}{\rightarrow} \chi^2_{\dim(\mathrm{B}) - \dim(\mathrm{B}_0)}} \tag{3.46}$$

154. Statistique du score

Si l'on se place dans les mêmes conditions que précédemment, le test du score proposé par Rao (1973) s'appuie sur la statistique suivante :

$$U = \tilde{\mathbf{S}}_\beta' \tilde{\mathbf{J}}_\beta^- \tilde{\mathbf{S}}_\beta , \tag{3.47}$$

où $\tilde{\mathbf{S}}_\beta$ est la valeur de la fonction score $\mathbf{S}_\beta = \mathbf{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = \partial \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\beta}$ évaluée au point des estimations ML, $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ et $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$ obtenues sous le modèle réduit et $\tilde{\mathbf{J}}_\beta$, la valeur de la matrice d'information de Fisher $\mathbf{J}_\beta = -\mathrm{E}\left[\partial^2 \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'\right]$ relative à $\boldsymbol{\beta}$, évaluée au même point soit $\tilde{\mathbf{J}}_\beta = \mathbf{J}_\beta(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$.

L'idée de ce test est très simple : si l'on évaluait la fonction $U(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{S}_\beta' \mathbf{J}_\beta^- \mathbf{S}_\beta$ au point des estimations ML $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ et $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$ obtenues sous le modèle complet, alors $U(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = 0$ puisque par définition, $\mathbf{S}_\beta(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}; \mathbf{y}) = \mathbf{0}$. Evaluée en $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ et $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$, cette forme quadratique s'interprète comme une distance à sa valeur de référence nulle. Si elle est proche de zéro, on aura tendance à accepter H_0 ; au contraire, plus sa valeur sera grande, plus on sera enclin à ne pas accepter cette hypothèse. Comme précédemment, sous l'hypothèse nulle, la statistique $U(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ tend asymptotiquement vers une loi de Khi-deux dont le nombre de degrés de liberté est la différence entre le nombre de paramètres du modèle complet et celui du modèle réduit

$$\boxed{U(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})\big|_{\mathrm{H}_0} \underset{\mathcal{L}}{\rightarrow} \chi^2_{\dim(\mathrm{B}) - \dim(\mathrm{B}_0)}} . \tag{3.48}$$

Un cas particulièrement intéressant est celui du test d'absence d'effets $\mathrm{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ résultant de la comparaison du modèle réduit : $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{e}$ et du modèle complet $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{e}$ où, sous les deux modèles, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$. Par ailleurs, $\mathbf{X}_1$ et $\mathbf{X}_2$ sont supposés de plein rang pour simplifier. Dans ce cas, la fonction du score s'écrit $\mathbf{S}_\beta = \mathbf{X}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ et sa valeur sous H_0 se réduit à $\tilde{\mathbf{S}}_\beta = \begin{bmatrix} \mathbf{0} \\ \mathbf{X}_2' \tilde{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}_1 \tilde{\boldsymbol{\beta}}_1) \end{bmatrix}$ où $\tilde{\mathbf{V}} = \mathbf{V}(\tilde{\boldsymbol{\gamma}})$ puisque, par définition, le score sous le modèle réduit est tel que $\mathbf{X}_1' \tilde{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}_1 \tilde{\boldsymbol{\beta}}_1) = \mathbf{0}$. Il en résulte que

$$U = (\mathbf{y} - \mathbf{X}_1 \tilde{\boldsymbol{\beta}}_1)' \tilde{\mathbf{V}}^{-1} \mathbf{X}_2 \left[\left(\mathbf{X}' \tilde{\mathbf{V}}^{-1} \mathbf{X} \right)^{-1} \right]_{22} \mathbf{X}_2' \tilde{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}_1 \tilde{\boldsymbol{\beta}}_1) . \tag{3.49}$$

Si l'on pose $\underline{\mathbf{P}}_1 = \mathbf{V}^{-1} (\mathbf{I}_N - \mathbf{Q}_1)$ avec $\mathbf{Q}_1 = \mathbf{X}_1 \left(\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1 \right)^{-1} \mathbf{X}_1' \mathbf{V}^{-1}$, (3.49) peut s'écrire aussi

$$U = \mathbf{y}'\underline{\tilde{\mathbf{P}}}_1\mathbf{X}_2\left(\mathbf{X}_2'\underline{\tilde{\mathbf{P}}}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\underline{\tilde{\mathbf{P}}}_1\mathbf{y} = \left[\hat{\boldsymbol{\beta}}_2(\tilde{\boldsymbol{\gamma}})\right]'\mathbf{X}_2'\underline{\tilde{\mathbf{P}}}_1\mathbf{y}, \qquad (3.50)$$

où $\hat{\boldsymbol{\beta}}_2(\tilde{\boldsymbol{\gamma}})$ est solution du système $\left(\mathbf{X}_2'\underline{\tilde{\mathbf{P}}}_1\mathbf{X}_2\right)\hat{\boldsymbol{\beta}}_2(\tilde{\boldsymbol{\gamma}}) = \mathbf{X}_2'\underline{\tilde{\mathbf{P}}}_1\mathbf{y}$ ou celle en $\boldsymbol{\beta}_2$ du système général $\left(\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X}\right)\hat{\boldsymbol{\beta}}(\tilde{\boldsymbol{\gamma}}) = \mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{y}$.

Il est intéressant de comparer cette statistique à celle de Wald appliquée au même test d'hypothèses. Par application de (3.44), il vient :

$$W = \left[\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}})\right]'\left(\mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{X}_2\right)\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}}), \qquad (3.51)$$

où $\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}})$ est solution du système $\left(\mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{X}_2\right)\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}}) = \mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{y}$ et $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\gamma}})$ avec $\hat{\boldsymbol{\gamma}}$ estimation ML sous le modèle complet. Il en résulte que

$$W = \left[\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}})\right]'\mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{y}. \qquad (3.52)$$

A l'examen de (3.50) et de (3.51), il s'avère que les statistiques de Wald et du score ont donc la même forme, la différence entre elles étant que la première est basée sur une estimation ML de $\mathbf{V}$ soit $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\gamma}})$ obtenue sous le modèle complet alors que la seconde utilise l'estimation $\tilde{\mathbf{V}} = \mathbf{V}(\tilde{\boldsymbol{\gamma}})$ sous le modèle réduit. Ces statistiques U et W peuvent se calculer aisément grâce à une formule développée par Harvey (1970, formule 3, p487).

155. Discussion

Les trois tests sont équivalents asymptotiquement (Rao, 1973 ; Gourieroux et Monfort, 1989). Le débat reste ouvert quant à leurs mérites respectifs à distance finie, avec toutefois une préférence de certains spécialistes pour le test de Neyman-Pearson notamment si l'on replace la comparaison de modèles dans un cadre plus général tel que celui adopté par les Bayésiens. S'agissant de conditions asymptotiques, il importe également de s'assurer que la structure particulière des modèles étudiés autorise bien une application raisonnable de celles-ci. Le nombre d'observations ou d'unités expérimentales (individus par ex) est-il suffisant? D'une part, que se passe-t-il quand le nombre d'observations augmente? Est-ce que la dimension p de $\boldsymbol{\beta}$ augmente corrélativement ou non? Si oui, comment varie N/p ?

Le test du rapport de vraisemblance nécessite de contraster deux modèles: le modèle complet (C) et le modèle réduit (R) correspondant à H_0 alors que la statistique de Wald ne requiert que la mise en œuvre du modèle complet. La statistique de Wald offre toutefois le désavantage de ne pas être invariante par transformation non linéaire des paramètres. Enfin,

avec les formules de calcul du maximum de la logvraisemblance présentées précédemment, la différence en terme de difficulté et temps de calcul entre les deux n'est pas si grande.

Il est important de souligner que les deux modèles contrastés vis-à-vis des effets fixes $\boldsymbol{\beta}$ comportent la même structure de variance-covariance $\mathbf{V}(\boldsymbol{\gamma})$. De la même façon, toute comparaison de structures de $\mathbf{V}(\boldsymbol{\gamma})$ se fera à structure d'espérance identique. Cette contrainte technique inhérente à la procédure de test n'est pas sans poser des interrogations sur la méthode de choix de ces deux structures dans les modèles linéaires mixtes. Pour contourner cette circularité, on pourra être amené à développer des tests robustes d'une des structures qui soient peu sensibles à l'autre.

Ainsi, dans le cas de données répétées $\mathbf{y}_i = \{y_{ij}\}$ sur une même unité expérimentale i, le test robuste des effets fixes de Liang et Zeger (1986) permet de s'affranchir, dans une certaine mesure, de l'incertitude qui existe sur la structure de variance covariance des observations. Il se fonde sur l'estimateur « sandwich » de la variance d'échantillonnage de l'estimateur $\hat{\boldsymbol{\beta}}=\left(\sum_{i=1}^{I}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i\right)^{-}\sum_{i=1}^{I}\mathbf{X}_i'\mathbf{W}_i\mathbf{y}_i$ des moindres carrés pondérés, soit

$$\mathrm{Var}(\mathbf{k}'\hat{\boldsymbol{\beta}}) = \mathbf{k}'\left(\sum_{i=1}^{I}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i\right)^{-}\left(\sum_{i=1}^{I}\mathbf{X}_i'\mathbf{W}_i\mathbf{V}_i\mathbf{W}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{I}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i\right)^{-}\mathbf{k}$$

où $\mathbf{W}_i$ est une matrice de travail et la variance $\mathbf{V}_i = \mathrm{var}(\mathbf{y}_i)$ est remplacée par une estimation convergente $\hat{\mathbf{V}}_i = \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)'$.

Enfin, pour des raisons de concision, la discussion des tests relatifs aux structures de dispersion est reportée à la suite de l'exposé de la méthode REML ce qui n'exclut pas qu'on puisse les envisager dans le cadre d'une estimation ML de tels paramètres.

2. Méthode dite REML

21. Exemple simple

211. Estimateur

Pourquoi REML plutôt que ML? Nous allons aborder cette question à travers un exemple simple: celui de l'estimation de la variance à partir d'un échantillon de N observations $y_i \sim_{\text{iid}} \mathcal{N}(\mu,\sigma^2)$ supposées indépendantes et de même loi normale d'espérance μ et de variance. σ^2. Du fait de l'indépendance des y_i, la logvraisemblance se met sous la forme suivante :

$$-2\,\mathrm{L}\left(\mu,\sigma^2;\mathbf{y}\right) = N\left(\ln 2\pi + \ln\sigma^2\right) + \sum_{i=1}^{N}\left(y_i-\mu\right)^2/\sigma^2 . \tag{3.53}$$

On peut décomposer $\sum_{i=1}^{N}\left(y_i-\mu\right)^2$ en la somme

$$\sum_{i=1}^{N}\left(y_i-\mu\right)^2 = N\left[s^2+\left(\overline{y}-\mu\right)^2\right], \tag{3.54}$$

où $\overline{y}=\left(\sum_{i=1}^{N} y_i\right)/N$ est la moyenne des observations, et $s^2=\sum_{i=1}^{N}\left(y_i-\overline{y}\right)^2/N$, la variance de l'échantillon, d'où

$$-2\,\mathrm{L}\left(\mu,\sigma^2;\mathbf{y}\right) = N\left[\ln 2\pi + \ln\sigma^2 + \frac{s^2+\left(\overline{y}-\mu\right)^2}{\sigma^2}\right], \tag{3.55}$$

et les dérivées partielles par rapport à μ et σ^2 :

$$\partial\left(-2\,\mathrm{L}/N\right)/\partial\mu = -2\left(\overline{y}-\mu\right)/\sigma^2, \tag{3.56a}$$

$$\partial\left(-2\,\mathrm{L}/N\right)/\partial\sigma^2 = \frac{1}{\sigma^2} - \frac{s^2+\left(\overline{y}-\mu\right)^2}{\sigma^4}. \tag{3.56b}$$

Par annulation de ces dérivées, on obtient:

$$\hat{\mu} = \overline{y}\ , \tag{3.57}$$

Et, pour $N \geq 2$,

$$\hat{\sigma}^2_{ML} = s^2 + \left(\overline{y}-\hat{\mu}\right)^2 = s^2 . \tag{3.58}$$

Or

$$\mathrm{E}\left(\hat{\sigma}^2_{ML}\right) = \left(N-1\right)\sigma^2/N \tag{3.59}$$

indiquant que l'estimateur s^2 du maximum de vraisemblance de σ^2 est biaisé par défaut, la valeur du biais étant de $-\sigma^2/N$. C'est la constatation de ce biais qui est à l'origine du développement du concept de vraisemblance restreinte (ou résiduelle).

212. Correction du biais

L'estimation de μ interférant avec celle de σ^2, on va faire en sorte d'éliminer μ. Pour ce faire deux approches sont envisageables qui préfigurent les méthodes générales exposées par la suite.

2121. Factorisation de la vraisemblance

Le principe est le suivant : on factorise la vraisemblance en deux parties et on ne retient pour l'estimation de la variance que celle qui ne dépend pas de μ. A cet égard, on considère la transformation biunivoque suivante :

$$\mathbf{y}_{(N\text{x}1)} = \{y_i\} \leftrightarrow \mathbf{y}^*_{(N\text{x}1)} = \left(\mathbf{z}'_{N-1}, \bar{y}\right)' \tag{3.60}$$

où $\mathbf{z}_{N-1} = \{z_i = y_i - \bar{y}; i = 1,2,...,N-1\}$, le vecteur des $N-1$ écarts élémentaires à la moyenne. S'agissant d'une transformation biunivoque, on peut donc relier les densités de $\mathbf{y}$ et de $\mathbf{y}^*$ par l'expression:

$$\mathrm{p}_Y\left(\mathbf{y}\middle|\mu,\sigma^2\right) = \mathrm{p}_{Y^*}\left(\mathbf{y}^*\middle|\mu,\sigma^2\right)|J| \tag{3.61}$$

où $|J|$ est la valeur absolue du jacobien $J = \det\left(\dfrac{\partial \mathbf{y}^{*'}}{\partial \mathbf{y}}\right)$ de la transformation.

Or, $\bar{y}$ et $\mathbf{z}_{N-1}$ sont indépendantes et la loi de $\mathbf{z}_{N-1}$ ne dépend pas de μ d'où la factorisation de la densité de $\mathbf{y}^*$ en:

$$\mathrm{p}_{Y^*}\left(\mathbf{y}^*\middle|\mu,\sigma^2\right) = \mathrm{p}_Z\left(\mathbf{z}_{N-1}\middle|\sigma^2\right)\mathrm{p}_{\bar{Y}}\left(\bar{y}\middle|\mu,\sigma^2\right), \tag{3.62}$$

Par ailleurs, eu égard à la définition de la transformation (3.60), la valeur du jacobien J ne dépend pas des paramètres; on en déduit donc la décomposition suivante de la logvraisemblance $\mathrm{L}(\mu,\sigma^2;\mathbf{y}) = \ln \mathrm{p}_Y\left(\mathbf{y}\middle|\mu,\sigma^2\right)$:

$$\boxed{\mathrm{L}(\mu,\sigma^2;\mathbf{y}) = \mathrm{L}_1(\sigma^2;\mathbf{z}_{N-1}) + \mathrm{L}_2(\mu,\sigma^2;\bar{y}) + cste}, \tag{3.63}$$

où $\mathrm{L}_1(\sigma^2;\mathbf{z}_{N-1}) = \ln \mathrm{p}_Z\left(\mathbf{z}_{N-1}\middle|\sigma^2\right)$, $\mathrm{L}_2(\mu,\sigma^2;\bar{y}) = \ln \mathrm{p}_{\bar{Y}}\left(\bar{y}\middle|\mu,\sigma^2\right)$, la constante étant égale à $\ln|J|$.

L'idée sous-jacente à REML consiste à n'utiliser que $\mathrm{L}_1(\sigma^2;\mathbf{z}_{N-1})$ pour faire inférence sur σ^2, d'où le nom de (log)vraisemblance résiduelle ou restreinte (la restriction portant sur l'espace d'échantillonnage) donné par Thompson (1989) à cette fonction ou de (log)vraisemblance de «contrastes d'erreur» selon la terminologie d'Harville.

Par spécification directe de la loi de $\mathbf{z}_{N-1} \sim \mathcal{N}(0, \mathbf{V}_Z)$ avec $\mathbf{V}_Z = \sigma^2(\mathbf{I}_{N-1} - \mathbf{J}_{N-1}/N)$, (par définition $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N^{'}$) ou indirectement, compte tenu de (3.63), on montre que :

$$\boxed{-2\,\mathrm{L}_1(\sigma^2; \mathbf{z}_{N-1}) = (N-1)\left(\ln 2\pi + \ln \sigma^2\right) - \ln N + \left[\sum_{i=1}^{N} (y_i - \overline{y})^2\right]/\sigma^2} \quad (3.64)$$

Il s'en suit que :

$$\frac{\partial(-2\,\mathrm{L}_1)}{\partial \sigma^2} = \left[(N-1)\sigma^2 - Ns^2\right]/\sigma^4,$$

et, par annulation :

$$\hat{\sigma}^2 = Ns^2/(N-1);\ N \geq 2 \quad (3.65)$$

qui est l'estimateur usuel, sans biais, de σ^2.

2122. Remplacement de μ par son espérance conditionnelle

Le point de départ du raisonnement réside dans la remarque suivante: si μ était connu, l'estimateur ML de σ^2 serait, comme indiqué en (3.58): $\hat{\sigma}^2 = s^2 + (\overline{y} - \mu)^2$ dont la valeur est toujours supérieure ou égale à l'estimateur $\hat{\sigma}^2 = s^2$; μ est généralement inconnu, mais on peut prédire sa contribution au terme $(\overline{y} - \mu)^2$ en remplaçant ce dernier par son espérance conditionnelle sachant les observations $\mathrm{E}\left[(\overline{y} - \mu)^2 \middle| \mathbf{y}, \sigma^2\right]$ à l'instar de ce qui est fait avec l'algorithme EM (Foulley, 1993).

L'écriture du pivot normal réduit $\dfrac{\overline{y} - \mu}{\sqrt{\sigma^2/N}} \sim \mathcal{N}(0,1)$ peut s'interpréter à la fois, en statistique classique, comme $\overline{y} | \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2/N)$ ou, en statistique fiduciaire (au sens de Fisher), comme $\mu | \overline{y}, \sigma^2 \sim \mathcal{N}(\overline{y}, \sigma^2/N)$. Si l'on admet cette dernière interprétation, on a

$$\mathrm{E}\left[(\overline{y} - \mu)^2 \middle| \mathbf{y}, \sigma^2\right] = \mathrm{Var}\left(\mu | \overline{y}, \sigma^2\right) = \sigma^2/N,$$

et l'équation à résoudre devient: $\hat{\sigma}^2 = s^2 + \hat{\sigma}^2/N$ qui a pour solution la même expression que celle obtenue en (3.67) par maximisation de la logvraisemblance résiduelle. Cette approche illustre bien le fait que le biais de l'estimateur de σ^2 tire son origine de la mauvaise prise en compte par ML de l'incertitude liée à la fluctuation de μ autour de son estimation $\overline{y}$.

22. Cas général

221. Concept de vraisemblance marginale

Ce concept a été formalisé en statistique classique par Kalbfleisch et Sprott (1970). En résumé, le problème revient à chercher une transformation biunivoque de $\mathbf{y}$ en $(\mathbf{u}', \mathbf{v}')'$ telle que les deux conditions suivantes portant sur l'expression de la densité conjointe $f(\mathbf{u}, \mathbf{v}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = f(\mathbf{v}|\boldsymbol{\beta}, \boldsymbol{\gamma})\, g(\mathbf{u}|\mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ soient réalisées:

a) $f(\mathbf{v}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = f(\mathbf{v}|\boldsymbol{\gamma})$

b) $g(\mathbf{u}|\mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ «contains no available information concerning $\boldsymbol{\gamma}$ in the absence of knowledge of $\boldsymbol{\beta}$»

La densité en a) permet ainsi de définir la vraisemblance «marginale» de $\boldsymbol{\gamma}$. On dit corrélativement que $\mathbf{v}$ est une statistique «ancillaire» de $\boldsymbol{\beta}$, considéré ici comme paramètre parasite, alors que $\boldsymbol{\gamma}$ est le paramètre d'intérêt.

