JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN-LOUIS FOULLEY
CÉLINE DELMAS
CHRISTÈLE ROBERT-GRANIÉ

Méthodes du maximum de vraisemblance en modèle linéaire mixte

Journal de la société française de statistique, tome 143, n° 1-2 (2002), p. 5-52

<http://www.numdam.org/item?id=JSFS_2002__143_1-2_5_0>

MÉTHODES DU MAXIMUM DE VRAISEMBLANCE EN MODÈLE LINÉAIRE MIXTE

Jean-Louis FOULLEY *,
Céline DELMAS * et Christèle ROBERT-GRANIÉ **

RÉSUMÉ

Cet article présente à des fins pédagogiques une synthèse sur l'utilisation des méthodes du maximum de vraisemblance dans les modèles linéaires mixtes. On y développe tout d'abord l'approche classique de la fonction de vraisemblance des paramètres de position $\boldsymbol{\beta}$ et de dispersion $\boldsymbol{\gamma}$ dans le cadre d'observations gaussiennes $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_\gamma)$. Sur cette base, on rappelle comment on obtient les estimateurs du maximum de vraisemblance (dit ML) et comment on les calcule en pratique. On en déduit les statistiques classiques des tests asymptotiques (Wald, rapport de vraisemblance et des scores) relatifs aux effets fixes.

On aborde ensuite la méthode dite de la vraisemblance restreinte ou résiduelle (REML) à partir de l'exemple simple de N observations gaussiennes indépendantes et identiquement distribuées en s'appuyant sur la factorisation de la vraisemblance en deux composantes dont l'une ne dépend pas du paramètre parasite. On explique comment ce concept dit de vraisemblance marginale peut s'appliquer au cas général et permet de définir une vraisemblance des paramètres de dispersion $\boldsymbol{\gamma}$ basée non plus sur les observations brutes mais sur des projections de celles-ci ou résidus (ou « contrastes d'erreur ») libres des effets fixes $\boldsymbol{\beta}$. On montre aussi que l'estimateur REML peut s'interpréter dans le cadre bayésien comme un maximum de vraisemblance marginale après intégration des effets fixes selon un a priori uniforme sur ceux-ci. On rappelle qu'il peut également être considéré comme un estimateur MINQUE (« Minimum Norm Quadratic Unbiased Estimator ») itéré. On fait état ensuite des procédures de tests des effets fixes qui s'appuient sur les estimateurs REML ou sur le concept de vraisemblance résiduelle.

On aborde enfin la question des tests d'hypothèse sur les effets aléatoires qui s'avère plus délicate du fait de la spécification de valeurs des paramètres en bordure de l'espace.

Mots clés : Modèle linéaire mixte; Composantes de variance; ML; REML; Tests; EM.

ABSTRACT

This paper presents a pedagogical review about maximum likelihood procedures in linear mixed models. We deal at first with the classical approach based on

* Auteur à qui adresser toute correspondance (foulley@jouy.inra.fr), INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy-en-Josas cedex.

** Station d'Amélioration Génétique des Animaux, 31326 Castanet Tolosan cedex.

the likelihood function of both location $\boldsymbol{\beta}$ and dispersion $\boldsymbol{\gamma}$ parameters under the Gaussian model $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_\gamma)$. Then, it is reminded how to get and compute maximum likelihood estimators (ML) and classical asymptotic procedures of tests for fixed effects are deduced (Wald, likelihood ratio and scores statistics).

Next, we address the method of residual (or restricted) likelihood (REML) starting from a simple example of N Gaussian, independently and identically distributed observations by factorizing the likelihood into two components, one of which being free of the nuisance parameter. It is explained how this concept of marginal likelihood can be applied in the general case by defining a likelihood function of the dispersion $\boldsymbol{\gamma}$ parameters based no longer on the raw observations but on some projections of them or residuals (the so called "error contrasts") which are free of the location $\boldsymbol{\beta}$ parameters. It is also shown how REML can be interpreted within the Bayesian framework as a marginalized likelihood after integrating out fixed effects via a uniform prior distribution on them. It is recalled that REML can be viewed as well as an iterative MINQUE (Minimum Norm Quadratic Unbiased Estimator). Then, testing procedures of fixed effects using REML or the concept of residual likelihood are reviewed.