Il faut bien admettre que la formulation de la condition b) reste quelque peu obscure surtout en l'absence de critère rigoureux de vérification. Mc Cullagh et Nelder (1989) reconnaîtront eux-mêmes la difficulté de justifier clairement l'inutilité de cette information[5] en l'appliquant au cas du modèle mixte gaussien. Dans la discussion de cet article, un des rapporteurs (Barnard) mit en avant le caractère indissociable des informations imputables à $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$ dans $g(\mathbf{u}|\mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ («This information is inextricably mixed up with the nuisance parameters»). Toujours est-il que c'est bien ce concept de vraisemblance marginale qui est à l'origine de la théorie classique de REML comme en atteste bien l'acronyme MMLE (Marginal Maximum Likelihood Estimator) proposé par Rao pour désigner cet estimateur (Rao, 1979).

222. Application au modèle linéaire mixte gaussien

Dans le cadre du modèle $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, Patterson et Thompson (1971) proposèrent le choix suivant pour la transformation $\mathbf{y} \leftrightarrow (\mathbf{u}', \mathbf{v}')' : \mathbf{u} = \underline{\mathbf{H}}\mathbf{y}$ et $\mathbf{v} = \mathbf{S}\mathbf{y} = (\mathbf{I}_N - \underline{\mathbf{H}})\mathbf{y}$, où $\underline{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ est le projecteur classique des moindres carrés «simples» qui est appelé aussi «hat matrix»[6] dans la littérature anglo-saxonne. Par définition, la variable $\mathbf{v}$ est

[5] "In this example, there appears to be no loss of information on $\boldsymbol{\gamma}$ by using $R(\text{"}\mathbf{v}\text{"})$ in place of $\mathbf{Y}$, though it is difficult to give a totally satisfactory justification of this claim"

[6] Car $\underline{\mathbf{H}}\mathbf{y} = \hat{\mathbf{y}}$

ancillaire de $\boldsymbol{\beta}$ puisque $\mathbf{SX}=\mathbf{0}$ et va donc servir à définir la vraisemblance «marginale» de $\boldsymbol{\gamma}$.

Deux remarques méritent l'attention à ce stade:

1) En fait, peu importe le choix du projecteur pourvu que celui-ci ne dépende pas des paramètres. On aurait pu prendre aussi bien $\tilde{\mathbf{v}}=\tilde{\mathbf{S}}\mathbf{y}=\left(\mathbf{I}_N-\tilde{\mathbf{H}}\right)\mathbf{y}$ où $\tilde{\mathbf{H}}=\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}$, ($\mathbf{W}$ étant une matrice symétrique connue définie-positive) puisque alors $\tilde{\mathbf{v}}=\tilde{\mathbf{S}}\mathbf{v}$.

2) $\mathbf{v}$ comporte N éléments dont certains sont linéairement dépendants. Pour éliminer cette information redondante, Harville (1977) proposa de ne considérer dans la vraisemblance marginale qu'un sous-vecteur noté $\mathbf{K}'\mathbf{y}$ formé de $N-\mathrm{r}(\mathbf{X})$ éléments linéairement indépendants appelés «contrastes d'erreur». Pour ce faire, il suffit comme l'ont montré Searle et al (1992, p251) de prendre $\mathbf{K}'$ sous la forme $\mathbf{WS}$ où $\mathbf{W}$ est une matrice $\left[N-\mathrm{r}(\mathbf{X})\right]\mathrm{x}\,N$ de plein rang suivant les lignes. Un choix possible consiste (Searle, 1979) à bâtir $\mathbf{K}$ avec les $N-\mathrm{r}(\mathbf{X})$ premiers vecteurs propres du projecteur $\mathbf{S}=\mathbf{I}_N-\underline{\mathbf{H}}$; soit $\mathbf{A}$ de dimension $N\,\mathrm{x}\left[N-\mathrm{r}(\mathbf{X})\right]$ cette matrice, elle satisfait alors $\mathbf{A}'\mathbf{A}=\mathbf{I}_{N-\mathrm{r}(\mathbf{X})}$ et $\mathbf{A}\mathbf{A}'=\mathbf{S}$ et on vérifie aisément que $\mathbf{A}'$ peut se mettre sous la forme $\mathbf{WS}$ indiquée ci-dessus.

Sur cette base, on peut exprimer la logvraisemblance résiduelle comme la logvraisemblance de $\boldsymbol{\gamma}$ basée sur $\mathbf{K}'\mathbf{y}$, soit :

$$-2\,\mathrm{L}\left(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y}\right)=\left[N-\mathrm{r}(\mathbf{X})\right]\ln 2\pi+\ln\left|\mathbf{K}'\mathbf{V}\mathbf{K}\right|+\mathbf{y}'\mathbf{K}\left(\mathbf{K}'\mathbf{V}\mathbf{K}\right)^{-1}\mathbf{K}'\mathbf{y}\,. \qquad (3.66)$$

Cette expression va grandement se simplifier du fait des relations suivantes (Searle, 1979, p2.14 à 2.17; Quaas, 1992 ; Rao et Kleffe, 1988, p247) :

$$\left|\mathbf{K}'\mathbf{V}\mathbf{K}\right|=\left|\mathbf{V}\right|\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|\left|\underline{\mathbf{X}}'\underline{\mathbf{X}}\right|^{-1}\left|\mathbf{K}'\mathbf{K}\right|, \qquad (3.67a)$$

$$\mathbf{K}\left(\mathbf{K}'\mathbf{V}\mathbf{K}\right)^{-1}\mathbf{K}'=\underline{\mathbf{P}} \qquad (3.67b)$$

où $\underline{\mathbf{X}}$ est une matrice d'incidence correspondant à une paramétrisation de plein rang, ($\underline{\mathbf{X}}$ correspond à toute matrice formée par $\mathrm{r}(\mathbf{X})$ colonnes de $\mathbf{X}$ linéairement indépendantes si bien que $\mathrm{r}(\underline{\mathbf{X}})=p$) et $\underline{\mathbf{P}}=\mathbf{V}^{-1}\left(\mathbf{I}_N-\mathbf{Q}\right)$ avec $\mathbf{Q}=\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$.

En insérant (3.67ab) dans (3.66), et en isolant la constante C, on obtient l'expression suivante de la logvraisemblance :

$$\boxed{-2\,\mathrm{L}\left(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y}\right)=C+\ln\left|\mathbf{V}\right|+\ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|+\mathbf{y}'\underline{\mathbf{P}}\mathbf{y}} \qquad (3.68a)$$

avec

$$C = \left[N - \mathrm{r}(\mathbf{X})\right]\ln 2\pi - \ln\left|\underline{\mathbf{X}}'\underline{\mathbf{X}}\right| + \ln\left|\mathbf{K}'\mathbf{K}\right|. \tag{3.68b}$$

Dans certains ouvrages et articles, on trouve d'autres valeurs de constantes, telles que :

$$C' = \left[N - \mathrm{r}(\mathbf{X})\right]\ln 2\pi - \ln\left|\underline{\mathbf{X}}'\underline{\mathbf{X}}\right|, \tag{3.69a}$$

$$C'' = \left[N - \mathrm{r}(\mathbf{X})\right]\ln 2\pi. \tag{3.69b}$$

La première (3.69a) (Welham et Thompson, 1997) est liée au choix particulier de $\mathbf{K} = \mathbf{A}$ proposé par Searle (1979) et tel que $\mathbf{A}'\mathbf{A} = \mathbf{I}_{N-\mathrm{r}(\mathbf{X})}$. La valeur C'' résulte de l'interprétation bayésienne de la vraisemblance marginale et sera développée dans le paragraphe suivant.

Si l'on dérive maintenant (3.68a) par rapport à γ_k, il vient :

$$\frac{\partial\left[-2\,\mathrm{L}\left(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y}\right)\right]}{\partial\gamma_k} = \frac{\partial\ln\left|\mathbf{V}\right|}{\partial\gamma_k} + \frac{\partial\ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial\gamma_k} + \mathbf{y}'\frac{\partial\underline{\mathbf{P}}}{\partial\gamma_k}\mathbf{y} \tag{3.70}$$

Par manipulation algébrique, on montre que :

$$\frac{\partial\ln\left|\mathbf{V}\right|}{\partial\gamma_k} + \frac{\partial\ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial\gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\right). \tag{3.71}$$

De même, à partir de $\underline{\mathbf{P}} = \mathbf{V}^{-1}\left(\mathbf{I}_N - \mathbf{Q}\right)$, il vient (cf démonstration en annexe II)

$$\frac{\partial\underline{\mathbf{P}}}{\partial\gamma_k} = -\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}},$$

d'où

$$\frac{\partial\left[-2\,\mathrm{L}\left(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y}\right)\right]}{\partial\gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\right) - \mathbf{y}'\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\mathbf{y} \tag{3.72}$$

Si $\mathbf{V}$ a une structure linéaire, soit $\mathbf{V} = \sum_{l=0}^{K}\mathbf{V}_l\gamma_l$ avec $\partial\mathbf{V}/\partial\gamma_k = \mathbf{V}_k$, et sachant que $\underline{\mathbf{P}}\mathbf{V}\underline{\mathbf{P}} = \underline{\mathbf{P}}$, alors $\mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{V}_k\right) = \sum_{l=0}^{K}\mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{V}_l\right)\gamma_l$ et le système des équations REML s'écrit:

$$\boxed{\sum_{l=0}^{K}\mathrm{tr}\left(\hat{\underline{\mathbf{P}}}\mathbf{V}_k\hat{\underline{\mathbf{P}}}\mathbf{V}_l\right)\hat{\gamma}_l = \mathbf{y}'\hat{\underline{\mathbf{P}}}\mathbf{V}_k\hat{\underline{\mathbf{P}}}\mathbf{y}}. \tag{3.73}$$

En posant

$$\tilde{\mathbf{F}} = \left\{\tilde{f}_{kl}\right\} = \left\{\mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{V}_l\right)\right\} \tag{3.74a}$$

$$\mathbf{g} = \left\{g_k\right\} = \left\{\mathbf{y}'\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{y}\right\}. \tag{3.74b}$$

Le système en (3.73) peut être résolu numériquement par un algorithme itératif ayant la forme d'un système linéaire en $\boldsymbol{\gamma}$:

$$\boxed{\tilde{\mathbf{F}}\left(\boldsymbol{\gamma}^{[n]}\right)\boldsymbol{\gamma}^{[n+1]} = \mathbf{g}\left(\boldsymbol{\gamma}^{[n]}\right)}, \tag{3.75}$$

La remarque faite à propos de ML s'applique également ici quant au calcul des éléments de $\tilde{\mathbf{F}}$ qui se simplifie en tirant avantage de la forme que prend la trace du produit d'une matrice et de sa transposée. Ainsi, $\tilde{f}_{kl} = \sum_{ij} \left\{ \mathbf{Z}_k' \underline{\mathbf{P}} \mathbf{Z}_l \right\}_{ij}^2$.

Au vu de ces équations, tout se passe de ML à REML comme si la matrice $\underline{\mathbf{P}}$ était substituée à $\mathbf{V}^{-1}$ dans la matrice des coefficients du système (82), ces deux matrices partageant la propriété d'avoir $\mathbf{V}$ comme inverse (respectivement généralisée et classique). Mais, cette substitution a son importance sur les propriétés de ML et de REML. Ainsi, l'espérance du score $\frac{\partial \left[\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}) \right]}{\partial \gamma_k} = \frac{1}{2} \left[\mathbf{y}' \underline{\mathbf{P}} \mathbf{V}_k \underline{\mathbf{P}} \mathbf{y} - \mathrm{tr}\left(\underline{\mathbf{P}} \mathbf{V}_k \right) \right]$ des équations REML est par définition nulle, alors que celle du score relatif à la vraisemblance profilée $\frac{\partial \left[\mathrm{L}_P(\boldsymbol{\gamma}; \mathbf{y}) \right]}{\partial \gamma_k} = \frac{1}{2} \left[\mathbf{y}' \underline{\mathbf{P}} \mathbf{V}_k \underline{\mathbf{P}} \mathbf{y} - \mathrm{tr}\left(\mathbf{V}^{-1} \mathbf{V}_k \right) \right]$ ne peut l'être. Cette différence de propriété est mise en avant par Cressie et Lahiri (1993) pour expliquer le meilleur comportement de REML par rapport à ML en terme de non biais.

Enfin, le système (3.75) appliqué une seule fois est formellement identique à celui des équations du MINQUE (Rao, 1971ab, LaMotte, 1970, 1973). Il montre en outre que l'estimateur REML peut s'interpréter aussi comme un estimateur dit MINQUE itéré pour lequel les estimations premières servent de poids a priori pour des estimations ultérieures et ainsi de suite (Searle, 1979, p6.7 ; Rao et Kleffe, 1988, p236).

Dans le cas général, on procédera comme pour ML, en utilisant le hessien de la logvraisemblance ou la matrice d'information de Fisher dans un algorithme de Newton-Raphson ou des scores de Fisher. Ces matrices ont pour expression (cf annexe II) :

$$\begin{aligned} -\frac{\partial^2 \mathrm{L}}{\partial \gamma_k \partial \gamma_l} = & \frac{1}{2} \mathrm{tr}\left(\underline{\mathbf{P}} \frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} \right) - \frac{1}{2} \mathrm{tr}\left(\underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_l} \right) \\ & - \frac{1}{2} \mathbf{y}' \underline{\mathbf{P}} \left(\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} - 2 \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_l} \right) \underline{\mathbf{P}} \mathbf{y} \end{aligned}, \tag{3.76}$$

$$\boxed{\mathrm{E}\left(-\frac{\partial^2 \mathrm{L}}{\partial \gamma_k \partial \gamma_l} \right) = \frac{1}{2} \mathrm{tr}\left(\underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_l} \right)}. \tag{3.77}$$

Comme pour ML, on montre aisément que le système des équations des scores de Fisher équivaut dans le cas linéaire au système (3.75) au coefficient ½ près.

La complémentarité des formules (3.76) et (3.77) a incité Gilmour et al (1995) à proposer pour les modèles linéaires mixtes, un algorithme de second d'ordre dit AI-REML basé sur la moyenne de ces deux matrices d'information soit

$$AI_{kl} = \frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\underline{\mathbf{P}}\mathbf{y}\,. \tag{3.78}$$

Cet algorithme est d'ailleurs appliqué dans le logiciel ASREML qui a été développé par les mêmes auteurs.

223. Interprétation bayésienne

C'est à Harville (1974) que l'on doit l'interprétation bayésienne de REML. Celle-ci repose sur le concept de vraisemblance marginale, cette fois au sens bayésien du terme (Dawid, 1980), comme outil d'élimination des paramètres parasites par intégration de ceux-ci. Dans le cas qui nous concerne, la vraisemblance marginale de $\boldsymbol{\gamma}$ se définit par:

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}) = \int \mathrm{p}(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\beta}\,, \tag{3.79}$$

où $\mathrm{d}\boldsymbol{\beta}$ est le symbole représentant $\mathrm{d}\beta_1\,\mathrm{d}\beta_2 ... \mathrm{d}\beta_p$.

L'intégrale en (86) peut se décomposer aussi en

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}) = \int \mathrm{p}(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\beta}\,, \tag{3.80}$$

où $\mathrm{p}(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})$ est la densité usuelle des observables sachant les paramètres et $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ est la densité a priori de $\boldsymbol{\beta} \in \mathbb{R}^p$ sachant $\boldsymbol{\gamma}$.

L'équivalence avec la vraisemblance résiduelle s'obtient en considérant une distribution uniforme pour cette dernière densité comme prouvé ci-dessous.

Dans le cadre du modèle $\mathbf{y} \sim \mathcal{N}(\underline{\mathbf{X}}\boldsymbol{\beta}, \mathbf{V})$, la densité $\mathrm{p}(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})$ s'écrit:

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}) = (2\pi)^{-N/2}|\mathbf{V}|^{-1/2}\exp\left[-(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})/2\right].$$

Or, $(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})$ peut se décomposer en (Gianola, Foulley et Fernando, 1986):

$$(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta}) = (\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), \tag{3.81}$$

où $\hat{\boldsymbol{\beta}}$ correspond à l'estimateur GLS de $\boldsymbol{\beta}$.

Le premier terme de cette décomposition ne dépend pas de $\boldsymbol{\beta}$ et l'intégration de cette partie par rapport à $\boldsymbol{\beta}$ est donc une constante qui se factorise, d'où

$$\begin{aligned}\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}) = {} & (2\pi)^{-N/2}|\mathbf{V}|^{-1/2}\exp\left[-(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})/2\right] \\ & \int \exp\left[-(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})/2\right]\mathrm{d}\boldsymbol{\beta}\end{aligned} \tag{3.82}$$

L'expression sous le signe «somme» est le noyau de $\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\gamma}$ qui est distribuée selon $\mathcal{N}\left[\hat{\boldsymbol{\beta}},\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-1}\right]$ ce qui implique que:

$$(2\pi)^{-p/2}\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|^{1/2}\int\exp\left[-\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)/2\right]\mathrm{d}\boldsymbol{\beta}=1.$$

L'intégrale en (3.82) est donc égale à $(2\pi)^{p/2}\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|^{-1/2}$ d'où l'expression de la densité marginale:

$$\boxed{\begin{aligned}\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma})=(2\pi)^{-(N-p)/2}|\mathbf{V}|^{-1/2}\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|^{-1/2}\\ \exp\left[-\left(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)/2\right]\end{aligned}},\qquad(3.83)$$

dont moins deux fois le logarithme est bien identique à (3.68a) avec une constante égale à $(N-p)\ln 2\pi$.

On en déduit donc que le REML de $\boldsymbol{\gamma}$, s'il existe, est le mode de la densité marginale de $\mathbf{y}$ (ou maximum de vraisemblance marginale de $\boldsymbol{\gamma}$). On montrerait de la même façon que c'est aussi le mode de la densité marginale a posteriori $\pi(\boldsymbol{\gamma}|\mathbf{y})$ de $\boldsymbol{\gamma}$ sous l'hypothèse additionnelle d'une densité uniforme de $\boldsymbol{\gamma}$.

En résumé:

$$\hat{\boldsymbol{\gamma}}_{REML}=\mathrm{argmax}_{\boldsymbol{\gamma}\in\Gamma}\ln\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}),\qquad(3.84a)$$

$$\hat{\boldsymbol{\gamma}}_{REML}=\mathrm{argmax}_{\boldsymbol{\gamma}\in\Gamma}\ln\pi(\boldsymbol{\gamma}|\mathbf{y}).\qquad(3.84b)$$

23. Aspects calculatoires

231. Algorithme «type-Henderson» et d'Harville

Sans entrer dans le détail des démonstrations, on montre que les algorithmes d'Henderson et d'Harville (38abc) relatifs au calcul des estimations ML des composantes de variance présentent des pendants REML de forme similaire soit:

$$\sigma_k^{2[t+1]}=\left\{\hat{\mathbf{u}}_k'\left(\boldsymbol{\eta}^{[t]}\right)\hat{\mathbf{u}}_k\left(\boldsymbol{\eta}^{[t]}\right)+\mathrm{tr}\left[\mathbf{C}_{kk}\left(\boldsymbol{\eta}^{[t]}\right)\right]\sigma_0^{2[t]}\right\}/q_k,\qquad(3.85a)$$

$$\sigma_0^{2[t+1]}=\left[\mathbf{y}'\mathbf{y}-\hat{\boldsymbol{\beta}}'\left(\boldsymbol{\eta}^{[t]}\right)\mathbf{X}'\mathbf{y}-\hat{\mathbf{u}}'\left(\boldsymbol{\eta}^{[t]}\right)\mathbf{Z}'\mathbf{y}\right]/\left[N-\mathrm{r}(\mathbf{X})\right]\qquad(3.85b)$$

et, pour l'algorithme d'Harville:

$$\sigma_k^{2[t+1]}=\left[\hat{\mathbf{u}}_k'\left(\boldsymbol{\eta}^{[t]}\right)\hat{\mathbf{u}}_k\left(\boldsymbol{\eta}^{[t]}\right)\right]/\left\{q_k-\mathrm{tr}\left[\mathbf{C}_{kk}\left(\boldsymbol{\eta}^{[t]}\right)\right]/\eta_k^{[t]}\right\},\qquad(3.85c)$$

où $\boldsymbol{\eta}^{[t]} = \left\{\sigma_k^{2[t]} / \sigma_0^{2[t]}\right\}$ est, comme précédemment, le vecteur des rapports de variance des K facteurs aléatoires à la variance résiduelle à l'itération n, $\hat{\mathbf{u}}_k\left(\boldsymbol{\eta}^{[t]}\right)$ est le BLUP de $\mathbf{u}_k$ conditionnellement à ces valeurs courantes des ratios de variance et $\mathbf{C}_{kk}$ est le bloc correspondant au facteur k dans l'inverse $\mathbf{C} = \begin{bmatrix} \underline{\mathbf{X}}'\underline{\mathbf{X}} & \underline{\mathbf{X}}'\mathbf{Z} \\ \mathbf{Z}'\underline{\mathbf{X}} & \mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1} \end{bmatrix}^{-1}$ de la matrice des coefficients des équations du modèle mixte d'Henderson (Henderson, 1973, 1984) (après factorisation de $1/\sigma_0^2$). Hormis cette différence portant sur la définition de $\mathbf{C}_{kk}$, les formules (3.85ac) restent inchangées. Il en est de même pour la variance résiduelle à la nuance importante près que, pour REML, $\left[N - \mathrm{r}(\mathbf{X})\right]$ se substitue à N au dénominateur de (3.85b).

232. Calcul de –2RL

Reprenons l'expression (3.68ab) de la logvraisemblance résiduelle soit, en reprenant la notation de Welham et Thompson:

$$-2\,\mathrm{RL} = \left[N - \mathrm{r}(X)\right]\ln 2\pi + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \qquad (3.86)$$

On a déjà montré (cf (3.33)) que :

$$\mathbf{y}'\underline{\mathbf{P}}\mathbf{y} = \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y},$$

où $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}'\right)'$ est la solution des équations dites du modèle mixte $\left(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right)\hat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{R}^{-1}\mathbf{y}$

avec $\mathbf{T} = (\underline{\mathbf{X}}, \mathbf{Z})$ et $\boldsymbol{\Sigma}^{-} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} \end{bmatrix}$.

Par ailleurs, les règles de calcul du déterminant d'une matrice partitionnée permettent d'établir que :

$$\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right| = \left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|. \qquad (3.87)$$

On a aussi montré (41) que:

$$\left|\mathbf{V}\right| = \left|\mathbf{R}\right|\left|\mathbf{G}\right|\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|.$$

d'où

$$\left|\mathbf{V}\right|\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| = \left|\mathbf{R}\right|\left|\mathbf{G}\right|\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right|. \qquad (3.88)$$

On en déduit le résultat général suivant, applicable à tout modèle linéaire gaussien de type $\mathbf{y} \sim \mathcal{N}(\underline{\mathbf{X}}\boldsymbol{\beta},\ \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})$:

$$\boxed{\begin{aligned}-2\,\mathrm{RL} = \left[N - \mathrm{r}(X)\right]\ln 2\pi + \ln\left|\mathbf{R}\right| + \ln\left|\mathbf{G}\right| + \ln\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right| \\ + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}\end{aligned}}. \qquad (3.89)$$

Comme pour la logvraisemblance profilée, cette formule permet de simplifier grandement le calcul de la logvraisemblance résiduelle notamment de son maximum grâce au recours aux équations du modèle mixte d'Henderson. Il suffit, pour calculer cet extremum, de remplacer dans (3.89), $\mathbf{R}$ et $\mathbf{G}$ par leurs estimations REML soit :

$$-2\,\mathrm{RL}_m = -2\,\mathrm{RL}(\mathbf{G} = \hat{\mathbf{G}}_{REML}, \mathbf{R} = \hat{\mathbf{R}}_{REML}).$$

Cette formule peut aussi se simplifier dans maintes situations par la prise en compte des structures particulières de $\mathbf{R}$ et de $\mathbf{G}$. Le seul terme susceptible de poser quelques difficultés de calcul est $\ln\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right|$. Celles-ci se résorbent en partie en ayant recours à une transformation de Cholesky $\mathbf{E}\mathbf{E}' = \mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}$ de la matrice des coefficients, si bien que $\ln\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right| = 2\sum_{j=1}^{rg(\mathbf{E})}\ln e_{jj}$ où les e_{jj} sont les termes diagonaux de $\mathbf{E}$.

24. Vraisemblance résiduelle et tests

241. Approximation de Kenward et Roger

Dans le cas où $\mathbf{V}$ dépend de paramètres inconnus $\boldsymbol{\gamma}$, la précision de $\hat{\boldsymbol{\beta}}$ est obtenue comme l'inverse de la matrice d'information de Fisher évaluée à la valeur estimée $\hat{\boldsymbol{\gamma}}$. Cette approche ignore l'incidence du bruit généré par les fluctuations d'échantillonnage de $\hat{\boldsymbol{\gamma}}$ si bien que la valeur de la précision qui en découle est surestimée (erreur-standard sous-estimée). En conséquence, les propriétés du test de Wald sont aussi affectées pour les petits échantillons. Comme les variances d'échantillonnage sont sous-estimées, les statistiques du test sont surévaluées et on a donc tendance à rejeter trop souvent l'hypothèse nulle (niveau effectif supérieur au niveau nominal ou P-value trop petite).

Kenward et Roger (1997) ont proposé récemment des ajustements de l'estimation de la précision et de la construction des tests relatifs aux effets fixes visant à améliorer leurs propriétés pour des petits échantillons. Pour ce faire, ils se placent délibérément dans le cadre d'un estimateur de $\boldsymbol{\beta}$ de type GLS où $\boldsymbol{\gamma}$ est remplacé par son estimation $\hat{\boldsymbol{\gamma}}_{REML}$.

Soit $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) = \left\{\mathbf{X}'\left[\mathbf{V}(\hat{\boldsymbol{\gamma}})\right]^{-1}\mathbf{X}\right\}^{-}$, l'estimateur GLS de $\boldsymbol{\beta}$ basé sur REML s'écrit:

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\mathbf{X}'\left[\mathbf{V}(\hat{\boldsymbol{\gamma}})\right]^{-1}\mathbf{y}. \qquad (3.90)$$

et sa variance d'échantillonnage :

$$\text{var}\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}})\right] = \text{var}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right] + \text{E}\left\{\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right]\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right]'\right\} \qquad (3.91)$$

Cette formule montre clairement que l'estimateur usuel $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) = \left[\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\gamma}})\mathbf{X}\right]^{-1}$ pose problème puisqu'à la fois $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ diffère du premier terme $\text{var}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right] = \boldsymbol{\Phi}(\boldsymbol{\gamma})$ (la différence $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) - \boldsymbol{\Phi}(\boldsymbol{\gamma})$ étant une matrice négative-définie) et que le second terme est ignoré.

Partant de l'expression ajustée, notée en bref $\boldsymbol{\Phi}_A(\boldsymbol{\gamma}) = \boldsymbol{\Phi}(\boldsymbol{\gamma}) + \boldsymbol{\Lambda}(\boldsymbol{\gamma})$, Kenward et Roger construisent un estimateur $\hat{\boldsymbol{\Phi}}_A$ de $\boldsymbol{\Phi}_A(\boldsymbol{\gamma})$ à partir de l'estimateur usuel $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ et d'un estimateur $\hat{\boldsymbol{\Lambda}}$ de la correction $\boldsymbol{\Lambda}$. Comme $\mathbf{E}\left[\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\right] \neq \boldsymbol{\Phi}(\boldsymbol{\gamma})$, il faut faire également un ajustement pour le biais $\mathbf{B} = \mathbf{E}\left[\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\right] - \boldsymbol{\Phi}(\boldsymbol{\gamma})$. Pour ce faire, Kenward et Roger procèdent comme Kackar et Harville (1984) en formant un développement limité de $\hat{\boldsymbol{\Phi}} = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ au second ordre au voisinage de la valeur vraie du paramètre soit

$$\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) \approx \boldsymbol{\Phi}(\boldsymbol{\gamma}) + \sum_{k=1}^{K}(\hat{\gamma}_k - \gamma_k)\frac{\partial\boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial\gamma_k} + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}(\hat{\gamma}_k - \gamma_k)(\hat{\gamma}_l - \gamma_l)\frac{\partial^2\boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial\gamma_k\partial\gamma_l}$$

conduisant à $\mathbf{B} \approx \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}W_{kl}\frac{\partial^2\boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial\gamma_k\partial\gamma_l}$

où W_{kl} est l'élément kl de $\mathbf{W} = \text{Var}(\hat{\boldsymbol{\gamma}})$ et,

$$\frac{\partial^2\boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial\gamma_k\partial\gamma_l} = \boldsymbol{\Phi}(\mathbf{P}_k\boldsymbol{\Phi}\mathbf{P}_l + \mathbf{P}_l\boldsymbol{\Phi}\mathbf{P}_k - \mathbf{Q}_{kl} - \mathbf{Q}_{lk} + \mathbf{R}_{kl})\boldsymbol{\Phi} \qquad (3.92)$$

avec $\mathbf{P}_k = \mathbf{X}'\frac{\partial\mathbf{V}^{-1}}{\partial\gamma_k}\mathbf{X}$, $\mathbf{Q}_{kl} = \mathbf{X}'\frac{\partial\mathbf{V}^{-1}}{\partial\gamma_k}\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\gamma_k}\mathbf{X}$ et $\mathbf{R}_{kl} = \mathbf{X}'\mathbf{V}^{-1}\frac{\partial^2\mathbf{V}(\boldsymbol{\gamma})}{\partial\gamma_k\partial\gamma_l}\mathbf{V}^{-1}\mathbf{X}$.

On peut procéder d'une façon similaire vis-à-vis de $\boldsymbol{\Lambda}$ en faisant un développement limité au premier ordre de $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}})$ autour de $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$, soit $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) \approx \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}) + \sum_{k=1}^{K}(\hat{\gamma}_k - \gamma_k)\partial\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})/\partial\gamma_k$.

Comme $\frac{\partial\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})}{\partial\gamma_k} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\frac{\partial\mathbf{V}^{-1}}{\partial\gamma_k}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ et $\text{var}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{V} - \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'$, on en déduit, à l'instar de Kackar et Harville (1984), que:

$$\boldsymbol{\Lambda} \approx \boldsymbol{\Phi}\left[\sum_{k=1}^{K}\sum_{l=1}^{K}W_{kl}\left(\mathbf{Q}_{kl} - \mathbf{P}_k\boldsymbol{\Phi}\mathbf{P}_l\right)\right]\boldsymbol{\Phi} \qquad (3.93)$$

Avec une structure linéaire de $\mathbf{V}$ telle que $\mathbf{V} = \sum_{k=1}^{K}\mathbf{V}_k\gamma_k$, les termes $\mathbf{R}_{kl}$ sont nuls et il vient $\mathbf{B} = -\boldsymbol{\Lambda}$ ce qui implique $\hat{\boldsymbol{\Phi}} = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) - \hat{\mathbf{B}}$; comme $\hat{\boldsymbol{\Phi}}_A = \hat{\boldsymbol{\Phi}} + \hat{\boldsymbol{\Lambda}}$, on aboutit en définitive à :

$$\boxed{\hat{\boldsymbol{\Phi}}_A = \boldsymbol{\Phi}\left(\hat{\boldsymbol{\gamma}}\right) + 2\hat{\boldsymbol{\Lambda}}}. \qquad (3.94)$$

Rappelons que $\mathbf{W}$ peut être approché par l'inverse $\mathbf{J}^{-1}\left(\boldsymbol{\gamma}\right)$ de la matrice d'information de Fisher soit avec REML: $\mathbf{J}\left(\boldsymbol{\gamma}\right) = 1/2\tilde{\mathbf{F}}\left(\boldsymbol{\gamma}\right) = \left\{1/2\,\mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{V}_l\right)\right\}$. Mais on peut également utiliser la matrice d'information observée (3.76) ou d'information moyenne (3.78). Des approximations similaires ont été développées par Monod (2000) dans le cadre de dispositifs «bloc-traitement» équilibrés de petite taille.

Soit à tester l'hypothèse $\mathrm{H}_0 : \mathbf{k}'\boldsymbol{\beta} = \mathbf{0}$ de rang r contre son alternative contraire H_1, Kenward et Roger proposent de bâtir une statistique de test de la forme

$$\boxed{F^* = \lambda F} \qquad (3.95)$$

où

- F est la statistique classique basée sur un pivot de Wald ($F = \hat{W}/r$ avec $\hat{W} = \hat{\boldsymbol{\beta}}'\mathbf{k}\left(\mathbf{k}'\hat{\boldsymbol{\Phi}}_A\mathbf{k}\right)^{-1}\mathbf{k}'\hat{\boldsymbol{\beta}}$) et qui prend en compte l'ajustement de la variance d'échantillonnage;

- λ un facteur d'échelle ($0 < \lambda \leq 1$) de la forme $\lambda = m/\left(m + r - 1\right)$ où m joue le rôle d'un nombre de degrés de liberté du dénominateur d'un F de Fisher-Snedecor.

Kenward et Roger déterminent m tel que F^* soit distribué approximativement sous l'hypothèse nulle comme un $\mathrm{F}\left(r, m\right)$; ils s'imposent de surcroît que ce soit une distribution exacte $\mathrm{F}\left(r, m\right)$ dans le cas où $\hat{W}$ est un T^2 d'Hotelling ou dans d'autres situations d'anova en dispositif équilibré.

Une situation typique relevant d'une statistique de Hotelling découle du test de l'hypothèse $\mathrm{H}_0 : \mathbf{k}'\boldsymbol{\mu} = \mathbf{0}$ sous le modèle multidimensionnel $\mathbf{Y}_i \sim_{\mathrm{iid}} \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; $i = 1, 2, ..., N$, (Rao, 1973, p564-565). La statistique de Hotelling s'écrit alors:

$$T^2 = \min_{H_0}\left(\bar{\mathbf{Y}} - \boldsymbol{\mu}\right)'\left(\mathbf{S}/N\right)^{-1}\left(\bar{\mathbf{Y}} - \boldsymbol{\mu}\right), \qquad (3.96)$$

où $\bar{\mathbf{Y}} = \left(\sum_{i=1}^{N} Y_i\right)/N$ et $\mathbf{S} = \left(N-1\right)^{-1}\sum_{i=1}^{N}\left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right)'$ sont les estimateurs usuels de $\boldsymbol{\mu}$ et de $\boldsymbol{\Sigma}$. Alors $F^* = \lambda T^2/r$ avec $\lambda = \left(N - r\right)/\left(N - 1\right)$ et l'on peut montrer qu'ici $T^2 = \hat{W} = \hat{\boldsymbol{\mu}}'\mathbf{k}\left[\mathbf{k}'\hat{\mathrm{V}}\left(\hat{\boldsymbol{\mu}}\right)\mathbf{k}\right]^{-1}\mathbf{k}'$ où $\hat{\boldsymbol{\mu}} = \bar{\mathbf{Y}}$ et $\hat{\mathrm{V}}\left(\hat{\boldsymbol{\mu}}\right) = \mathbf{S}/N$

Kenward et Roger donnent les valeurs suivantes de m et de λ à utiliser:

$$m = 4 + \frac{r+2}{r\rho - 1} \text{ et } \lambda = \frac{m}{E^*(m-2)} \tag{3.97}$$

où $\rho = V^* / 2E^{*2}$ (3.98a)

avec $E^* = (1 - A_2 / r)^{-1}$; $V^* = \frac{2}{r} \frac{1 + c_1 B}{(1 - c_2 B)^2 (1 - c_3 B)}$ (3.98b)

$$c_1 = g / [3r + 2(1-g)]; c_2 = (r-g) / [3r + 2(1-g)]$$
$$c_3 = (r + 2 - g) / [3r + 2(1-g)] \tag{3.98c}$$

pour $g = [(r+1)A_1 - (r+4)A_2] / [(r+2)A_2]$. (3.98d)

$$B = (A_1 + 6A_2) / 2r \tag{3.98e}$$

$$A_1 = \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \operatorname{tr}(\mathbf{\Theta\Phi P}_k \mathbf{\Phi}) \operatorname{tr}(\mathbf{\Theta\Phi P}_l \mathbf{\Phi}) \tag{3.98f}$$

$$A_2 = \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \operatorname{tr}(\mathbf{\Theta\Phi P}_k \mathbf{\Phi\Theta\Phi P}_l \mathbf{\Phi}) \tag{3.98g}$$

sachant que

$$\mathbf{\Theta} = \mathbf{k}(\mathbf{k}'\mathbf{\Phi k})^{-1} \mathbf{k}'. \tag{3.98h}$$

Dans le cas d'un seul contraste à tester ($r = 1$), λ vaut 1 et l'approximation de Kenward et Roger se ramène au carré d'un T de Student dont le nombre de degrés de liberté se calcule comme une variante de la méthode de Satterthwaite. Quoiqu'il en soit, l'approximation proposée conduit à une meilleure adéquation entre le niveau nominal et le niveau effectif que celle observée avec les tests de Wald et de type F non ajusté qui, appliqués à de petits échantillons, rejettent trop souvent l'hypothèse nulle (tests trop libéraux). Il est à remarquer que cette méthode est maintenant disponible dans la procédure Proc-mixed de SAS (version 8).

142. Approche de Welham et Thompson

Dans le cas de ML, le test des effets fixes dit du rapport de vraisemblance est basé sur la variation de $-2\,\mathrm{L}_m$ entre un modèle réduit et un modèle complet correspondant respectivement à l'hypothèse nulle H_0 et à la réunion $\mathrm{H}_0 \cup \mathrm{H}_1$ de celle-ci et de son alternative. Malheureusement, la transposition immédiate de cette technique à la logvraisemblance résiduelle $-2\,\mathrm{R\,L}_m$ n'a guère de sens puisque cela revient à contraster deux types d'ajustement des mêmes effets aléatoires mais qui utilisent des informations différentes:

$\mathbf{S}_0\mathbf{y}$ pour le modèle réduit $\mathrm{E}_R(\mathbf{y}) = \underline{\mathbf{X}}_0\boldsymbol{\beta}_0$ et $\mathbf{Sy}$ pour le modèle complet $\mathrm{E}_C(\mathbf{y}) = \underline{\mathbf{X}}\boldsymbol{\beta} = \underline{\mathbf{X}}_0\boldsymbol{\beta}_0 + \underline{\mathbf{X}}_1\boldsymbol{\beta}_1$ où $\mathbf{S}_0 = \mathbf{I}_N - \underline{\mathbf{X}}_0(\underline{\mathbf{X}}_0'\underline{\mathbf{X}}_0)^{-1}\underline{\mathbf{X}}_0$ et $\mathbf{S} = \mathbf{I}_N - \underline{\mathbf{X}}(\underline{\mathbf{X}}'\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}$.

Pour rendre le procédé cohérent, Welham et Thompson (1997) proposent de contraster les deux modèles sur la base d'une même projection en l'occurrence $\mathbf{S}_0\mathbf{y}$ (ou $\mathbf{K}_0'\mathbf{y}$ en écriture de plein rang) soit:

$$-2\,\mathrm{L}\left(\boldsymbol{\beta}_0, \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = \left(N - p_0\right)\ln 2\pi + \ln\left|\mathbf{K}_0'\mathbf{V}\mathbf{K}_0\right| + \left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}_0\boldsymbol{\beta}_0\right)'\left[\mathrm{Var}\left(\mathbf{K}_0'\mathbf{y}\right)\right]^{-1}\left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}_0\boldsymbol{\beta}_0\right)$$

et

$$-2\,\mathrm{L}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = \left(N - p_0\right)\ln 2\pi + \ln\left|\mathbf{K}_0'\mathbf{V}\mathbf{K}_0\right| + \left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}\boldsymbol{\beta}\right)'\left[\mathrm{Var}\left(\mathbf{K}_0'\mathbf{y}\right)\right]^{-1}\left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}\boldsymbol{\beta}\right)$$

où $p_0 = \mathrm{r}(\underline{\mathbf{X}}_0)$.