Eventually, testing random effects is discussed emphasizing the non standard approach to it due to the specification of parameter values on the border of the parameter space.

Keywords : Linear mixed models; Variance components; ML; REML; Testing procedures; EM.

Plan de l'article

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). Cet article marque la rupture avec les estimateurs quadratiques inspirés 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. Rappels sur le modèle mixte

11. Définition

L'histoire du modèle mixte remonte, comme souvent en statistiques à Ronald Fisher (1925), et plus précisément, à ses travaux sur l'analyse de variance et sur le coefficient de corrélation intra-classe. Cependant, il fallut attendre le papier d'Eisenhart (1947) dans Biometrics pour qu'une distinction claire apparût entre effets fixes (Classe I) et effets aléatoires (Classe II).

Il y a plusieurs façons d'introduire le modèle mixte; nous retiendrons les deux approches suivantes :

111. Approche de Rao et Kleffe (1988)

Un modèle linéaire mixte est un modèle linéaire de type $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\varepsilon, \varepsilon \sim (\mathbf{0}, \mathbf{V})$ dans lequel la variable aléatoire ε relative à l'écart entre les observations et le modèle explicatif de l'espérance $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, est décomposée en une combinaison linéaire de variables aléatoires structurales $\mathbf{u}_k$; $k = 0, 1, 2, \ldots, K$ non observables :

$$\varepsilon = \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k = \mathbf{Z}\mathbf{u}, \tag{1}$$

où $\mathbf{Z}_{N\times q_+} = (\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k, \ldots, \mathbf{Z}_K)$ est une concaténation de matrices connues d'incidence $\mathbf{Z}_k$ de dimension $(N \times q_k)$ et $\mathbf{u}_{(q_+\times 1)} = (\mathbf{u}_0', \mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_k', \ldots, \mathbf{u}_K')'$ est le vecteur correspondant des variables structurales $\mathbf{u}_k = \{u_{kl}\}$; $l = 1, 2, \ldots, q_k$ tel que $\mathbf{u} \sim (\mathbf{0}, \Sigma_u)$. Si l'on suppose en outre que Σ_u est

une fonction linéaire de paramètres $\Sigma_u = \sum_{m=1}^{M} \theta_m \mathbf{F}_m$, on obtient une structure linéaire pour la matrice de variance covariance $\mathbf{V} = \mathbf{Z}\Sigma_u\mathbf{Z}'$; cette propriété est caractéristique de ce que l'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 des paramètres.

112. Approche marginale de modèles hiérarchiques

À l'instar de Lindley and Smith (1972), on considère un processus d'échantillonnage gaussien en deux étapes relatives respectivement aux données et aux paramètres de position :

$$\text{a) } \mathbf{y} \mid \boldsymbol{\theta}_1\mathbf{C}_1 \sim N(\mathbf{A}_1\boldsymbol{\theta}_1, \mathbf{C}_1), \text{ b) } \boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \mathbf{C}_2 \sim N(\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_2). \tag{2}$$

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

$$\text{c) } \mathbf{y} \mid \boldsymbol{\theta}_2, \mathbf{C}_1, \mathbf{C}_2 \sim N(\mathbf{A}_1\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_1 + \mathbf{A}_1\mathbf{C}_2\mathbf{A}_1') \tag{3}$$

qui génère ainsi la structure mixte classique $\mathbf{y} = \mathbf{A}_1\mathbf{A}_2\theta_2 + \mathbf{A}_1\mathbf{u} + \mathbf{e}$ où $\mathbf{e} \sim N(\mathbf{0}, \mathbf{C}_1)$ et $\mathbf{u} \sim N(\mathbf{0}, \mathbf{C}_2)$.

Une illustration immédiate de cette approche réside dans la modélisation de profils longitudinaux par des modèles à coefficients aléatoires (Diggle *et al.*, 1994; Laird et Ware, 1982). S'agissant par exemple de données de croissance faciale mesurées chez 11 filles et 16 garçons à 4 âges équidistants (8, 10, 12 et 14 ans), Verkebe et Molenberghs (2000) proposent le modèle de régression linéaire suivant : $y_{ijk} = A_{ik} + B_{ik}t_j + e_{ijk}$, où i désigne l'indice du sexe, j celui de la mesure, t_j l'âge correspondant et k l'individu intra sexe. Dans ce modèle, l'ordonnée à l'origine A_{ik} et la pente B_{ik} sont décomposées à leur tour en $A_{ik} = \alpha_i + a_{ik}$ et $B_{ik} = \beta_i + b_{ik}$ c'est-à-dire en une espérance propre à chaque sexe (respectivement α_i, β_i) et en un écart individuel (respectivement a_{ik}, b_{ik}) considéré comme aléatoire eu égard à l'échantillonnage des individus.

Avec cette formulation, un modèle linéaire mixte se définit alors comme un modèle linéaire dans lequel toute ou partie des paramètres associés à certaines unités expérimentales sont traités comme des variables aléatoires eu égard à l'échantillonnage de ces unités dans une population plus large.

12. Notations

Les modèles mixtes étant définis, il reste à préciser les notations génériques qui les désignent. Pour des raisons de commodité, nous nous en tiendrons à l'usage et utiliserons les notations consacrées par Henderson reprises actuellement dans nombre d'ouvrages et de logiciels (SAS par exemple).

Sous la forme la plus générale, le modèle linéaire mixte s'écrit :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{4}$$

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 $\boldsymbol{\beta} \in R^p$ 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 :

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{5a}$$

et

$$\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}, \tag{5b}$$

où $\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$ et $\mathrm{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$.

Cette écriture générale peut rendre compte de la plupart des situations particulières rencontrées en pratique, notamment celle d'un modèle de type « analyse de variance » à plusieurs facteurs aléatoires $k = 1, 2, \ldots, K$ tel que :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{u}_k + \mathbf{e}, \tag{6a}$$

En situation de non corrélation entre les $\mathbf{u}_k$ et entre ceux-ci et la résiduelle, et en supposant $\mathbf{u}_k \sim (\mathbf{0}, \sigma_k^2 \mathrm{I}_{qk})$ et $\mathbf{e} \sim (\mathbf{0}, \sigma_0^2 \mathrm{I}_N)$, la matrice de variance covariance $\mathbf{V}$ des observations s'écrit :

$$\mathbf{V} = \sum_{k=1}^{K} \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k' + \sigma_0^2 \mathbf{I}_N. \tag{6b}$$

Les paramètres σ_0^2 et $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ sont appelées les composantes de variance et la formule (6b) illustre bien la linéarité de $\mathbf{V}$ par rapport à celles-ci.

Comme dans la présentation de Rao et Kleffe, on peut – si besoin est – amalgamer la résiduelle $\mathbf{e}$ à un vecteur $\mathbf{u}_0$ d'effets aléatoires soit : $\mathbf{u}_0 = e$, $\mathbf{Z}_0 = \mathbf{I}_N$ d'où

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k, \tag{7a}$$

et

$$\mathbf{V} = \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{Z}_k' \sigma_k^2. \tag{7b}$$

2. Méthode dite ML

21. Fonction de vraisemblance

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

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_\gamma) \tag{8}$$

où $\mathbf{X}$ est la matrice $(N \times p)$ des p variables explicatives relatives aux N éléments du vecteur $\mathbf{y}$ des observations et $\boldsymbol{\beta}$, le vecteur $(p \times 1)$ des coefficients

de ces variables ou effets fixes. $\mathbf{V}_\gamma$ est la matrice $(N \times N)$ 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 y s'écrit :

$$p_Y(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = (2\pi)^{-N/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right], \qquad (9)$$

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

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad (10a)$$

ou, sous sa forme « $-2L$ »

$$\boxed{-2L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = N\ln(2\pi) + \ln|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}. \qquad (10b)$$