Comme $\mathbf{K}_0'\mathbf{X}_0 = \mathbf{0}$, la première expression est celle classique d'une vraisemblance résiduelle (cf 3.68ab) qu'on peut écrire sous la forme:

$$-2\,\mathrm{L}\left(\boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = C(\underline{\mathbf{X}}_0) + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0\right| + \left(\mathbf{y} - \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0\right), \qquad (3.99)$$

où $C(\underline{\mathbf{X}}_0)$ est une constante fonction de la matrice $\underline{\mathbf{X}}_0$ telle que définie en (75b) et $\hat{\boldsymbol{\beta}}_0$ l'estimateur GLS de $\boldsymbol{\beta}_0$.

En ce qui concerne la seconde expression, on remarque que $\mathbf{K}_0'\underline{\mathbf{X}}\boldsymbol{\beta} = \mathbf{K}_0'\underline{\mathbf{X}}_1\boldsymbol{\beta}_1$ et $\mathbf{K}_0\left(\mathbf{K}_0'\mathbf{V}\mathbf{K}_0\right)^{-1}\mathbf{K}_0' = \underline{\mathbf{P}}_0$ où $\underline{\mathbf{P}}_0 = \mathbf{V}^{-1}\left(\mathbf{I} - \mathbf{Q}_0\right)$, d'où

$$-2\,\mathrm{L}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = C(\underline{\mathbf{X}}_0) + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0\right| + \left(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1\right)'\underline{\mathbf{P}}_0\left(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1\right). \qquad (3.100)$$

Par ailleurs, $\max_{\boldsymbol{\beta},\boldsymbol{\gamma}}\mathrm{L}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = \max_{\boldsymbol{\gamma}}\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right]$ où $\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right]$ est la vraisemblance profilée $\mathrm{L}_P\left(\boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right)$ de $\boldsymbol{\gamma}$ basée sur $\mathbf{K}_0'\mathbf{y}$ et définie par:

$$-2\,\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right] = C(\underline{\mathbf{X}}_0) + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0\right| + \min_{\boldsymbol{\beta}_1}\left(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1\right)'\underline{\mathbf{P}}_0\left(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1\right)$$

Or, on peut montrer par manipulation matricielle que:

$$\min_{\boldsymbol{\beta}_1}\left(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1\right)'\underline{\mathbf{P}}_0\left(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1\right) = \min_{\boldsymbol{\beta}}\left(\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta}\right) = \left(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right), \quad (3.101)$$

où $\tilde{\boldsymbol{\beta}}$ est une solution du système GLS : $\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \underline{\mathbf{X}}'\mathbf{V}^{-1}\mathbf{y}$. En définitive:

$$-2\,\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right] = C(\underline{\mathbf{X}}_0) + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0\right| + \left(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right). \quad (3.102)$$

Welham et Thompson proposent la statistique A du logarithme du rapport de vraisemblance qui, mesure comme dans le cas classique, la variation de moins deux fois la

logvraisemblance maximum quand on passe du modèle réduit au modèle complet à partir, non plus de l'information sur $\mathbf{y}$, mais de celle sur $\mathbf{S}_0\mathbf{y}$, soit:

$$A = -2\max_{\gamma} \mathrm{L}\left(\boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) + 2\max_{\gamma} \mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right]. \tag{3.103a}$$

ou, encore

$$A = -2\,\mathrm{L}\left(\hat{\boldsymbol{\gamma}}; \mathbf{K}_0'\mathbf{y}\right) + 2\,\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\gamma}}), \tilde{\boldsymbol{\gamma}}; \mathbf{K}_0'\mathbf{y}\right], \tag{3.103b}$$

où $\tilde{\boldsymbol{\gamma}} = \arg\max_{\gamma} \mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right]$.

Si, à l'instar de Welham et Thompson, on introduit la notation suivante :

$$-2\,\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}_j\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{S}(\underline{\mathbf{X}}_i)\right] = C(\underline{\mathbf{X}}_i) + \ln\left|\underline{\mathbf{X}}_i'\mathbf{V}^{-1}\underline{\mathbf{X}}_i\right| + \ln|\mathbf{V}| + \left(\mathbf{y} - \underline{\mathbf{X}}_j\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}_j\boldsymbol{\beta}\right) \tag{3.104}$$

qui est celle d'une vraisemblance obtenue en ajustant le modèle $\mathrm{E}(\mathbf{y}) = \underline{\mathbf{X}}_j\boldsymbol{\beta}$, corrigée forfaitairement en fonction de l'information procurée par le projecteur $\mathbf{S}(\underline{\mathbf{X}}_i)$, la statistique A s'écrit comme

$$\boxed{A = -2\,\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0(\hat{\boldsymbol{\gamma}}), \hat{\boldsymbol{\gamma}}, \mathbf{S}(\underline{\mathbf{X}}_0)\right] + 2\,\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\gamma}}), \tilde{\boldsymbol{\gamma}}, \mathbf{S}(\underline{\mathbf{X}}_0)\right]}. \tag{3.105}$$

A la lumière de cette expression, on peut considérer la formule homologue obtenue en ajustant le modèle complet $\underline{\mathbf{X}}\boldsymbol{\beta}$ à partir du projecteur correspondant $\mathbf{S}(\underline{\mathbf{X}})$ soit

$$-2\,\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{S}(\underline{\mathbf{X}})\right] = C(\underline{\mathbf{X}}) + \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| + \ln|\mathbf{V}| + \left(\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta}\right)$$

puis, en passant au modèle réduit sur la base de l'expression (3.104) correspondante:

$$-2\,\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}_0\boldsymbol{\beta}_0, \boldsymbol{\gamma}, \mathbf{S}(\underline{\mathbf{X}})\right] = C(\underline{\mathbf{X}}) + \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| + \ln|\mathbf{V}| + \left(\mathbf{y} - \underline{\mathbf{X}}_0\boldsymbol{\beta}_0\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}_0\boldsymbol{\beta}_0\right)$$

conduisant à la statistique

$$\boxed{D = -2\,\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}_0\tilde{\tilde{\boldsymbol{\beta}}}_0(\tilde{\tilde{\boldsymbol{\gamma}}}), \tilde{\tilde{\boldsymbol{\gamma}}}, \mathbf{S}(\underline{\mathbf{X}})\right] + 2\,\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}), \hat{\boldsymbol{\gamma}}, \mathbf{S}(\underline{\mathbf{X}})\right]}, \tag{3.106}$$

où $\hat{\boldsymbol{\gamma}} = \arg\max_{\gamma} \mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}, \mathbf{S}(\underline{\mathbf{X}})\right]$ avec $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$ solution du système $\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \underline{\mathbf{X}}'\mathbf{V}^{-1}\mathbf{y}$ et, de façon similaire, $\tilde{\tilde{\boldsymbol{\gamma}}} = \arg\max_{\gamma} \mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}_0\tilde{\tilde{\boldsymbol{\beta}}}_0(\boldsymbol{\gamma}), \boldsymbol{\gamma}, \mathbf{S}(\underline{\mathbf{X}})\right]$ avec $\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0\tilde{\tilde{\boldsymbol{\beta}}}_0(\boldsymbol{\gamma}) = \underline{\mathbf{X}}_0'\mathbf{V}^{-1}\mathbf{y}$.

Il est important de noter que, dans le cas de la statistique D, $\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}_0\tilde{\tilde{\boldsymbol{\beta}}}_0(\tilde{\tilde{\boldsymbol{\gamma}}}), \tilde{\tilde{\boldsymbol{\gamma}}}, \mathbf{S}(\underline{\mathbf{X}})\right]$ n'a plus d'interprétation en terme de maximum d'une fonction classique de logvraisemblance obtenue en ajustant le modèle $\underline{\mathbf{X}}_0\boldsymbol{\beta}$ aux observations $\mathbf{K}'\mathbf{y}$ utilisant le projecteur $\mathbf{S}(\underline{\mathbf{X}})$. Contrairement à ce qui advenait avec A, cette statistique n'est donc pas le logarithme d'un

rapport de vraisemblances maximisées, mais seulement celui d'un rapport de vraisemblances profilées ajustées. Toutefois, au vu de résultats de simulation effectués sur des petits échantillons, Welham et Thompson concluent à de meilleures performances du test basé sur D par rapport à celles utilisant A et la statistique de Wald, et cela en terme d'approximation de ces statistiques à une loi Khi deux sous l'hypothèse nulle.

243. Tests des effets aléatoires

Le test de l'existence de certains effets aléatoires doit retenir l'attention car il pose des problèmes particuliers dans la théorie des tests de rapport de vraisemblance du fait que les paramètres spécifiés dans l'hypothèse nulle se trouvent à la frontière de l'espace paramétrique général. Cette question a été abordée d'un point de vue théorique par Self et Liang (1987) et son application au modèle linéaire mixte d'analyse de données longitudinales par Stram et Lee (1994, 1995). Un condensé des principaux résultats théoriques figure en annexe I.

Nous nous plaçons dans le cadre du modèle linéaire mixte gaussien $\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{Z}'\sigma_u^2 + \mathbf{I}_N\sigma_e^2\right)$ et considérons le test: $\mathrm{H}_0 : \sigma_u^2 = 0$ vs $\mathrm{H}_1 : \sigma_u^2 > 0$. La statistique du test du rapport de vraisemblance s'écrit alors: $\lambda = -2L_R + 2L_C$ où $L_R = \mathrm{Max}_{\sigma_e^2>0}\,\mathrm{RL}(\sigma_u^2 = 0, \sigma_e^2; \mathbf{y})$ et $L_C = \mathrm{Max}_{\sigma_u^2\geq 0, \sigma_e^2>0}\,\mathrm{RL}(\sigma_u^2, \sigma_e^2; \mathbf{y})$ si l'on utilise la fonction de vraisemblance résiduelle RL. L'utilisation de celle-ci se justifie parfaitement eu égard à la propriété de normalité asymptotique de l'estimateur REML qui a été formellement établie par Cressie et Lahiri (1993). Une statistique homologue basée sur la vraisemblance classique $\mathrm{L}(\boldsymbol{\beta}; \sigma_u^2, \sigma_e^2; \mathbf{y})$ est également envisageable même si celle-ci s'avère en pratique moins efficace (Morell, 1998).

L'utilisation usuelle de ce test se réfère alors à une distribution asymptotique de λ sous H_0 qui est une loi de Khi-deux à 1 degré de liberté. Cette assertion est inexacte et cela pour la simple raison de bon sens suivante. En effet, il est fort possible que sous le modèle complet (C: $\sigma_u^2 \geq 0, \sigma_e^2 > 0$), l'estimateur REML de σ_u^2 soit nul ($\hat{\sigma}_u^2 = 0$) si bien que $L_C = L_R$ et $\lambda = 0$. Sous H_0, un tel événement survient asymptotiquement une fois sur deux du fait de la propriété de normalité asymptotique de l'estimateur non contraint de σ_u^2 autour de sa valeur centrale nulle. La distribution asymptotique correcte à laquelle il faut se référer sous H_0 est donc celle d'un mélange en proportions égales, d'une loi de Dirac en zéro (D_0 notée aussi quelquefois χ_0^2) et d'une loi de Khi-deux à un degré de liberté (χ_1^2) soit en abrégé:

$$\boxed{\lambda \underset{\mathcal{L}}{\rightarrow} 1/2\,\mathrm{D}_0 + 1/2\,\chi_1^2}. \qquad (3.107)$$

En conséquence, le test «naïf» est trop conservateur et le seuil s du test correct au niveau α correspond à:

$$\Pr(\chi_1^2 \geq s) = 2\alpha \qquad (3.108)$$

puisque, sous H_0, la décision de rejet est prise lorsque la statistique est positive (une fois sur deux) et que celle-ci, alors de loi de Khi-deux à un degré de liberté, dépasse le seuil s. En définitive, la procédure correcte revient à effectuer un test unilatéral au lieu d'un test bilatéral en utilisant le rapport de vraisemblance.

Ce résultat se généralise au test $\mathrm{H}_0 : \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & 0 \end{pmatrix}$ vs $\mathrm{H}_1 : \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$, cette dernière hypothèse correspondant au modèle $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}$ où $\mathrm{var}\left(\mathbf{u}_1', \mathbf{u}_2'\right)' = \boldsymbol{\Sigma} \otimes \mathbf{I}_q$. Ce modèle se rencontre dans l'analyse de données longitudinales (Laird et Ware, 1982 ; Diggle et al, 1994).

Si l'on contraint $\boldsymbol{\Sigma}$ sous H_1 à être définie semi-positive, alors, (Stram et Lee , 1994)

$$\lambda \underset{\mathcal{L}}{\rightarrow} 1/2\,\chi_1^2 + 1/2\,\chi_2^2 \qquad (3.109).$$

De la même façon, on généralise ensuite au cas du test $\mathrm{H}_0 : \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11(q\,x\,q)} & 0 \\ 0 & 0 \end{pmatrix}$ vs $\mathrm{H}_1 : \boldsymbol{\Sigma}_{(q+1)\mathrm{x}(q+1)} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \sigma_{22} \end{pmatrix}$ définie semi-positive pour lequel

$$\lambda \underset{\mathcal{L}}{\rightarrow} 1/2\,\chi_q^2 + 1/2\,\chi_{q+1}^2. \qquad (3.110)$$

Discussion-Conclusion

La théorie de la vraisemblance qui, à la suite de Fisher, est devenue le paradigme central de la statistique inférentielle paramétrique trouve dans le modèle linéaire une de ses applications les plus démonstratives. Il est apparu également que les techniques ML et REML avaient des liens profonds avec la théorie du BLUP et les équations du modèle mixte d'Henderson (Henderson et al, 1959 ; Henderson, 1973, 1984 ; Goffinet, 1983), relations qui s'explicitent clairement grâce à la théorie EM (Dempster et al, 1977, McLachlan and Krishnan, 1997). Ce relais permet de développer des algorithmes de calcul performants applicables à des échantillons de grande taille et des dispositifs déséquilibrés et relativement complexes.

Développée au départ pour estimer les poids à affecter à l'information intra et inter blocs dans l'analyse en blocs incomplets déséquilibrés, la méthode REML s'est avérée rapidement comme un passage obligé et une référence dans l'inférence des composantes de la variance en modèle linéaire mixte au point qu'elle a supplanté en pratique les estimateurs quadratiques d'Henderson (1953) et du MINQUE (Rao et Kleffe, 1988 ; LaMotte, 1970, 1973). Cette place privilégiée de REML a été d'autant mieux affirmée et acceptée que les interprétations qu'on pouvait en faire (vraisemblance de contrastes d'erreur, inférence conditionnelle, vraisemblance marginalisée par rapport aux effets fixes, MINQUE itéré) se révélaient diverses et complémentaires enrichissant ainsi la compréhension de la méthode. A la lumière des travaux récents de Kenward et Roger (1997) ainsi que de Welham et Thompson (1997), on peut gager que la place qu'occupe REML va dépasser le cadre strict de l'estimation des composantes de la variance pour intervenir également dans l'inférence des effets fixes.

L'offre logicielle est relativement abondante (SAS Proc-Mixed, ASREML, Splus) et permet de traiter un grand éventail de structures de variances covariances avec accès aussi bien à ML qu'à REML. Ces logiciels généralistes s'appuient sur des algorithmes de second ordre (Newton Raphson, Fisher ou information moyenne) de convergence rapide. Toutefois, comme le notait récemment Thompson (2002) lui-même lors d'une comparaison de ces différents algorithmes, les techniques EM se montrent en constant progrès ; elles s'avèrent aussi plus fiables et quasi incontournables dans certaines situations ou avec certains modèles (van Dyk, 2000 ; Delmas et al, 2002).

La disponibilité des logiciels explique pour une grande part le succès grandissant du modèle mixte et des méthodes du maximum de vraisemblance auprès des utilisateurs et l'on ne saurait que s'en féliciter. Celui-ci d'ailleurs ne pourra aller que grandissant eu égard à l'ampleur du domaine d'application du modèle mixte; ses extensions au modèle linéaire généralisé (Mc Cullagh et Nelder, 1989) et au modèle non linéaire (Davidian et Giltinian, 1995) le prouvent à l'évidence. On a pu aussi montrer que maintes techniques particulières pouvaient faire l'objet d'une interprétation en terme de modèle mixte; on peut citer par exemple le krigeage, le filtre de Kalman (Robinson, 1991) l'ajustement par splines (Verbyla et al, 1999) et l'hétérogénéité de variance (Foulley et Quaas, 1995 ; San Cristobal, Robert-Granié et Foulley, 2002) ; cette vision unificatrice ne peut qu'enrichir l'ensemble et stimuler l'esprit de tous.

REFERENCES

Anderson R.L., Bancroft T.A. (1952), *Statistical theory in research*. Mc Graw-Hill, New-York

Berger J.O., Liseo B., Wolpert R.L. (1999), Integrated Likelihood methods for eliminating nuisance parameters, Statistical Science, 14, 1-28

Cox D.R., Reid N. (1987), Parameter orthogonality and approximate conditional inference, Journal of the Royal Statistical Society B, 49, 1-39

Cox D.R., Hinkley D.V. (1974), *Theoretical statistics*, Chapman & Hall, London

Cressie N., Lahiri S.N. (1993), The asymptotic distribution of REML estimators, Journal of Multivariate Analysis, 45, 217-233

Crump S.L. (1947), The estimation of components of variance in multiple classifications, PhD thesis, Iowa State University, Ames

Davidian M., Giltinian D.M. (1995), *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London

Dawid A.P. (1980), A Bayesian look at nuisance parameters. Proceedings of the first international meeting held in Valencia. (Bernardo J.M., DeGroot M.H., Lindley D.V., Smith A.F.M., eds) University Press, Valencia, Spain, 167-184.

Delmas C., Foulley J.L., Robert-Granié C. (2002), Further insights into tests of variance components and model selection, Proceedings of the 7th World Congress of Genetics applied to Livestock Production, Montpellier, France, 19-23 August 2002.

Diggle P.J., Liang K.Y., Zeger S.L. (1994), *Analysis of longitudinal data*, Oxford Science Publications, Clarendon Press, Oxford

Edwards A.W.F. (1972), *Likelihood*, Cambridge University Press, Cambridge

Eisenhart C. (1947), The assumptions underlying the analysis of variance. Biometrics 3, 1-21

Fisher R.A. (1922), On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London Series A 222, 309-368

Fisher R.A. (1925), *Statistical methods for research workers*, Oliver and Boyd. Edinburgh and London

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

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

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

Gianola D., Foulley J.L, Fernando R. (1986), Prediction of breeding values when variances are not known, Genetics Selection Evolution, 18, 485-498

Gilmour A.R., Thompson R., Cullis B.R. (1995), An efficient algorithm for REML estimation in linear mixed models, Biometrics 51, 1440-1450

Goffinet B. (1983), Risque quadratique et sélection : quelques résultats appliqués à la sélection animale et végétale, Thèse de Docteur Ingénieur, Université Paul Sabatier, Toulouse.

Gourieroux C., Montfort A. (1989), *Statistique et modèles économétriques*, Economica, France.

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

Harvey W.R. (1970), Estimation of variance and covariance components in the mixed model, Biometrics 61, 485-504

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

Harville D.A. (1997) *Matrix algebra from a statistician's perspective*, Springer, Berlin

Harville D.A., Callanan T.P. (1990), Computational aspects of likelihood based inference for variance components, Advances in Statistical Methods for Genetic Improvement of Livestock (Gianola D, Hammond K, eds), Springer Verlag, 136-176

Henderson C.R. (1953), Estimation of variance and covariance components. Biometrics 9, 226-252

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

Henderson C.R., Kempthorne O., Searle S.R., von Krosigk C.N. (1959), Estimation of environmental and genetic trends from records subject to culling, Biometrics 13, 192-218

Kackar A.N., Harville D.A. (1984), Approximation for standard errors of estimators of fixed and random effects in mixed linear models, Journal of the American Statistical Association, 79, 853-862

Kalbfleisch J.D., Sprott D.A. (1970), Application of the likelihood methods to models involving large numbers of parameters. Journal of the Royal Statistical Society B 32, 175-208

Kenward M.G., Roger J.H. (1997), Small sample inference for fixed effects from restricted maximum likelihood, Biometrics 53, 983-997

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

LaMotte L.R. (1970), A class of estimators of variance components, Technical report 10, Department of Statistics, University of Kentucky, Lexington, KE

LaMotte L.R. (1973), Quadratic estimation of variance components, Biometrics 29, 311-330

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

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

Mardia K.V., Marshall R.J. (1985), Maximum likelihood estimation of models for residual covariance in spatial regression, Biometrika 71, 135-146

McCullagh P., Nelder J. (1989), *Generalized linear models*, 2nd edition, Chapman and Hall, London

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

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

Monod H. (2000), On the efficiency of generally balanced designs analysed by restricted maximum likelihood, Proceedings of Optimum Design, Cardiff, 12-14 april 2000, Kluwer Academic

Mood A.M., Graybill F.A., Boes D.C. (1974), *Introduction to the theory of statistics*, Third edition, International student edition

Morrell C.H. (1998), Likelihood ratio of variance components in the linear mixed-effects model using restricted maximum likelihood, Biometrics 54, 1560-1568

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

Quaas R.L. (1992), *REML Notebook*, Mimeo, Cornell University, Ithaca, New York.

Rao C.R. (1971a), Estimation of variance components-Minque theory, Journal of Multivariate Analysis, 1, 257-275

Rao C.R. (1971b), Minimum variance quadratic unbiased estimation of variance components. Journal of Multivariate Analysis, 1, 445-456

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

Rao C.R. (1979), MIQE theory and its relation to ML and MML estimation of variance components, Sankhya B 41, 138-153

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

Robinson G.K. (1991), The estimation of random effects. Statistical Science, 6, 15-51

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, à paraître.

Searle S.R. (1979), *Notes on variance component estimation. A detailed account of maximum likelihood and kindred methodology*, Paper BU-673-M, Cornell University, Ithaca

Searle SR (1982*) Matrix algebra useful for statistics*. John Wiley and Sons, New York

Searle S.R. (1989), Variance components-some history and a summary account of estimation methods, Journal of Animal Breeding and Genetics, 106, 1-29

Searle S.R., Casella G., Mc Culloch C.E. (1992), *Variance components*, J Wiley and Sons, New-York

Self S.G., Liang K.Y. (1987), Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions, Journal of the American Statistical Association, 82, 605-610

Stram D.O., Lee J.W. (1994), Variance components testing in the longitudinal mixed effects model, Biometrics 50, 1171-1177

Stram D.O., Lee J.W. (1995), Correction to "Variance components testing in the longitudinal mixed effects model", Biometrics 51, 1196

Sweeting T.J. (1980), Uniform asymptotic normality of the maximum likelihood estimator, The Annals of Statistics, 8, 1375-1381

Thompson W.A. (1962), The problem of negative estimates of variance components, Annals of Mathematical Statistics, 33, 273-289

Thompson R. (1989), REML, Biometric Bulletin, 6, (3), 4-5.

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

Verbeke G., Molenberghs G. (2000), *Linear mixed models for longitudinal data*, Springer Verlag, New York

Verbyla A.P., Cullis B.R., Kenward M.G., Welham S.J. (1999), The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion), Applied Statistics, 48, 269-311

Welham S.J., Thompson R. (1997), A likelihood ratio test for fixed model terms using residual maximum likelihood, Journal of the Royal Statistical Society B, 59, 701-714

1) Optimisation avec contraintes

De manière générale supposons que $\hat{\alpha}$ est un minimum local de $-2L(\alpha; y)$ sur un espace paramétrique Γ contraint par un ensemble d'égalités et d'inégalités:

$$\Gamma = \{\alpha \in \mathbb{R}^n : g(\alpha) \leq 0; h(\alpha) = 0\} \tag{1}$$

avec $g : \mathbb{R}^n \to \mathbb{R}^p$ et $h : \mathbb{R}^n \to \mathbb{R}^q$. Supposons que g, h et L sont suffisamment régulières et que les contraintes ne sont pas redondantes.

Alors il existe des nombres $\lambda_i \geq 0$ pour $i = 1$ à p et des $\mu_j \in \mathbb{R}$ pour $j = 1$ à q tels que:

$$\begin{cases} \nabla(-2L(\hat{\alpha}; y)) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\hat{\alpha}) + \sum_{j=1}^{q} \mu_j \nabla h_j(\hat{\alpha}) = 0 \\ \lambda_i g_i(\hat{\alpha}) = 0 \ \forall i = 1 \text{ à } p \end{cases} \tag{2}$$

Ce sont les conditions de Karush, Kühn, Tucker qui des conditions nécessaires d'optimalité qu'il convient de résoudre pour obtenir le minimum local $\hat{\alpha}$ lorsqu'il existe.

2) Test du rapport de vraisemblance

On suppose que Γ est convexe fermé et que différentes valeurs de α correspondent à différentes lois de probabilité. On s'intéresse au test du rapport de vraisemblance de l'hypothèse nulle $\alpha \in \Gamma_0$ contre l'hypothèse alternative $\alpha \in \Gamma \setminus \Gamma_0$ où Γ_0 est un sous ensemble de Γ. On note α_0 la vraie valeur du paramètre sous l'hypothèse nulle et $I(\alpha_0)$ la matrice d'information de Fisher supposée définie positive. On suppose que α_0 est sur la frontière de Γ. On rappelle qu'un cône de sommet α_0, C, est l'ensemble des points tels que si $x \in C$ alors $a(x - \alpha_0) + \alpha_0 \in C$ où $a \geq 0$. On suppose que Γ et Γ_0 sont suffisamment réguliers pour être approchés par des cônes de sommet α_0, C_Γ et C_{Γ_0} respectivement. C'est-à-dire que (cf. Chernoff (1954) et Self et Liang (1987)):

$$\inf_{x \in C_\Gamma} \|x - y\| = o(\|y - \alpha_0\|) \ \forall y \in \Gamma$$

$$\inf_{y \in \Gamma} \|x - y\| = o(\|x - \alpha_0\|) \ \forall x \in C_\Gamma$$

On obtient des conditions analogues pour Γ_0. Sous des conditions faibles de régularité de $L(\alpha; y)$ (cf. Self et Liang (1987)), on peut montrer que la loi asymptotique de la statistique de test du rapport de vraisemblance $2(L(\hat{\alpha}) - L(\hat{\alpha_0}))$ est la même que:

$$\sup_{\alpha \in (C_\Gamma - \alpha_0)} [-(Z-\alpha)^T I(\alpha_0)(Z-\alpha)] - \sup_{\alpha \in (C_{\Gamma_0} - \alpha_0)} [-(Z-\alpha)^T I(\alpha_0)(Z-\alpha)] \tag{3}$$

où Z suit une loi normale multivariée de moyenne 0 et de variance $I^{-1}(\alpha_0)$ et $C_\Gamma - \alpha_0$ désigne la translation du cône C_Γ de sommet α_0 de sorte qu'il soit de sommet 0. Ce qui se reécrit également:

$$\inf_{\alpha \in \tilde{C}_0} \|\tilde{Z} - \alpha\|^2 - \inf_{\alpha \in \tilde{C}} \|\tilde{Z} - \alpha\|^2 \tag{4}$$

où:

$$\tilde{C} = \{\tilde{\alpha} : \tilde{\alpha} = \Lambda^{1/2} P^T \alpha \ , \forall \alpha \in C_\Gamma - \alpha_0\}$$

$$\tilde{C}_0 = \{\tilde{\alpha} : \tilde{\alpha} = \Lambda^{1/2} P^T \alpha \ , \forall \alpha \in C_{\Gamma_0} - \alpha_0\}$$

et $\tilde{Z}$ suit une loi normale multivariée centrée de variance identité. $P\Lambda P^T$ est la décomposition spectrale de $I(\alpha_0)$.

3) Application au modèle mixte

On se place dans le cadre du modèle mixte à deux effets aléatoires

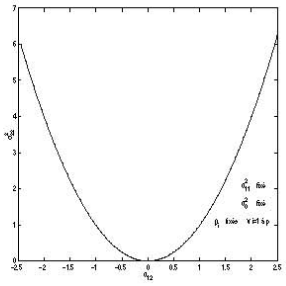
$$Y = X\beta + Z_1 u_1 + Z_2 u_2 + \epsilon \tag{5}$$

où X, Z_1 et Z_2 sont des matrices d'incidence connues; β est le vecteur des effets fixes inconnu; u_1 et u_2 sont les deux vecteurs des effets aléatoires inconnus et ϵ est le vecteur des erreurs résiduelles. On suppose que $(u_1, u_2)^T$ est un vecteur gaussien centré de variance

$$Var \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \sigma_{11}^2 I_q & \sigma_{12} I_q \\ \sigma_{12} I_q & \sigma_{22}^2 I_q \end{pmatrix} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix} \bigotimes I_q = \Sigma \bigotimes I_q \tag{6}$$

On suppose que ϵ est gaussien centré de variance $\sigma_\epsilon^2 I_N$ et indépendant de $(u_1, u_2)^T$. On s'intéresse alors au test de l'hypothèse nulle $\mathcal{H}_0, \Sigma = \begin{pmatrix} \sigma_{11}^2 & 0 \\ 0 & 0 \end{pmatrix}$ contre l'hypothèse alternative $\mathcal{H}_1, \Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix}$

avec $\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2 \geq 0$ et $\sigma_{22}^2 \neq 0$. On pose $\alpha = [\sigma_{22}^2, \sigma_{12}, \sigma_{11}^2, \sigma_0^2, \beta_1, \cdots, \beta_p]^T$. L'espace complet des paramètres est contraint par $\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2 \geq 0$. Sous l'hypothèse nulle, $\sigma_{22}^2 = 0$ et $\sigma_{12} = 0$, la vraie valeur du paramètre, α_0, est notée $[0, 0, \sigma_{11,0}^2, \sigma_{0,0}^2, \beta_{1,0}, \cdots, \beta_{p,0}]^T$. On suppose que $\sigma_{11,0}^2$ et $\sigma_{0,0}^2$ sont strictement positives. α_0 se trouve alors en bordure de l'espace paramétrique.



Au point α_0, l'espace complet des paramètres peut être approché par le cône C de sommet α_0 tel que $C - \alpha_0 = [0, +\infty[\times \mathbb{R}^{p+3}$. L'espace réduit des paramètres peut être approché au point α_0 par le cône C_0, de sommet α_0 tel que $C_0 - \alpha_0 = \{0\} \times \{0\} \times \mathbb{R}^{p+2}$. On se trouve alors dans le cas 6 de l'article de Self et Liang (1987) qui nous dit que la loi asymptotique de la statistique de test est $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$. Ceci s'obtient aisément à partir des éléments donnés dans les sections 1 et 2 précédentes. En effet il suffit de remarquer qu'il s'agit dans un premier cas de minimiser $||\tilde{Z} - \alpha||^2$ sans contrainte et dans un second cas de minimiser $||\tilde{Z} - \alpha||^2$ sous la contrainte $\alpha_1 \geq 0$. On obtient alors:

$$\inf_{\alpha \in \tilde{C}_0} ||\tilde{Z} - \alpha||^2 - \inf_{\alpha \in \tilde{C}} ||\tilde{Z} - \alpha||^2 = \tilde{Z}_1^2 I_{\tilde{Z}_1 > 0} + \tilde{Z}_2^2 \tag{7}$$

qui suit une loi $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$. Ce résultat se généralise au test de q contre $q+1$ effets aléatoires pour lequel la loi asymptotique de la statistique de test est $\frac{1}{2}\chi_q^2 + \frac{1}{2}\chi_{q+1}^2$ sous l'hypothèse que les q premiers effets aléatoires sont linéairement indépendants sous l'hypothèse nulle.

ANNEXE II

Matrices d'information

1. Estimation ML

Le point de départ est l'expression de la logvraisemblance sous la forme

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = N\ln(2\pi) + \ln|\mathbf{V}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \qquad \text{(II.1)}$$

où $l(\boldsymbol{\beta},\boldsymbol{\gamma}) = -2\,\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) = -2\ln \mathrm{p}_Y(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})$.

Nous avons vu que les dérivées premières s'écrivent :

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \qquad \text{(II.2)}$$

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \ . \qquad \text{(II.3)}$$

On en déduit l'expression des dérivées partielles secondes

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \qquad \text{(II.4)}$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}\partial \gamma_k} = 2\mathbf{X}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \qquad \text{(II.5)}$$

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} &= \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l}\right) - \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right) \\ &-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\left(\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} - 2\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right)\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\end{aligned} \qquad \text{(II.6)}$$

En divisant par deux, ces formules fournissent les termes qui permettent de calculer la matrice d'information dite observée $\mathbf{I}(\hat{\boldsymbol{\alpha}};\mathbf{y}) = -\left.\frac{\partial^2 \mathrm{L}(\boldsymbol{\alpha};\mathbf{y})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}$ où $\boldsymbol{\alpha} = (\boldsymbol{\beta}',\boldsymbol{\gamma}')'$ qui interviennent par exemple, dans l'algorithme de Newton-Raphson.

En prenant l'espérance de $\mathbf{I}(\boldsymbol{\alpha};\mathbf{y})$, on obtient les termes de la matrice d'information de Fisher $\mathbf{J}(\boldsymbol{\alpha}) = \mathrm{E}[\mathbf{I}(\boldsymbol{\alpha};\mathbf{y})]$ soit :

$$\mathbf{J}_{\beta\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \qquad \text{(II.7)}$$

$$\mathbf{J}_{\beta\gamma} = \mathbf{0}, \qquad \text{(II.8)}$$

$$\left(\mathbf{J}_{\gamma\gamma}\right)_{kl} = \frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \qquad \text{(II.9)}$$

Deux remarques importantes méritent d'être formulées à ce stade. Premièrement, les estimations ML de $\boldsymbol{\beta}$ et de $\boldsymbol{\gamma}$ sont asymptotiquement non corrélées. Deuxièmement, les formules (7-8-9) s'appliquent aussi bien aux modèles linéaires qu'aux modèles non linéaires en $\mathbf{V}$ ce qui n'est pas le cas pour $\mathbf{I}(\boldsymbol{\alpha};\mathbf{y})$.

2. Estimation REML

La logvraisemblance résiduelle s'écrit

$$r(\boldsymbol{\gamma}) = [N - \mathrm{r}(\mathbf{X})]\ln 2\pi + \ln|\mathbf{V}| + \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \qquad \text{[II.10]}$$

où $r(\boldsymbol{\gamma}) = -2\,\mathrm{L}\left(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}\right)$.

En différenciant par rapport à γ_k, on obtient :

$$\frac{\partial r(\boldsymbol{\gamma})}{\partial \gamma_k} = \frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} + \frac{\partial \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial \gamma_k} + \mathbf{y}'\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k}\mathbf{y} \qquad \text{[II.11]}$$

Or,

$$\frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} = tr\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right)$$

$$\frac{\partial \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial \gamma_k} = -\mathrm{tr}\left[\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\underline{\mathbf{X}}\right] = -\mathrm{tr}\left[\mathbf{V}^{-1}\underline{\mathbf{X}}\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right],$$

ce qui permet de faire apparaître et de factoriser la matrice $\underline{\mathbf{P}}$, soit

$$\frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} + \frac{\partial \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial \gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) \qquad \text{[II.12]}$$

Il reste à expliciter $\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k}$. Par définition, $\mathbf{V}\underline{\mathbf{P}} = \left(\mathbf{I} - \mathbf{Q}\right)$ avec $\mathbf{Q} = \underline{\mathbf{X}}\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}$. Par dérivation de cette expression, on a :

$$\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}} + \mathbf{V}\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\frac{\partial \mathbf{Q}}{\partial \gamma_k}. \qquad \text{[II.13]}$$

Or, la dérivée de l'expression explicite de $\mathbf{Q}$ conduit à :

$$\frac{\partial \mathbf{Q}}{\partial \gamma_k} = -\mathbf{Q}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}},$$

d'où, en remplaçant dans [II.13], $\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\mathbf{V}^{-1}\left(\mathbf{I} - \mathbf{Q}\right)\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}$, c'est-à-dire

$$\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}. \qquad \text{[II.14]}$$

Il s'en suit l'expression suivante du score :

$$\frac{\partial r(\boldsymbol{\gamma})}{\partial \gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - \mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\mathbf{y}. \qquad \text{[II.15].}$$

On vérifie bien au passage que l'espérance du score est nulle puisque

$$\mathrm{E}\left(\frac{\partial r(\boldsymbol{\gamma})}{\partial \gamma_k}\right) = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - \mathrm{tr}\left[\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\,\mathrm{E}(\mathbf{y}\mathbf{y}')\right]$$

Or, $\mathrm{E}(\mathbf{y}\mathbf{y}') = \underline{\mathbf{X}}\boldsymbol{\beta}\boldsymbol{\beta}'\underline{\mathbf{X}}' + \mathbf{V}$. Comme $\underline{\mathbf{P}}\underline{\mathbf{X}} = \mathbf{0}$ et $\underline{\mathbf{P}}\mathbf{V}\underline{\mathbf{P}} = \underline{\mathbf{P}}$, le deuxième terme est égal au premier, QED.

En dérivant à nouveau terme à terme [II.15], on obtient l'expression du hessien qui peut s'écrire sous une forme similaire à celle présentée en [II.6] avec ML, soit :

$$\begin{aligned}\frac{\partial^2 r(\boldsymbol{\gamma};\mathbf{y})}{\partial \gamma_k \partial \gamma_l} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l}\right) - \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right) \\ -\mathbf{y}'\underline{\mathbf{P}}\left(\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} - 2\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right)\underline{\mathbf{P}}\mathbf{y}\end{aligned}. \qquad \text{[II.16]}$$

La matrice d'information de Fisher s'en déduit immédiatement

$$\left(\mathbf{J}_{\gamma\gamma}\right)_{kl} = \frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \qquad \text{[II.17]}$$

ANNEXE III

Calcul de $|\mathbf{V}|$

On considère la partition suivante:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{R}^{-1} & \mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix},$$

alors on sait que: (cf par ex Searle, 1982, Ch 10, page 257-271

$$|\mathbf{A}| = |\mathbf{A}_{11}|\left|\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right|$$

Ici, $|\mathbf{A}| = \left|\mathbf{R}^{-1}\right|\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\right| = 1/|\mathbf{R}||\mathbf{G}|$.

De même, par symétrie

$$|\mathbf{A}| = \left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|\left|\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\right|$$

$$\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right| / |\mathbf{V}|$$

d'où $\boxed{|\mathbf{V}| = |\mathbf{R}||\mathbf{G}|\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|}$

Chapitre IV : Algorithme EM

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 $\mathrm{L}(\boldsymbol{\phi};\mathbf{y}) = \ln \mathrm{p}(\mathbf{y} \mid \boldsymbol{\phi})$ 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

$$\mathrm{L} = \sum_{j=1}^{4} y_j \ln \pi_j + Cste\,, \qquad (1)$$

où π_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 $\mathbf{S} = \{s_k = \partial \mathrm{L} / \partial \phi_k\}$

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

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

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

où $N = \sum_{j=1}^{4} y_j$.

Comme les π_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.