22. Maximisation

221. Dérivées premières

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

$$\widehat{\boldsymbol{\alpha}} = \arg\max_{\alpha \in A = \mathbf{R}^p \times \Gamma} L(\boldsymbol{\alpha}; \mathbf{y}) \qquad (11)$$

se fait habituellement en annulant les dérivées premières :

$$\frac{\partial L(\boldsymbol{\alpha}; \mathbf{y})}{\partial \boldsymbol{\alpha}} = \mathbf{0} \qquad (12)$$

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 $\dfrac{\partial^2 L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} < \mathbf{0}$ forment une matrice définie-négative. Si la condition $\boldsymbol{\beta} \in \mathbf{R}^p$ 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 $\Gamma = \{\sigma_0^2 > 0;\ \sigma_k^2 \geqslant 0,\ \forall k = 1, \ldots, K\}$.

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 :

$$\frac{\partial(-2\mathrm{L})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{13}$$

$$\frac{\partial(-2\,\mathrm{L})}{\partial \gamma_k} = \frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} + (\mathbf{y} - \mathbf{X}\beta)'\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{14}$$

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

$$\frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) \tag{15}$$

$$\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k} = -\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}. \tag{16}$$

d'où

$$\frac{\partial(-2\mathrm{L})}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{17}$$

L'annulation des dérivées premières en (13) et (14) conduit au système suivant :

$$\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{18a}$$

$$\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right)_{\mathbf{V}=\hat{\mathbf{V}}} - (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\Big|_{\mathbf{V}=\hat{\mathbf{V}}}\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0. \tag{18b}$$

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 (18b) en reportant son expression $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ de (18a) dans (18b) et en remarquant que : $\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{V}}^{-1}(\mathbf{I} - \hat{\mathbf{Q}})\mathbf{y} = \underline{\hat{\mathbf{P}}}\mathbf{y}$ où $\underline{\hat{\mathbf{P}}}$ représente la notation abrégée de la valeur de la matrice

$$\underline{\mathbf{P}} = \mathbf{V}^{-1}(\mathbf{I} - \mathbf{Q}) = \mathbf{V}^{-1} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}, \tag{19}$$

(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.

222. Cas général

Le système en (18ab) 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 $\mathbf{J}(\boldsymbol{\alpha}) = \mathrm{E}_{Y|\alpha}[-\ddot{\mathbf{L}}(\boldsymbol{\alpha};\mathbf{y})]$ (*cf.* annexe II), soit , pour cette dernière

$$\boxed{\mathbf{J}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}/2 \end{bmatrix}}. \tag{20}$$

où

$$(\mathbf{F})_{kl} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \tag{21}$$

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

$$\mathbf{J}(\boldsymbol{\gamma}^{[n]})\Delta^{[n+1]} = \dot{\mathbf{L}}(\boldsymbol{\gamma}^{[n]}) \tag{22}$$

où

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

$$\dot{\mathbf{L}}(\boldsymbol{\gamma}^{[n]}) = \left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) + \frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\mathbf{y}\right\}\Big|_{\gamma=\gamma^{[n]}}, \tag{23}$$

$$\mathbf{J}(\boldsymbol{\gamma}^{[n]}) = 1/2\mathbf{F}(\boldsymbol{\gamma}^{[n]}). \tag{24}$$

L'estimation $\hat{\boldsymbol{\gamma}}$ étant obtenue, on en déduit $\hat{\boldsymbol{\beta}}$ par résolution de (18a) 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, $\hat{\mathbf{V}}$.

223. Cas du modèle linéaire 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 $(N \times N)$ connue, par exemple : $\mathbf{V}_k = \mathbf{Z}_k\mathbf{Z}_k'$ dans le cas du modèle linéaire mixte considéré en (6ab), et l'équation (18b) devient

$$\mathrm{tr}(\hat{\mathbf{V}}^{-1}\mathbf{V}_k) - \mathbf{y}'\hat{\mathbf{P}}\mathbf{V}_k\hat{\mathbf{P}}\mathbf{y} = 0. \tag{25}$$

Du fait de la linéarité de $\mathbf{V}$, on peut expliciter le terme de gauche de (25) en