A 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:

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

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 $x_{AA}^{\#} = \frac{p^2}{p^2 + 2pr} y_A$ et $x_{AO}^{\#} = \frac{2pr}{p^2 + 2pr} y_A$ ou après simplification:

$$x_{AA}^{\#} = \frac{p}{p + 2r} y_A \; ; \; x_{AO}^{\#} = \frac{2r}{p + 2r} y_A \, . \qquad (5)$$

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

$$p'' = \left(2x_{AA}^{\#} + y_{AB} + x_{AO}^{\#}\right)/2N \; ; \; q'' = \left(2x_{BB}^{\#} + y_{AB} + x_{BO}^{\#}\right)/2N \qquad (6)$$

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 $\hat{p} = 0.251560$, $\hat{q} = 0.050012$ et $\hat{r} = 0.698428$. 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[7], 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):

[7] ou variables latentes, supplémentaires ou cachées selon les circonstances et les auteurs

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

formule qui traduit simplement le fait que la dérivée de la logvraisemblance $L(\boldsymbol{\phi};\mathbf{y}) = \ln g(\mathbf{y}|\,\boldsymbol{\phi})$ de $\boldsymbol{\phi}$ basée sur $\mathbf{y}$ par rapport au paramètre est l'espérance conditionnelle de la dérivée de la logvraisemblance $L(\boldsymbol{\phi};\mathbf{x}) = \ln f(\mathbf{x}|\,\boldsymbol{\phi})$ des données dites augmentées ($\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$). Cette espérance, notée $E_C(.)$, 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:

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

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 $\partial\left[\ln f(\mathbf{y},\mathbf{z}\,|\,\boldsymbol{\phi})\right]/\partial\boldsymbol{\phi}$ par rapport à la densité de $\mathbf{z}\,|\,\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]}$ qu'on note $E_C^{[t]}\left[\frac{\partial \ln f(\mathbf{y},\mathbf{z}\,|\,\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right]$.

Cette espérance s'écrit:

$$E_C^{[t]}\left[\frac{\partial \ln f(\mathbf{y},\mathbf{z}\,|\,\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right] = \int_{\mathcal{Z}} \frac{\partial \ln f(\mathbf{y},\mathbf{z}\,|\,\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} h\left(\mathbf{z}\,|\,\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]}\right) d\mathbf{z}$$

où $\mathcal{Z}$ désigne l'espace d'échantillonnage de $\mathbf{z}$ et $h\left(\mathbf{z}\,|\,\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]}\right)$ 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 $d\mathbf{z}$.

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

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

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

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

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). $E_C^{[t]}\left[\ln f(\mathbf{y},\mathbf{z}\,|\,\boldsymbol{\phi})\right]$ 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 $Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})$ et son établissement correspond précisément à l'étape E (dite «Expectation») de l'algorithme. L'annulation de sa dérivée première $\frac{\partial}{\partial\boldsymbol{\phi}}\left[Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})\right]=0$ 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 :

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

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

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

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 $\mathbf{e} = \{e_i\} \sim \left(0, \mathbf{I}\sigma_e^2\right)$ 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 $g(\mathbf{y} \mid \mu, \sigma_a^2, \sigma_e^2) = \int_{\mathcal{X}_y} f(\mathbf{a}, \mathbf{e} \mid \mu, \sigma_a^2, \sigma_e^2)\, d\mathbf{a}\, d\mathbf{e}$ avec $\mathcal{X}_y = \{\mathcal{X}; \mathbf{a} + \mathbf{e} = \mathbf{y} - \mathbf{1}\mu\}$. 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

$$\boxed{Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}) = E_C^{[t]}\left[L(\boldsymbol{\phi}; \mathbf{x})\right]}. \tag{13}$$

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

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

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

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

où

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

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

$$\boxed{\boldsymbol{\phi}^{[t+1]} = \arg\max_{\Phi} Q(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]})}. \tag{16}$$

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 $Q(\boldsymbol{\phi}^{[t+1]}; \boldsymbol{\phi}^{[t]}) \geq Q(\boldsymbol{\phi}^{[t]}; \boldsymbol{\phi}^{[t]}), \forall t$.

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 μ_j et de variance σ_j^2 en proportion p_j pour chacune des composantes $j=1,...,J$ du mélange.

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

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

où $\mathbf{p}_{J\,\mathrm{x}\,1}=\left\{p_j\right\}$, $\boldsymbol{\theta}_j=\left(\mu_j,\sigma_j^2\right)'$, $\boldsymbol{\phi}=\left(\mathbf{p}',\boldsymbol{\theta}_1',...,\boldsymbol{\theta}_j',...,\boldsymbol{\theta}_J'\right)'$ représentent les paramètres et $f_j\left(y;\boldsymbol{\theta}_j\right)$ 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 :

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

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 $g\left(x_i;\boldsymbol{\phi}\right)$ du couple $x_i=\left(y_i,z_i\right)'$ peut alors s'écrire

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

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

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

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

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

On en déduit immédiatement l'expression de la logvraisemblance correspondante

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

En prenant l'espérance de $\mathrm{L}\left(\boldsymbol{\phi};\mathbf{x}\right)$ 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 $\mathrm{Q}\left(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}\right)$ à la phase E

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

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

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

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

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

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

Par annulation, on obtient les solutions à savoir

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

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

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

Si l'on avait fait l'hypothèse de variances homogènes, $\left(\sigma_j^2=\sigma^2,\forall j\right)$, la formule (23c) se serait écrite

$$\sigma^{2[t+1]}=\left[\sum_{i=1}^{N}\sum_{j=1}^{J}\alpha_{ij}^{[t]}\left(y_i-\mu_j^{[t+1]}\right)^2\right]/N\,.$$

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 :

$$\mathrm{f}\left(\mathbf{x} \mid \boldsymbol{\phi}\right) = \mathrm{b}(\mathbf{x}) \exp\left[\boldsymbol{\phi}' \mathbf{t}(\mathbf{x})\right] / a\left(\boldsymbol{\phi}\right), \qquad (24)$$

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

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

$$\mathrm{E}\left[\mathbf{t}(\mathbf{x}) \mid \boldsymbol{\phi}\right] = \partial \ln\left[\mathrm{a}\left(\boldsymbol{\phi}\right)\right] / \partial \boldsymbol{\phi}, \qquad (25a)$$

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

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

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

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

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

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

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

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 Φ, 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 $t_B = 2x_{BB} + x_{AB} + x_{BO}$. A la phase M, on résout l'équation (27) $\mathrm{E}\left(t_A \mid p^{[t+1]}, q^{[t+1]}\right) = \mathrm{E}\left(t_A \mid \mathbf{y}, p^{[t]}, q^{[t]}\right)$ soit

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

avec

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

puisque, conditionnellement à y_A, x_{AA} a une distribution binomiale de paramètres y_A et $p/\left(p+2r\right)$. 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 $\mathrm{p}\left(\boldsymbol{\phi} \mid \mathbf{y}\right)$. Il existe pour la logdensité a posteriori l'homologue de la formule (7) pour la logvraisemblance,

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

où $\mathrm{E}_C\left(.\right)$ indique comme précédemment une espérance conditionnelle prise par rapport à $\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}$.

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

$$Q^*(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) = E_C^{[t]}\left[\ln p\left(\boldsymbol{\phi}|\ \mathbf{y},\mathbf{z}\right)\right], \tag{31}$$

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

$$Q^*(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) = Q(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}) + \ln p\left(\boldsymbol{\phi}\right) + Cste\ , \tag{32}$$

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

A 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 $\boldsymbol{\phi}^{[t+1]} = \arg\max_{\Phi} Q^*(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})$.

1.6. Quelques propriétés

1.6.1. Accroissement monotone de la vraisemblance

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

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

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

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

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):

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

Le logarithme étant une fonction concave, on peut majorer cette quantité par application de l'inégalité de Jensen[8].

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

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

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

ce qui établit le théorème de départ (33).

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

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

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 :

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

Ce théorème découle d'un corollaire de (7) et (9). En effet, par définition $Q(\boldsymbol{\phi};\boldsymbol{\phi}_0) = E_C^0\left[L(\boldsymbol{\phi};\mathbf{x}) \right]$ où $E_C^0(.)$ indique une espérance conditionnelle prise par rapport à la

[8] Si X est une variable aléatoire d'espérance μ et si $f(x)$ est une fonction concave, alors $E\left[f(X) \right] \le f(\mu)$

distribution de $\mathbf{z}|\ \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}_0$, et $\dfrac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}} = \dfrac{\partial}{\partial \boldsymbol{\phi}}\left\{ E_C^0\left[L(\boldsymbol{\phi};\mathbf{x})\right]\right\}$. Du fait de l'égalité (9), on peut intervertir les opérateurs de dérivation et d'espérance si bien que $\dfrac{\partial Q(\boldsymbol{\phi};\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}} = E_C^0\left[\dfrac{\partial L(\boldsymbol{\phi};\mathbf{x})}{\partial \boldsymbol{\phi}}\right]$ et en évaluant ces deux fonctions de $\boldsymbol{\phi}$ au point $\boldsymbol{\phi}_0$, il vient compte tenu de (7)

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

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

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

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 $£(L_0) = \left\{\boldsymbol{\phi} \in \Phi; L(\boldsymbol{\phi};\mathbf{y}) = L_0\right\}$ le sous-ensemble de Φ 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 : $\boldsymbol{\phi}^{[0]}, \boldsymbol{\phi}^{[1]}, \boldsymbol{\phi}^{[2]}, ..., \boldsymbol{\phi}^{[t]}, \boldsymbol{\phi}^{[t+1]}, ...$, 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 $\boldsymbol{\phi}^{[t]} \to \boldsymbol{\phi}*$ quand $t \to +\infty$ où $\boldsymbol{\phi}*$ est un point stationnaire (il vérifie $L'(\boldsymbol{\phi}*;\mathbf{y}) = \mathbf{0}$) qui est tel que $L(\boldsymbol{\phi}*;\mathbf{y}) = L* = \operatorname{Lim} L\left(\boldsymbol{\phi}^{[t]};\mathbf{y}\right)$ si l'une ou l'autre des conditions suivantes est remplie:

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

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

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

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

Ce théorème ne garantit donc pas la convergence vers un maximum global de la logvraisemblance $\mathrm{L}(\boldsymbol{\phi}; \mathbf{y})$. Si $\mathrm{L}(\boldsymbol{\phi}; \mathbf{y})$ 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 $\mathrm{L}(\boldsymbol{\phi}; \mathbf{y})$ 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 $\left[\partial \mathrm{Q}(\boldsymbol{\phi}; \boldsymbol{\Psi})\right] / \partial \boldsymbol{\phi}$. 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 $\mathrm{L}(\boldsymbol{\phi}; \mathbf{y})$ est unimodale avec un seul point stationnaire $\boldsymbol{\phi}^*$ à l'intérieur de Φ. On est donc dans le cas a) d'un singleton et, si la fonction $\dfrac{\partial \mathrm{Q}(\boldsymbol{\phi}; \boldsymbol{\Psi})}{\partial \boldsymbol{\phi}}$ 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 $\mathrm{L}(\boldsymbol{\phi}; \mathbf{y})$.

164. Partition de l'information

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

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

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 $\mathrm{E}_C(.)$, d'où

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

Cette formule peut s'écrire symboliquement sous la forme

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

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 $\mathbf{y}$,

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

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

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

Le terme noté

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

s'identifie, eu égard à (44), à la perte d'information $\mathcal{I}_c(\boldsymbol{\phi};\mathbf{x}) - \mathrm{I}(\boldsymbol{\phi};\mathbf{y})$ 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 $\mathbf{S}(\boldsymbol{\phi};\mathbf{y}) = \dfrac{\partial \mathrm{L}(\boldsymbol{\phi};\mathbf{y})}{\partial\boldsymbol{\phi}}$ et $\mathbf{S}(\boldsymbol{\phi};\mathbf{x}) = \dfrac{\partial \mathrm{L}(\boldsymbol{\phi};\mathbf{x})}{\partial\boldsymbol{\phi}}$ les fonctions de score relatives respectivement aux données observées et aux données complètes, on montre (cf. annexe A) que

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

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)

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

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

Comme $\mathrm{E}_C\left[\mathbf{S}(\boldsymbol{\phi};\mathbf{x})\right]=\mathbf{S}(\boldsymbol{\phi};\mathbf{y})$ (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 $\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}$ tel que $\mathbf{S}\left(\hat{\boldsymbol{\phi}};\mathbf{y}\right)=\mathbf{0}$, on a

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

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

$$\mathrm{I}\left(\hat{\boldsymbol{\phi}};\mathbf{y}\right)=\mathcal{I}_c\left(\hat{\boldsymbol{\phi}};\mathbf{x}\right)-\mathcal{I}_m\left(\hat{\boldsymbol{\phi}};\mathbf{y}\right). \tag{50}$$

165. Vitesse de convergence

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

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

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

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

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

Dans cette formule, $\mathbf{J}(\boldsymbol{\phi})$ est la matrice jacobienne $(r\,\mathrm{x}\,r)$ 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 ϕ_j le jème élément du vecteur $\boldsymbol{\phi}$. Si $\boldsymbol{\phi}^{[k-1]}\to\boldsymbol{\phi}^{*}$ alors, sous les conditions de continuité habituelles, $\mathbf{J}\left(\boldsymbol{\phi}^{[k-1]}\right)\to\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)$ si bien qu'à partir d'un certain rang, on pourra écrire $\boldsymbol{\phi}^{[k+1]}-\boldsymbol{\phi}^{[k]}\approx\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)\left(\boldsymbol{\phi}^{[k]}-\boldsymbol{\phi}^{[k-1]}\right)$.

La vitesse de convergence

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

est alors gouvernée par la plus grande valeur propre de $\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)$, $v=\max_{1\le i\le r}\lambda_i$, 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

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

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

$$\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)=\mathcal{I}_{c}^{-1}\left(\boldsymbol{\phi}^{*};\mathbf{x}\right)\mathcal{I}_{m}\left(\boldsymbol{\phi}^{*};\mathbf{y}\right), \tag{55}$$

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 $\mathcal{I}_{m}\left(\boldsymbol{\phi};\mathbf{y}\right)=\mathcal{I}_{c}\left(\boldsymbol{\phi};\mathbf{x}\right)-\mathrm{I}\left(\boldsymbol{\phi};\mathbf{y}\right)$, la formule (55) peut s'écrire aussi

$$\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)=\mathbf{I}_{r}-\mathcal{I}_{c}^{-1}\left(\boldsymbol{\phi}^{*};\mathbf{x}\right)\mathrm{I}\left(\boldsymbol{\phi}^{*};\mathbf{y}\right). \tag{56}$$

Pour être en conformité avec la littérature numérique, c'est la matrice $\mathbf{I}_{r}-\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)=\mathcal{I}_{c}^{-1}\left(\boldsymbol{\phi}^{*};\mathbf{x}\right)\mathrm{I}\left(\boldsymbol{\phi}^{*};\mathbf{y}\right)$ 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

$$\mathrm{I}\left(\boldsymbol{\phi}^{*};\mathbf{y}\right)=\mathcal{I}_{c}\left(\boldsymbol{\phi}^{*};\mathbf{x}\right)\left[\mathbf{I}_{r}-\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)\right], \tag{57}$$

et, pour l'inverse:

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

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

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 $\boldsymbol{\phi}^{[k+1]}=\mathbf{M}\left(\boldsymbol{\phi}^{[k]}\right)$ au premier ordre

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

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

$$\boldsymbol{\phi}^{*} - \boldsymbol{\phi}^{[k-1]} = \left(\boldsymbol{\phi}^{[k]} - \boldsymbol{\phi}^{[k-1]}\right) + \left(\boldsymbol{\phi}^{[k+1]} - \boldsymbol{\phi}^{[k]}\right) + \left(\boldsymbol{\phi}^{[k+2]} - \boldsymbol{\phi}^{[k+1]}\right) + \ldots + \left(\boldsymbol{\phi}^{[k+h-1]} - \boldsymbol{\phi}^{[k+h]}\right) + ..$$

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

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

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

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

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

$$\boxed{\boldsymbol{\phi}^{*} = \boldsymbol{\phi}^{[k-1]} + \left[\mathbf{I}_{r} - \mathbf{J}\left(\boldsymbol{\phi}^{*}\right)\right]^{-1}\left(\boldsymbol{\phi}^{[k]} - \boldsymbol{\phi}^{[k-1]}\right)}. \qquad (60)$$

Laird et al. (1987) ont proposé une approximation numérique de $\mathbf{J}\left(\boldsymbol{\phi}^{*}\right)$ 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

A 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.

171. « 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

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

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 $E\left[\ddot{Q}(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]})\right]$ 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

$$\boldsymbol{\gamma}^{[t+1]} = \arg\max Q(\boldsymbol{\gamma}, \boldsymbol{\theta}^{[t]}; \boldsymbol{\phi}^{[t]}), \tag{62a}$$

$$\boldsymbol{\theta}^{[t+1]} = \arg\max Q(\boldsymbol{\gamma}^{[t+1]}, \boldsymbol{\theta}; \boldsymbol{\phi}^{[t]}). \tag{62b}$$

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,

$$\boldsymbol{\theta}^{[t+1]} = \arg\max L(\boldsymbol{\gamma}^{[t+1]}, \boldsymbol{\theta}; \mathbf{y}). \tag{63}$$

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 $L(\boldsymbol{\phi};\mathbf{x}) = \ln\left[f(\mathbf{x} \mid \boldsymbol{\phi})\right]$ des données complètes à partir, non pas de son expression analytique, mais grâce à une évaluation numérique de celle-ci via le calcul de $\ln\left[f(\mathbf{y}, \mathbf{z}^{[t]} \mid \boldsymbol{\phi})\right]$ où $\mathbf{z}^{[t]}$ est un échantillon simulé de données manquantes tiré dans la distribution conditionnelle de celles-ci de densité $h\left(\mathbf{z}^{[t]} \mid \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}\right)$. 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 $Q\left(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}\right)$ 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 $\mathbf{z}_1,...,\mathbf{z}_j,...,\mathbf{z}_m$ extraits de la loi de densité $h\left(\mathbf{z}\,|\,\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]}\right)$;

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

$$\tilde{Q}\left(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}\right)=\frac{1}{m}\sum_{j=1}^{m}\ln f\left(\mathbf{y},\mathbf{z}_j\,|\,\boldsymbol{\phi}\right). \qquad (64)$$

On remarque de suite que pour $m=1$, MCEM se ramène exactement à SEM, et que pour $m\to\infty$, 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 :

$$\boldsymbol{\phi}^{[t+1]}=\gamma_{t+1}\boldsymbol{\phi}_{SEM}^{[t+1]}+\left(1-\gamma_{t+1}\right)\boldsymbol{\phi}_{EM}^{[t+1]}, \qquad (65)$$

où les γ_t forment une suite de nombres réels décroissant de $\gamma_0=1$ à $\gamma_\infty=0$ avec les deux conditions suivantes : $\mathrm{Lim}\left(\gamma_t/\gamma_{t+1}\right)=1$ et $\sum_t\gamma_t\to\infty$ quand $t\to\infty$. 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 $\underline{Q}\left(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}\right)$ et de la partie simulée en (64) selon la formule

$$\underline{\mathrm{Q}}\left(\boldsymbol{\phi};\boldsymbol{\phi}^{[t+1]}\right)=\gamma_{t+1}\left[\frac{1}{m_{t+1}}\sum\nolimits_{j=1}^{m_{t+1}}\ln f\left(\mathbf{y},\mathbf{z}_{j,t+1}\mid\boldsymbol{\phi}\right)\right]+\left(1-\gamma_{t+1}\right)\underline{\mathrm{Q}}\left(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}\right). \qquad (66)$$

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 $\hat{\boldsymbol{\phi}}$.