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

Le système en (25) s'écrit alors

$$\boxed{\sum_{l=0}^{K} \mathrm{tr}(\hat{\mathbf{V}}^{-1}\mathbf{V}_k\hat{\mathbf{V}}^{-1}\mathbf{V}_l)\hat{\gamma}_l = \mathbf{y}'\underline{\hat{\mathbf{P}}}\mathbf{V}_k\underline{\hat{\mathbf{P}}}\mathbf{y}}, (k = 0, 1, \ldots, K) \tag{26a}$$

soit encore, sous forme matricielle :

$$\boxed{\hat{\mathbf{F}}\hat{\boldsymbol{\gamma}} = \hat{\mathbf{g}}}, \tag{26b}$$

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

$$\mathbf{F} = \{f_{kl}\} = \{\mathrm{tr}(\mathbf{V}^{-1}\mathbf{V}_k\mathbf{V}^{-1}\mathbf{V}_l)\}, \tag{27a}$$
$$\mathbf{g} = \{g_k\} = \{\mathbf{y}'\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{y}\}, \tag{27b}$$

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

Le système en (26ab) 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 $\boldsymbol{\gamma}$:

$$\boxed{\mathbf{F}(\boldsymbol{\gamma}^{[n]})\boldsymbol{\gamma}^{[n+1]} = \mathbf{g}(\boldsymbol{\gamma}^{[n]})}, \tag{28}$$

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 système ainsi obtenu en $\boldsymbol{\gamma}$ de façon à obtenir la valeur du paramètre à l'itération suivante.

On montre aisément que le système (28) équivaut à celui des équations des scores de Fisher (22) au coefficient 1/2 près.

Lorsque $\mathbf{V}_k = \mathbf{Z}_k\mathbf{Z}_k'$, le calcul des éléments de $\mathbf{F}$ et de $\mathbf{g}$ en (27ab) et (28) 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, *i.e.* $\mathrm{tr}(\mathbf{AA}') = \sum_{ij} a_{ij}^2$. Ainsi, $f_{kl} = \sum_{ij}(\mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{Z}_l)_{ij}^2$ et $g_k = \sum_i(\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{y})_i^2$.

23. Variantes

231. Vraisemblance profilée

L'idée à la base de la vraisemblance profilée est de maximiser la vraisemblance par étapes successives. On va d'abord maximiser $\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$ par rapport à $\boldsymbol{\beta}$, puis la fonction ainsi obtenue $\mathrm{L}_P(\boldsymbol{\gamma}; \mathbf{y}) = \mathrm{L}(\hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}; \mathbf{y})$ (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

$$\begin{aligned} \mathrm{Max}_{\beta,\gamma}\, \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) &= \mathrm{Max}_\gamma[\mathrm{Max}_\beta\, \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})] \\ &= \mathrm{Max}_\gamma\, \mathrm{L}(\hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}; \mathbf{y}), \\ &= \mathrm{Max}_\gamma\, \mathrm{L}_P(\boldsymbol{\gamma}; \mathbf{y}) \end{aligned} \tag{29}$$

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

Compte tenu de (10b), il vient immédiatement :

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

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

$$\boxed{-2\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y}) = N\ln(2\pi) + \ln|\mathbf{V}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y}}. \tag{30}$$

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 :

$$\frac{\partial[-2\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y})]}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - \mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\mathbf{y} \tag{31}$$

qui coïncide bien (au coefficient 1/2 près) avec (18b).

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 – et en dépit de son appellation – une fonction de logvraisemblance même si, à l'occasion, elle conserve certaines de ses propriétés (Berger *et al.*, 1999).

232. Formulation de Hartley-Rao

Hartley et Rao (1967) se placent dans le cadre du modèle linéaire mixte gaussien usuel décrit au §12.

Au lieu de paramétrer $\mathbf{V}$ en terme de variances $\boldsymbol{\sigma}^2 = \{\sigma_k^2\}$, Hartley et Rao isolent la variance résiduelle σ_0^2 et introduisent le vecteur $\boldsymbol{\eta}_{K\times 1} = \{\eta_k = \sigma_k^2/\sigma_0^2\}$ 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 $|\mathbf{V}| = |\mathbf{H}|\sigma_0^{2N}$, la logvraisemblance s'écrit :

$$\begin{aligned}-2\mathrm{L}(\boldsymbol{\beta},\sigma_0^2,\boldsymbol{\eta};\mathbf{y}) = {} & N\ln(2\pi) + \ln|\mathbf{H}| + N\ln\sigma_0^2 \\ & + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma_0^2.\end{aligned} \tag{32}$$

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

$$\frac{\partial(-2\mathrm{L})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma_0^2, \tag{33a}$$

$$\frac{\partial(-2\mathrm{L})}{\partial \sigma_0^2} = \frac{N}{\sigma_0^2} - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma_0^4}, \tag{33b}$$

$$\frac{\partial(-2\mathrm{L})}{\partial \eta_k} = \mathrm{tr}\left(\mathbf{H}^{-1}\frac{\partial \mathbf{H}}{\partial \eta_k}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}\frac{\partial \mathbf{H}}{\partial \eta_k}\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma_0^2. \tag{33c}$$

Par annulation de ces dérivées, on obtient immédiatement :

$$\mathbf{X}'\hat{\mathbf{H}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\hat{\mathbf{H}}^{-1}\mathbf{y}, \tag{34a}$$

$$\hat{\sigma}_0^2 = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}')\hat{\mathbf{H}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/N, \tag{34b}$$

$$\mathrm{tr}(\hat{\mathbf{H}}^{-1}\mathbf{H}_k) - (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{H}}^{-1}\mathbf{H}_k\hat{\mathbf{H}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}_0^2 = 0, \tag{34c}$$

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

On retrouve en (34a) 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 : $\mathrm{tr}(\mathbf{H}^{-1}\mathbf{H}_k) = \sum_{l=1}^{K} \mathrm{tr}(\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}\mathbf{H}_l)\eta_l + \mathrm{tr}(\mathbf{H}^{-2}\mathbf{H}_k)$, d'où le système linéaire itératif suivant :

$$\begin{aligned}&\sum_{l=1}^{K} \mathrm{tr}(\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}\mathbf{H}_l)\big|_{\eta=\eta^{[n]}}\eta_l^{[n+1]} = \\ &\left[(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}_0^2 - \mathrm{tr}(\mathbf{H}^{-2}\mathbf{H}_k)\right]\big|_{\eta=\eta^{[n]}}\end{aligned} \tag{35}$$

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

La même remarque qu'en (28) s'applique ici quant à la simplification des calculs des éléments des traces intervenant en (35).

24. Aspects calculatoires

241. Algorithme d'Henderson

Henderson (1973) se place également dans le cadre du modèle linéaire mixte défini en (5ab; 6ab) et considère la dérivée de $-2\mathrm{L}_P$ par rapport à σ_k^2 (*cf.* (25)) qui s'écrit :

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

Or, le meilleur prédicteur linéaire sans biais (acronyme BLUP en anglais) $\hat{\mathbf{u}}_k$ de $\mathbf{u}_k$ s'écrit par définition : $\hat{\mathbf{u}}_k = \mathrm{Cov}(\mathbf{u}_k, \mathbf{y}')\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$, 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 : $\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}_k\mathbf{Z}_k') = \dfrac{q_k}{\sigma_k^2} - \dfrac{\mathrm{tr}(\underline{\mathbf{C}}_{kk})\sigma_0^2}{\sigma_k^4}$ où, en posant $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k, \ldots, \mathbf{Z}_K)$, $\underline{\mathbf{C}}_{kk} = \left[(\mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1})^{-1}\right]_{kk}$ est le bloc relatif au facteur k de taille $(q_k \times q_k)$ 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 :

$$q_k\hat{\sigma}_k^2 = \hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k + \mathrm{tr}(\hat{\underline{\mathbf{C}}}_{kk})\hat{\sigma}_0^2 \tag{36}$$