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),

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

en fonction de l'inverse $\mathcal{I}_c^{-1}\left(\hat{\boldsymbol{\phi}};\mathbf{x}\right)$ 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 $\mathbf{J}\left(\boldsymbol{\phi}\right)$ 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 $\mathcal{I}_c^{-1}\left(\hat{\boldsymbol{\phi}};\mathbf{x}\right)$. L'apport crucial de Meng et Rubin (1991) est d'avoir montré comment on pouvait évaluer numériquement la matrice $\mathbf{J}\left(\hat{\boldsymbol{\phi}}\right)$ à partir de la mise en œuvre de l'EM classique. Posons, à l'instar de McLachlan et Krishnan (1997) : $\boldsymbol{\phi}_{(j)}^{[t]}=\left(\hat{\phi}_1,\hat{\phi}_2,...,\hat{\phi}_{j-1},\phi_j^{[t]},...,\hat{\phi}_r\right)'$, r_{ij} peut s'écrire comme suit :

$$r_{ij}=\mathrm{Lim}_{t\to\infty}\frac{\mathrm{M}_i\left(\boldsymbol{\phi}_{(j)}^{[t]}\right)-\hat{\phi}_i}{\phi_j^{[t]}-\hat{\phi}_j}. \qquad (67)$$

En fait, l'algorithme EM réalise l'application M (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 $\mathrm{M}_i\left(\boldsymbol{\phi}_{(j)}^{[t]}\right)$ 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 $\mathbf{I}_r - \mathbf{J}\left(\hat{\boldsymbol{\phi}}\right)$ ainsi obtenues sont de bons outils de diagnostic de la solution $\hat{\boldsymbol{\phi}}$ 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 $\hat{\boldsymbol{\phi}}$ est de repartir de la formule générale $\mathrm{I}\left(\hat{\boldsymbol{\phi}};\mathbf{y}\right) = \mathcal{I}_c\left(\hat{\boldsymbol{\phi}};\mathbf{y}\right) - \mathcal{I}_m\left(\hat{\boldsymbol{\phi}};\mathbf{y}\right)$ et d'utiliser la formule de Louis vue en (49ab) soit $\mathcal{I}_m\left(\hat{\boldsymbol{\phi}};\mathbf{y}\right) = \mathrm{E}_C\left[\mathbf{S}\left(\boldsymbol{\phi};\mathbf{x}\right)\mathbf{S}'\left(\boldsymbol{\phi};\mathbf{x}\right)\right]\Big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$ qu'on peut évaluer par simulation en prenant la moyenne sur m échantillons du produit du score $\mathbf{S}\left(\boldsymbol{\phi};\mathbf{y},\mathbf{z}_j\right)$ 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 $\boldsymbol{\alpha}$ tel que $\boldsymbol{\varphi} = \left(\boldsymbol{\phi}_*', \boldsymbol{\alpha}'\right)'$ où $\boldsymbol{\phi}_*$ joue le même rôle dans la densité des données complètes $\mathrm{p}_X\left[\mathbf{x}|\boldsymbol{\varphi} = \left(\boldsymbol{\phi}_*, \boldsymbol{\alpha}\right)\right]$ du modèle étendu (noté X) que $\boldsymbol{\phi}$ dans celle $\mathrm{p}\left(\mathbf{x}|\boldsymbol{\phi}\right)$ 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 $\boldsymbol{\phi} = \mathrm{R}\left(\boldsymbol{\varphi}\right)$;

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 $\mathrm{p}_0\left(\mathbf{x}\,|\,\boldsymbol{\phi}\right)=\mathrm{p}_X\left(\mathbf{x}\,|\,\boldsymbol{\phi}^*=\boldsymbol{\phi},\boldsymbol{\alpha}=\boldsymbol{\alpha}_0\right)$. Autrement dit, si l'on pose $\boldsymbol{\phi}^*=\boldsymbol{\phi}^*\left(\boldsymbol{\alpha}\right)$ alors $\boldsymbol{\phi}^*\left(\boldsymbol{\alpha}_0\right)=\boldsymbol{\phi}$.

La première condition se traduit par le fait que la logvraisemblance reste inchangée $\mathrm{L}(\boldsymbol{\gamma};\mathbf{y})=\mathrm{L}\left(\boldsymbol{\gamma}_*,\boldsymbol{\alpha};\mathbf{y}\right)$ 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 $\mathrm{h}\left(\mathbf{z}\,|\,\mathbf{y},\boldsymbol{\varphi}=\boldsymbol{\varphi}^{[t]}\right)$ 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) $y\,|\,z\sim\mathcal{N}\left(z,1\right)$ où y désigne la variable observée et z la variable manquante ;

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

Remarquons que cela équivaut à écrire : 1) $y=z+e;\ e\sim\mathcal{N}\left(0,1\right)$, et 2) $z=\theta+u;\ u\sim\mathcal{N}\left(0,\sigma^2\right)$ 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 θ, on remplace z par son espérance conditionnelle $\tilde{z}^{[t]}=\mathrm{E}\left(z\,|\,\theta^{[t]},\sigma^2,y\right)$. Du fait de l'hypothèse de normalité des distributions, cette espérance s'écrit comme l'équation de régression de y en z,

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

$$\tilde{z}^{[t]}=\theta^{[t]}+\frac{\sigma^2}{1+\sigma^2}\left(y-\theta^{[t]}\right)=\frac{\theta^{[t]}+\sigma^2 y}{1+\sigma^2}. \qquad (68)$$

Phase M : On résout l'équation $\mathrm{E}\left(z\,|\,\theta^{[t+1]},\sigma^2\right)=\mathrm{E}\left(z\,|\,y,\theta^{[t]},\sigma^2\right)$ 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 σ^2 sera petit.

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

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

On retrouve alors la propriété selon laquelle z est une statistique exhaustive de θ_*. La phase E reste inchangée puisque la loi de $z \mid \theta_*, \alpha = 0, \sigma^2, y$ est identique à la loi de $z \mid \theta, \sigma^2, y$. A la phase M, on résout $\mathrm{E}\left(z \mid \theta_*^{[t+1]}, \sigma^2\right) = \tilde{z}^{[t]}$ soit $\theta_*^{[t+1]} = \frac{\theta^{[t]} + \sigma^2 y}{1+\sigma^2}$. Quant à α, on a, eu égard à l'expression (70), $\alpha^{[t+1]} = y - \tilde{z}^{[t]}$ soit, compte tenu de (68), $\alpha^{[t+1]} = y - \frac{\theta^{[t]} + \sigma^2 y}{1+\sigma^2}$ 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 :

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

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 $\left(\alpha^{[t+1]} - \alpha_0\right)$ 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 $\left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}'\right)' = \arg\max_{\boldsymbol{\beta},\boldsymbol{\gamma}} \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$, où

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

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 $N - \mathrm{r}(\mathbf{X})$ 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),

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

où C est une constante égale dans sa forme la plus simple à $\left[N - \mathrm{r}(\mathbf{X})\right]\ln 2\pi$, $\underline{\mathbf{X}}$ correspond à toute matrice formée par $\mathrm{r}(\mathbf{X})$ colonnes de $\mathbf{X}$ linéairement indépendantes et $\underline{\mathbf{P}} = \mathbf{V}^{-1}\left[\mathbf{I}_N - \mathbf{Q}\right]$ 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 $\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}) = \int \mathrm{p}(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\beta}$ 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 $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{u}_{(q\,\mathrm{x}\,1)} \sim \mathcal{N}(\mathbf{0},\mathbf{G})$, $\mathbf{e}_{(N\,\mathrm{x}\,1)} \sim \mathcal{N}(\mathbf{0},\mathbf{R})$, $\mathrm{Cov}(\mathbf{u},\mathbf{e}') = \mathbf{0}$ 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, $\boldsymbol{\phi} = \left(\sigma_1^2, \sigma_0^2\right)'$ et $\mathbf{x} = \left(\mathbf{y}', \boldsymbol{\beta}', \mathbf{u}'\right)'$ si bien que la densité de $\mathbf{x}$ se factorise en

$\mathrm{p}\left(\mathbf{x} \mid \boldsymbol{\phi}\right) \propto \mathrm{p}\left(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_0^2\right) \mathrm{p}\left(\mathbf{u} \mid \sigma_1^2\right)$. Dans le cas gaussien, on obtient immédiatement :

$$\ln \mathrm{p}\left(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_0^2\right) = \ln \mathrm{p}\left(\mathbf{e} \mid \sigma_0^2\right) = -½\left(N \ln 2\pi + \mathrm{N} \ln \sigma_0^2 + \mathbf{e}'\mathbf{e} / \sigma_0^2\right), \tag{71a}$$

$$\ln \mathrm{p}\left(\mathbf{u} \mid \sigma_1^2\right) = -½\left(q \ln 2\pi + q \ln \sigma_1^2 + \mathbf{u}'\mathbf{u} / \sigma_1^2\right). \tag{71b}$$

En désignant par $\mathrm{L}_0\left(\sigma_0^2; \mathbf{e}\right) = \ln \mathrm{p}\left(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_0^2\right)$, la logvraisemblance de σ_0^2 basée sur $\mathbf{e}$, et par $\mathrm{L}_1\left(\sigma_1^2; \mathbf{u}\right) = \ln \mathrm{p}\left(\mathbf{u} \mid \sigma_1^2\right)$, celle de σ_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 :

$$\mathrm{L}\left(\boldsymbol{\phi}; \mathbf{x}\right) = \mathrm{L}_0\left(\sigma_0^2; \mathbf{e}\right) + \mathrm{L}_1\left(\sigma_1^2; \mathbf{u}\right) + cste. \tag{72}$$

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

$$\mathrm{Q}\left(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}\right) = \mathrm{Q}_0\left(\sigma_0^2; \boldsymbol{\phi}^{[t]}\right) + \mathrm{Q}_1\left(\sigma_1^2; \boldsymbol{\phi}^{[t]}\right), \tag{73}$$

où

$$\mathrm{Q}_0\left(\sigma_0^2; \boldsymbol{\phi}^{[t]}\right) = \mathrm{E}_c^{[t]}\left[\mathrm{L}_0\left(\sigma_0^2; \mathbf{e}\right)\right] = -½\left[N \ln 2\pi + \mathrm{N} \ln \sigma_0^2 + \mathrm{E}_c^{[t]}\left(\mathbf{e}'\mathbf{e}\right) / \sigma_0^2\right], \tag{74a}$$

$$\mathrm{Q}_1\left(\sigma_1^2; \boldsymbol{\phi}^{[t]}\right) = \mathrm{E}_c^{[t]}\left[\mathrm{L}_1\left(\sigma_1^2; \mathbf{u}\right)\right] = -½\left[q \ln 2\pi + q \ln \sigma_1^2 + \mathrm{E}_c^{[t]}\left(\mathbf{u}'\mathbf{u}\right) / \sigma_1^2\right], \tag{74b}$$

$\mathrm{E}_c^{[t]}(.)$ 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 $\mathrm{Q}\left(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}\right)$ par rapport à $\boldsymbol{\phi}$, soit , compte tenu de (73), en la maximisation de $\mathrm{Q}_0\left(\sigma_0^2; \boldsymbol{\phi}^{[t]}\right)$ par rapport à σ_0^2 et en celle de $\mathrm{Q}_1\left(\sigma_1^2; \boldsymbol{\phi}^{[t]}\right)$ par rapport à σ_1^2. Les derivées premières de ces fonctions s'écrivent :

$$\frac{\partial(-2\mathrm{Q}_0)}{\partial\sigma_0^2} = \frac{N}{\sigma_0^2} - \frac{\mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})}{\sigma_0^4}, \tag{75a}$$

$$\frac{\partial(-2\mathrm{Q}_1)}{\partial\sigma_1^2} = \frac{q}{\sigma_1^2} - \frac{\mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u})}{\sigma_1^4}. \tag{75b}$$

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 σ_0^2 et $\mathbf{u}'\mathbf{u}$ de σ_1^2, puis en égalant les espérances de celles-ci à leurs espérances conditionnelles respectives soit $\mathrm{E}\left(\mathbf{e}'\mathbf{e} \mid \sigma_0^{2[t+1]}\right) = \mathrm{E}_c^{[t]}(\mathbf{e}'\mathbf{e})$ et $\mathrm{E}\left(\mathbf{u}'\mathbf{u} \mid \sigma_1^{2[t+1]}\right) = \mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u})$.

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

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

Or,

$$\mathrm{E}\left(\mathbf{u} \mid \mathbf{y}, \boldsymbol{\phi} = \boldsymbol{\phi}^{[t]}\right) = \hat{\mathbf{u}}^{[t]} \tag{78}$$

est le BLUP[9] de $\mathbf{u}$ basé sur $\boldsymbol{\phi}^{[t]} = \left(\sigma_0^{2[t]}, \sigma_1^{2[t]}\right)'$. Par définition, le BLUP a pour expression $\hat{\mathbf{u}} = \mathrm{Cov}(\mathbf{u}, \mathbf{y}')\left[\mathrm{Var}(\mathbf{y})\right]^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ où $\hat{\boldsymbol{\beta}}$ 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)

[9] Abréviation de « Best Linear Unbiased Prediction »

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

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

De même,

$$\operatorname{var}\left(\mathbf{u}\mid\mathbf{y},\boldsymbol{\phi}=\boldsymbol{\phi}^{[t]}\right)=\operatorname{var}\left(\hat{\mathbf{u}}^{[t]}-\mathbf{u}\right)=\mathbf{C}_{uu}^{[t]}\sigma_0^{2[t]}, \tag{80}$$

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 :

$$\sigma_1^{2[t+1]}=\left[\hat{\mathbf{u}}^{[t]'}\hat{\mathbf{u}}^{[t]}+\operatorname{tr}\left(\mathbf{C}_{uu}^{[t]}\right)\sigma_0^{2[t]}\right]/q. \tag{82}$$

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

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

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 :

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

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

où $\hat{\boldsymbol{\theta}}^{[t]}=\left(\hat{\boldsymbol{\beta}}^{[t]'},\hat{\mathbf{u}}^{[t]'}\right)'$ 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 :

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

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

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

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

On note au passage que cette expression diffère de celle de l'algorithme d'Henderson (1973) qui s'écrit simplement $\sigma_0^{2[t+1]}=\left(\mathbf{y}'\mathbf{y}-\hat{\boldsymbol{\theta}}^{[t]'}\mathbf{T}'\mathbf{y}\right)/\left[N-\mathrm{r}(\mathbf{X})\right]$, 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 $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}_{q_k})$; $(k = 1, 2, ..., K)$ 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 :

$$\boxed{\sigma_k^{2[t+1]} = \left[\hat{\mathbf{u}}_k^{[t]} \hat{\mathbf{u}}_k^{[t]'} + \mathrm{tr}\left(\mathbf{C}_{kk}^{[t]} \right) \sigma_0^{2[t]} \right] / q_k}, \tag{86}$$

et

$$\boxed{\begin{aligned} \sigma_0^{2[t+1]} = \{ & \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}^{[t]'} \mathbf{T}'\mathbf{y} - \sum_{k=1}^{K} \lambda_k^{[t]} \hat{\mathbf{u}}_k^{[t]'} \hat{\mathbf{u}}_k^{[t]} \\ & + \left[\mathrm{rang}(\mathbf{X}) + \sum_{k=1}^{K} q_k - \sum_{k=1}^{K} \lambda_k^{[t]} \mathrm{tr}\left(\mathbf{C}_{kk}^{[t]} \right) \right] \sigma_0^{2[t]} \} / N \end{aligned}}. \tag{87}$$

Une des difficultés d'application de cet algorithme réside dans la nécessité de calculer le terme $\mathrm{tr}\left(\mathbf{C}_{uu}^{[t]} \right)$ à chaque itération $[t]$. En fait on peut écrire $\mathbf{C}_{uu}$ sous la forme : $\mathbf{C}_{uu} = \left(\mathbf{Z}'\mathbf{S}\mathbf{Z} + \lambda \mathbf{I}_q \right)^{-1}$ où $\mathbf{S} = \mathbf{I}_N - \mathbf{X}\left(\mathbf{X}'\mathbf{X} \right)^{-} \mathbf{X}'$ 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

$\mathrm{tr}\left(\mathbf{C}_{uu} \right) = \sum_i \delta_i^{-1} \left(\mathbf{Z}'\mathbf{S}\mathbf{Z} + \lambda \mathbf{I}_q \right)$ 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 $\delta_i \left(\mathbf{Z}'\mathbf{S}\mathbf{Z} + \lambda \mathbf{I}_q \right) = \delta_i \left(\mathbf{Z}'\mathbf{S}\mathbf{Z} \right) + \lambda$, d'où $\mathrm{tr}\left(\mathbf{C}_{uu} \right) = \sum_i \left[\delta_i \left(\mathbf{Z}'\mathbf{S}\mathbf{Z} \right) + \lambda \right]^{-1}$. 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 $\boldsymbol{\phi} = \left(\sigma_u^2, \sigma_e^2, \boldsymbol{\beta}' \right)'$ et celui $\mathbf{z}$ des données manquantes par $\mathbf{z} = \mathbf{u}$. On décompose la densité des données complètes $\mathbf{x} = \left(\mathbf{y}', \mathbf{z}' \right)'$ comme précédemment de sorte que

$$\mathrm{L}\left(\boldsymbol{\phi}; \mathbf{x} \right) = \mathrm{L}_0 \left(\sigma_0^2, \boldsymbol{\beta}; \mathbf{e} \right) + \mathrm{L}_1 \left(\sigma_1^2; \mathbf{u} \right) + cste, \tag{88}$$

avec

$$-2\,\mathrm{L}_0\left(\sigma_0^2,\boldsymbol{\beta};\mathbf{e}\right)=N\ln 2\pi+\mathrm{N}\ln\sigma_0^2+\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)/\sigma_0^2,$$

et

$$-2\,\mathrm{L}_1\left(\sigma_1^2;\mathbf{u}\right)=q\ln 2\pi+q\ln\sigma_1^2+\mathbf{u}'\mathbf{u}/\sigma_1^2.$$

L'expression de $\mathrm{Q}_1\left(\sigma_1^2;\boldsymbol{\phi}^{[t]}\right)$ reste formellement inchangée si bien que, comme précédemment, $\sigma_1^{2[t+1]}=\mathrm{E}_c^{[t]}\left(\mathbf{u}'\mathbf{u}\right)/q$. En ce qui concerne $\mathrm{Q}_0\left(\sigma_0^2,\boldsymbol{\beta};\boldsymbol{\phi}^{[t]}\right)$, son expression s'explicite sous la forme suivante :

$$\begin{aligned}-2\,\mathrm{Q}_0\left(\sigma_0^2,\boldsymbol{\beta};\boldsymbol{\phi}^{[t]}\right)=&N\ln 2\pi+\mathrm{N}\ln\sigma_0^2+\\&\left[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\,\mathrm{E}(\mathbf{u}\mid\mathbf{y},\boldsymbol{\phi}^{[t]})\right]'\left[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\,\mathrm{E}(\mathbf{u}\mid\mathbf{y},\boldsymbol{\phi}^{[t]})\right]/\sigma_0^2\\&+\mathrm{tr}\left[\mathbf{Z}\,\mathrm{var}\left(\mathbf{u}\mid\mathbf{y},\boldsymbol{\phi}^{[t]}\right)\mathbf{Z}'\right]/\sigma_0^2\end{aligned}\qquad(89)$$

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

$$\partial\left(-2\,\mathrm{Q}_0\right)/\partial\boldsymbol{\beta}=-2\mathbf{X}'\left[\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\,\mathrm{E}(\mathbf{u}\mid\mathbf{y},\boldsymbol{\phi}^{[t]})\right]/\sigma_0^2.\qquad(90)$$

Par annulation, l'équation obtenue ne dépend pas de σ_0^2 et on peut résoudre en $\boldsymbol{\beta}$:

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

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

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

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

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

où $\hat{\mathbf{u}}^{[t+1]}=\mathrm{E}(\mathbf{u}\mid\mathbf{y},\sigma_0^{2[t]},\sigma_1^{2[t]},\boldsymbol{\beta}^{[t+1]})$ et $\lambda^{[t]}=\sigma_0^{2[t]}/\sigma_1^{2[t]}$.