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 :

$$\hat{\sigma}_0^2(\boldsymbol{\eta}) = \left[\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\boldsymbol{\eta})\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'(\boldsymbol{\eta})\mathbf{Z}'\mathbf{y}\right] / N \tag{37}$$

Henderson propose alors d'utiliser les expressions (36) et (37) comme bases d'un algorithme itératif de calcul des estimateurs ML de σ_k^2, soit :

$$\boxed{\sigma_k^{2[t+1]} = \left\{\hat{\mathbf{u}}_k'(\boldsymbol{\eta}^{[t]})\hat{\mathbf{u}}_k(\boldsymbol{\eta}^{[t]}) + \mathrm{tr}\left[\underline{\mathbf{C}}_{kk}(\boldsymbol{\eta}^{[t]})\right]\sigma_0^{2[t]}\right\} / q_k} \tag{38a}$$

$$\boxed{\sigma_0^{2[t+1]} = \left[\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\boldsymbol{\eta}^{[t]})\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'(\boldsymbol{\eta}^{[t]})\mathbf{Z}'\mathbf{y}\right] / N} \tag{38b}$$

où $\boldsymbol{\eta}^{[t]} = \{\sigma_k^{2[t]}/\sigma_0^{2[t]}\}$ est le vecteur $(K \times 1)$ 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 réécrire (36) sous la forme suivante : $q_k\hat{\sigma}_k^2 = \hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k + \mathrm{tr}(\underline{\hat{\mathbf{C}}}_{kk})\hat{\sigma}_k^2/\hat{\eta}_k$ et de factoriser $\hat{\sigma}_k^2$ à gauche d'où la formule :

$$\boxed{\sigma_k^{2[t+1]} = \left[\hat{\mathbf{u}}_k'(\boldsymbol{\eta}^{[t]})\hat{\mathbf{u}}_k(\boldsymbol{\eta}^{[t]})\right] / \left\{q_k - \mathrm{tr}\left[\underline{\mathbf{C}}_{kk}(\boldsymbol{\eta}^{[t]})\right]/\eta_k^{[t]}\right\}} \tag{38c}$$

qui est combinée pour la variance résiduelle avec (38b). 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 (30) de la logvraisemblance profilée (multipliée par moins deux)

$$-2\mathrm{L}_P = N\ln 2\pi + \ln|\mathbf{V}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \tag{39}$$

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 $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, $\underline{\mathbf{P}}\mathbf{y} = \mathbf{R}^{-1}\hat{\mathbf{e}}$, d'où il découle que :

$$\mathbf{y}'\underline{\mathbf{P}}\mathbf{y} = \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}, \tag{40}$$

où $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}')'$ 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 :

$$|\mathbf{V}| = |\mathbf{R}||\mathbf{G}||\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}|. \tag{41}$$

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

$$\boxed{-2\mathrm{L}_P = N\ln 2\pi + \ln|\mathbf{R}| + \ln|\mathbf{G}| + \ln|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}| + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}}. \tag{42}$$

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 :

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

et,

$$-2\mathrm{L}_P = N(\ln 2\pi + 1) + \ln|\mathbf{R}| + \ln|\mathbf{G}| + \ln|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}|. \tag{43}$$

En (43), $\mathrm{L}_P = \mathrm{L}[\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}), \hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta}]$, $\mathbf{R} = \mathbf{R}[\hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta}]$, de même pour $\mathbf{G} = \mathbf{G}[\hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta}]$, $\boldsymbol{\eta}$ 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 $Z'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} = (\mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1})/\sigma_0^2$, d'où

$$-2\mathrm{L}_P = N(\ln 2\pi + 1) + (N - q)\ln\hat{\sigma}_0^2 + \ln|\mathbf{G}| + \ln|\mathbf{Z}'\mathbf{Z} + \hat{\sigma}_0^2\mathbf{G}^{-1}|, \tag{44}$$

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).

$$-2\mathrm{L}_P = N(\ln 2\pi + 1) + \left(N - \sum_{k=1}^{K} q_k\right)\ln\hat{\sigma}_0^2 + \sum_{k=1}^{K} q_k \ln\sigma_k^2 + \sum_{k=1}^{K} \ln|\mathbf{A}_k| + \ln\left|\mathbf{Z}'\mathbf{Z} + \bigoplus_{k=1}^{K} \mathbf{A}_k^{-1}(\hat{\sigma}_0^2/\sigma_k^2)\right|. \tag{45}$$

Cet inventaire n'a aucune prétention à l'exhaustivité. Il faudrait également envisager les modèles multidimensionnels. Dans tous les cas, la formule générale (42) peut être appliquée.