Notons que cela revient à actualiser la phase E sur la base de $\boldsymbol{\phi}=\left(\sigma_0^{2[t]},\sigma_1^{2[t]},\boldsymbol{\beta}^{[t+1]\prime}\right)'$ 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 :

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

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

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

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

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

Par dérivation de (89) par rapport à σ_0^2, et en annulant, il vient :

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

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

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

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

d'où

$$\boxed{\sigma_0^{2[t+1]} = \{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}^{[t+1]'}\mathbf{T}'\mathbf{y} - \lambda^{[t]}\hat{\mathbf{u}}^{[t+1]'}\hat{\mathbf{u}}^{[t+1]} + \left(q - \lambda^{[t]}\,\mathrm{tr}\left[\left(\mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q\right)^{-1}\right]\sigma_0^{2[t]}\right)\}/N}\,. \tag{97}$$

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 $\mathbf{x} = \left(\mathbf{y}', \boldsymbol{\beta}', \mathbf{u}^{*'}\right)'$, on a $\mathrm{p}\left(\mathbf{x} \mid \boldsymbol{\phi}\right) \propto \mathrm{p}\left(\mathbf{e} \mid \boldsymbol{\phi}\right)$ puisque la densité $\mathrm{p}\left(\boldsymbol{\beta}, \mathbf{u}^*\right)$ est non informative vis-à-vis des paramètres. A la phase E, la fonction $\mathrm{Q}\left(\boldsymbol{\phi}; \boldsymbol{\phi}^{[t]}\right)$ s'écrit donc :

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

A la phase M, il vient par dérivation :

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

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

Rien ne change donc formellement pour l'actualisation de σ_0^2. Par contre en ce qui concerne σ_1, l'annulation de la dérivée conduit à l'expression suivante :

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

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 :

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

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

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

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

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

On peut aussi résoudre les équations du modèle mixte sous leur forme habituelle (cf. 77) puis calculer $\hat{\mathbf{u}}^* = \hat{\mathbf{u}}/\sigma_1$, $\tilde{\mathbf{C}}_{\beta u} = \mathbf{C}_{\beta u}/\sigma_1$ et $\tilde{\mathbf{C}}_{uu} = \mathbf{C}_{uu}/\sigma_1^2$.

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 $R^2 = n/(n+\alpha)$ (α désignant ici le ratio σ_0^2/σ_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).

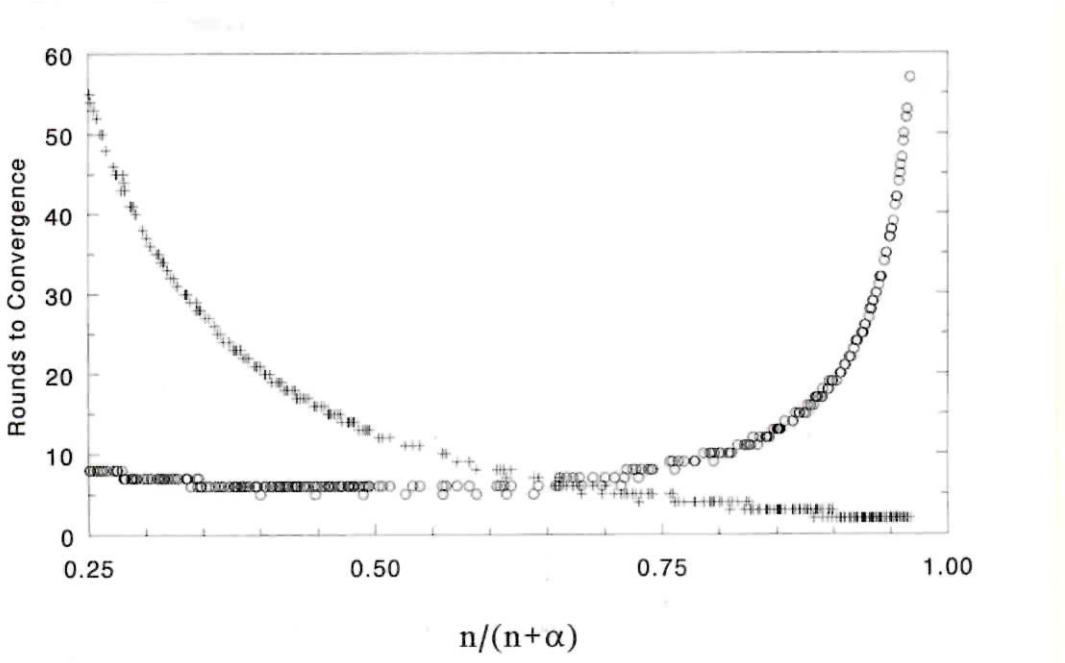


Fig1. Vitesse de convergence (nombre d'itérations) pour les algorithmes EM classique (croix) et « Scaled » (ronds) dans un dispositif de 100 familles de demi-frères de même taille (n) en fonction du rapport $R^2 = n/(n+\alpha)$ où $\alpha = \sigma_0^2 / \sigma_1^2$ est le ratio de la variance résiduelle 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 $\left(\sigma_1^2\right)$ tout comme la variance intra-familles $\left(\sigma_0^2\right)$ 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 $(n_i \text{ x } 1)$ des observations dans la strate $i = 1,2,...,I$; $\boldsymbol{\beta}$ est le vecteur $(p \text{ x } 1)$ des effets fixes associé à la matrice $(n_i \text{ x } p)$ 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 $(n_i \text{ x } q)$ d'incidence correspondante et $\sigma_{1,i}$ est la racine carrée de la composante **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 $\mathbf{A}$), $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2_{0,i}\mathbf{I}_{n_i})$ et $\mathrm{E}(\mathbf{u}^*\mathbf{e}_i') = \mathbf{0}$, $\forall i$.

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 $\left(\sigma^2_{0,i}, \sigma^2_{1,i}\right)$. 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 $\sigma^2_{0,i}$), 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 $(r \text{ x } 1)$ 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 $\boldsymbol{\lambda}$ $(s \text{ x } 1)$ des coefficients des variables explicatives $\mathbf{h}_i$ du logarithme du ratio τ_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 :

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

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)

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

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 :

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

où

$$\boldsymbol{\Delta\delta}^{[t,k+1]} = \boldsymbol{\delta}^{[t,k+1]} - \boldsymbol{\delta}^{[t,k]} \ , \boldsymbol{\Delta\lambda}^{[t,k+1]} = \boldsymbol{\lambda}^{[t,k+1]} - \boldsymbol{\lambda}^{[t,k]},$$

$$\mathbf{P}'_{(r\,\mathrm{x}\,I)} = \left(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_i, ..., \mathbf{p}_I\right) \ , \ \mathbf{H}'_{(s\,\mathrm{x}\,I)} = \left(\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_i, ..., \mathbf{h}_I\right).$$

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

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

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

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 $\left(I\,\mathrm{x}\,I\right)$ dont les éléments s'explicitent en

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

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

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

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 $S_{i,\varepsilon\varepsilon} = \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)'\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)$, $S_{i,\varepsilon u} = \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)'\mathbf{Z}_i\mathbf{u}^*$ et $S_{i,uu} = \mathbf{u}^{*\prime}\mathbf{Z}_i'\mathbf{Z}_i\mathbf{u}^*$,

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

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

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

Ici les équations du modèle mixte s'écrivent $\left(\sum_{i=1}^{I}\sigma_{0,i}^{-2}\mathbf{T}_i'\mathbf{T}_i+\boldsymbol{\Sigma}^{-}\right)\hat{\boldsymbol{\theta}}=\sum_{i=1}^{I}\sigma_{0,i}^{-2}\mathbf{T}_i'\mathbf{y}_i$ avec $\mathbf{T}_i=\left(\mathbf{X}_i,\tau_i\sigma_{0,i}\mathbf{Z}_i\right)$, $\boldsymbol{\theta}=\left(\boldsymbol{\beta}',\mathbf{u}^{*\prime}\right)'$, $\boldsymbol{\Sigma}^{-}=\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}^{-1}\end{pmatrix}$ 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 $\mathrm{E}_{\mathbf{y}}\left[\mathrm{E}\left(S\mid\mathbf{y},\boldsymbol{\phi}\right)\right]=\mathrm{E}\left(S\right)$ dont l'expression est particulièrement simple dans les cas abordés ici, soit

$$\bar{w}_{\delta\delta,ii}=½\left[n_i+\tau_i^2\,\mathrm{tr}\left(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i\right)/2\right],$$

$$\bar{w}_{\delta\lambda,ii}=\tau_i^2\,\mathrm{tr}\left(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i\right)/2,$$

$$\bar{w}_{\lambda\lambda,ii}=\tau_i^2\,\mathrm{tr}\left(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i\right).$$

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, $\mathrm{tr}\left(\mathbf{A}\mathbf{Z}_i'\mathbf{Z}_i\right)=n_i$ si bien que tous ces poids se simplifient en $\bar{w}_{\delta\delta,ii}=½n_i\left(1+\tau_i^2/2\right)$; $\bar{w}_{\delta\lambda,ii}=n_i\tau_i^2/2$ et $\bar{w}_{\lambda\lambda,ii}=n_i\tau_i^2$.

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) :

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

où

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

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

et

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

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 (EMO)

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 $k = 1,2,...,K$:

$$\mathbf{u}_{(q_+\,\mathrm{x}\,1)} = \left(\mathbf{u}_1', \mathbf{u}_2', ..., \mathbf{u}_k', ..., \mathbf{u}_K'\right)' \;;\; \mathbf{Z}_{(N\,\mathrm{x}\,q_+)} = \left(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k, ..., \mathbf{Z}_K\right).$$

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

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 $K = 2$:

$$\mathbf{G} = \mathrm{Var}\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \begin{pmatrix}\sigma_{11}\mathbf{I}_q & \sigma_{12}\mathbf{I}_q\\ \sigma_{12}\mathbf{I}_q & \sigma_{22}\mathbf{I}_q\end{pmatrix} = \mathbf{G}_0 \otimes \mathbf{I}_q \quad \text{où} \quad \mathbf{G}_0 = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix} \quad \text{et, de façon générale,}$$

$\mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$ avec $\mathbf{G}_0 = \left\{\sigma_{kl}\right\}$ pour $k, l = 1,2,...,K$ et $\mathbf{A}_q = \mathbf{I}_q$ si les unités expérimentales ($i = 1,2,...,q$; 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 $\mathbf{y}_i = \{y_{ij}\}; j = 1,...,n_i$ est le vecteur des n_i observations y_{ij} faites sur l'unité expérimentale i.

Ici $\mathbf{u}_{i(Kx1)} = (u_{i1}, u_{i2}, ..., u_{ik}, ..., u_{iK})'$ et $\mathbf{Z}_{i(n_i xK)} = [\mathbf{Z}_{i1}, \mathbf{Z}_{i2}, ..., \mathbf{Z}_{ik}, ..., \mathbf{Z}_{iK}]$ si bien que $\mathbf{u}_{i(K\mathrm{x}1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_0)$ 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 $\mathrm{h}_{i,jj'} = f\left(\rho, \left|t_{ij} - t_{ij'}\right|\right)$.

Si l'on pose $\mathbf{g}_0 = \mathrm{vech}(\mathbf{G}_0)$ [10], $\mathbf{r}$ 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,

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

et

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

Dans (112),

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

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

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

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

avec

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

A 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 :

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

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

[10] vech(**X**) est la notation de vectorisation d'une matrice, homoloque de vec(**X**), mais qui s'applique à une matrice symétrique, seuls les éléments distinctifs étant pris en compte (Searle, 1982).

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 $K = 2$:

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

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

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 $(K \text{ x } K)$ réelle inversible telle qu'au modèle d'origine en (110) (dit modèle O) se substitue le nouveau modèle (dit modèle X),

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

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 $\tilde{\mathbf{u}}_{i(K\text{x}1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_{0*})$ 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 $\boldsymbol{\varphi} = \left[\boldsymbol{\phi}_*', (\text{vec}\,\boldsymbol{\alpha})'\right]'$ avec $\boldsymbol{\phi}_* = \left(\mathbf{g}_{0*}', \mathbf{r}_*'\right)'$ et $\boldsymbol{\varphi}^{[t,0]} = \left[\left(\boldsymbol{\phi}_*^{[t]} = \boldsymbol{\phi}^{[t]}\right)', (\text{vec}\,\boldsymbol{\alpha} = \text{vec}\,\boldsymbol{\alpha}_0)'\right]'$.

L'étape E consiste en l'explicitation de $\text{Q}\left(\boldsymbol{\varphi}; \boldsymbol{\varphi}^{[t,0]}\right)$. 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 $\text{Q}_1\left(\mathbf{g}_{0*}; \boldsymbol{\varphi}^{[t,0]}\right)$ par rapport à $\mathbf{g}_{0*}$ revient à celle de $\mathbf{g}_0$ sous EMO soit

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

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

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

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

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

La résolution de ces K^2 équations ne fait pas intervenir σ_0^2 et se réduit, à une itération donnée, à celle du système linéaire $\mathbf{F}v\mathrm{ec}(\boldsymbol{\alpha})=\mathbf{h}$, soit encore

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

où

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

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

Soit $\mathbf{T}_{kl}=\mathbf{Z}_k^{'}\mathbf{Z}_l$, and $\mathbf{v}_{k_{(q\,\mathrm{x}\,1)}}=\mathbf{Z}_k^{'}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'$ 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)

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

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

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 $\boldsymbol{\beta}$; $\mathbf{C}_{u_k u_m}$ 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 $\mathbf{u}_k$; $\hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{u}}_k$ 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

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

Enfin, quant à σ_0^2, la maximisation de $\mathrm{Q}_0\left(\boldsymbol{\alpha}, \sigma_0^2 \mid \boldsymbol{\varphi}^{[t,0]}\right)$ conduit à:

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

la résiduelle $\mathbf{e}$ étant ajustée en fonction de $\boldsymbol{\alpha}^{[t+1]}$ via $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\alpha}^{[t+1]}\tilde{\mathbf{u}}_i$. 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).

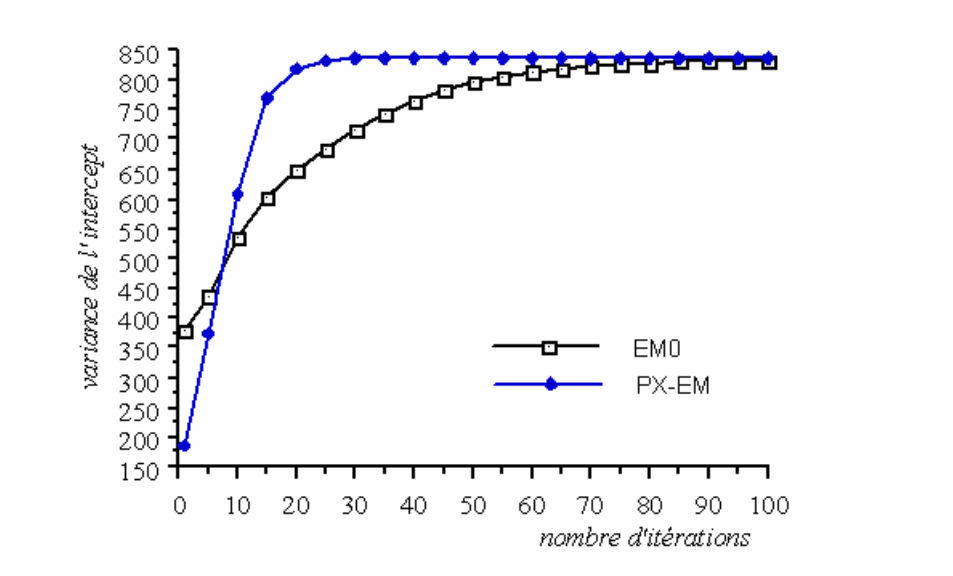


Fig 2 : Séquences typiques d'itérations EMO et PX-EM

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. A cet égard, dans le cadre d'un modèle très proche de (110),

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

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 vec($\boldsymbol{\alpha}$) 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).

REFERENCES

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 7th 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, 5-52.

Grimaud A., Huet S., Monod H., Jenczewski E., Eber F. (2002), Mélange de modèles mixtes : application à l'analyse des appariemments de chromosomes chez des haploïdes de colza, *Journal de la Société Française de Statistique*, 143, 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 7th 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

$$\frac{\partial \ln g(\mathbf{y}\,|\,\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{\partial g(\mathbf{y}\,|\,\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \frac{1}{g(\mathbf{y}\,|\,\boldsymbol{\phi})} \quad . \tag{A.1}$$

Or la densité marginale correspond à

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

d'où sa dérivée

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

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

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

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

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

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

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

soit après simplification,

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

ou encore,

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

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

2. Dérivée seconde

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

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

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

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

En reportant dans (A.7), on obtient

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

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

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

ANNEXE B

Eléments de l'expression de la variance résiduelle en EM

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

Partons des équations du modèle mixte sous leur forme condensée

$$\left(\mathbf{T}'\mathbf{T}+\boldsymbol{\Lambda}\right)\hat{\boldsymbol{\theta}}=\mathbf{T}'\mathbf{y}, \tag{B.2}$$

où $\mathbf{T}=\left(\mathbf{X},\mathbf{Z}\right)$, $\boldsymbol{\theta}=\left(\boldsymbol{\beta}',\mathbf{u}'\right)'$ 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 $\hat{\boldsymbol{\theta}}'$, il vient :

$$\hat{\boldsymbol{\theta}}'\left(\mathbf{T}'\mathbf{T}+\boldsymbol{\Lambda}\right)\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y}$$

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

$$\left(\mathbf{y}-\mathbf{T}\hat{\boldsymbol{\theta}}\right)'\left(\mathbf{y}-\mathbf{T}\hat{\boldsymbol{\theta}}\right)=\mathbf{y}'\mathbf{y}-\hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y}-\hat{\boldsymbol{\theta}}'\boldsymbol{\Lambda}\hat{\boldsymbol{\theta}}$$

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

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

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

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

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

Dans ces conditions,

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

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

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

ANNEXE C

Variances hétérogènes : dérivées intervenant à la phase M

La fonction Q à maximiser présente la forme suivante :

$$\mathrm{Q}\left(\boldsymbol{\phi};\boldsymbol{\phi}^{[t]}\right) = -½\left[N\ln 2\pi + \sum_{i=1}^{I} n_i \ln\sigma_{0,i}^2 + \sum_{i=1}^{I} \mathrm{E}_c\left(\mathbf{e}_i'\mathbf{e}_i\right)/\sigma_{0,i}^2 \right], \tag{C.1}$$

avec

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

$$\ln\tau_i = \mathbf{h}_i'\boldsymbol{\lambda}\,, \tag{C.3}$$

et,

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

1. Dérivée première par rapport à δ

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

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

Or

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

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

soit

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

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

et,

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

D'où

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

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

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

2.. Dérivée première par rapport à λ

En suivant la même démarche que précédemment, on a :

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

avec

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

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

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

et

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

D'où

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

avec

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

3. Dérivée seconde par rapport à δ

Posons

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

où

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

Or

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

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

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

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

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

De la même façon,

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

où

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

$$\frac{\partial v_{\lambda,i}}{\partial \tau_i} = \frac{\mathrm{E}_c\left(\mathbf{u}^{*'}\mathbf{Z}_i'\mathbf{e}_i\right)}{\sigma_{0,i}} + \frac{\tau_i}{\sigma_{0,i}}\mathrm{E}_c\left[\mathbf{u}^{*'}\mathbf{Z}_i'\left(\frac{\partial \mathbf{e}_i}{\partial \tau_i}\right)\right]$$

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

d'où

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

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

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

où

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

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

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

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

c'est-à-dire

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

On vérifie aisément la propriété de symétrie des dérivées, soit

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

6. Espérances des dérivées secondes

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

Par définition

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

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

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

De même,

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

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

$$\tilde{w}_{\delta\delta,ii} = ½ \left[n_i + \tau_i^2 \, \mathrm{tr}\left(\mathbf{Z}_i' \mathbf{Z}_i \mathbf{A} \right) / 2 \right].$$

De même : $\tilde{w}_{\delta\lambda,ii} = ½ \tau_i^2 \, \mathrm{tr}\left(\mathbf{Z}_i' \mathbf{Z}_i \mathbf{A} \right)$ et $\tilde{w}_{\lambda\lambda,ii} = \tau_i^2 \, \mathrm{tr}\left(\mathbf{Z}_i' \mathbf{Z}_i \mathbf{A} \right)$.