25. Tests d'hypothèses

251. 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 $\sqrt{N}(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha})$ converge en loi vers une distribution normale centrée, de matrice de variance-covariance $\operatorname{Lim} N[\mathbf{J}_N(\boldsymbol{\alpha})]^{-1}$ quand $N \to \infty$ (Sweeting, 1980; Mardia et Marshall, 1984) soit, en bref :

$$\sqrt{N}(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}) \underset{\mathrm{L}}{\to} \mathrm{N}\big(0, \operatorname{Lim} N[\mathbf{J}_N(\boldsymbol{\alpha})]^{-1}\big), \tag{46}$$

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

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

$$\hat{\mathbf{J}}_N^{T/2}(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}) \underset{\mathrm{L}}{\to} \mathrm{L}(\mathbf{0}, \mathbf{I}), \tag{47}$$

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 (46) se généralise à une fonction $\mathbf{g}(\boldsymbol{\alpha})$ continûment dérivable (de $\mathbf{R}^p$ dans $\mathbf{R}^q$)

$$\sqrt{N}\big[\mathbf{g}(\hat{\boldsymbol{\alpha}}_N) - \mathbf{g}(\boldsymbol{\alpha})\big] \underset{\mathrm{L}}{\to} \mathrm{N}\left(0, \operatorname{Lim} N \frac{\partial \mathbf{g}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'}[\mathbf{J}_N(\boldsymbol{\alpha})]^{-1}\frac{\partial \mathbf{g}'(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right). \tag{48}$$

252. Statistique de Wald

On va considérer le test de 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 $(r \times p)$ 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.* (20)) 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 (47) et (48) à $\mathbf{k}'\hat{\boldsymbol{\beta}}$, soit, sous l'hypothèse nulle,

$$\sqrt{N}(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}) \underset{\mathrm{L}}{\to} \mathrm{N}(0, \operatorname{Lim} N\, \mathbf{k}'\mathbf{J}_\beta^{-1}\mathbf{k}), \tag{49}$$

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

$$\left[(\mathbf{k}'\hat{\mathbf{J}}_\beta^{-1}\mathbf{k})^{-1}\right]^{T/2}(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}) \underset{\mathrm{L}}{\rightarrow} \mathrm{N}(\mathbf{0}, \mathbf{I}_r), \tag{50}$$

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

$$\boxed{(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m})'\left[\mathbf{k}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{k}\right]^{-1}(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}) \underset{\mathrm{L}}{\rightarrow} \chi_r^2}, \tag{51}$$

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 :

$$\mathbf{k}'\hat{\boldsymbol{\beta}} \approx \mathrm{N}\left[\mathbf{k}'\boldsymbol{\beta}, \mathbf{k}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{k}\right] \tag{52}$$

À 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 $(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m})'\left[\mathbf{k}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{k}\right]^{-1}(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m})$, on sait que, sous H_0, $\left[W(\hat{\sigma}_0^2)\right]/r \sim \mathrm{F}[r, N - r(\mathbf{X})]$. 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.

253. 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) :

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

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

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

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

$$\boxed{\lambda = -2\mathbf{L}_R + 2\mathbf{L}_C \mid_{H_0} \underset{L}{\rightarrow} \chi^2_{\dim(B)-\dim(B_0)}} \tag{53}$$

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 :

$$U = \tilde{\mathbf{S}}'_\beta \tilde{\mathbf{J}}_\beta \tilde{\mathbf{S}}_\beta, \tag{54}$$

où $\tilde{\mathbf{S}}_\beta$ est la valeur de la fonction score $\mathbf{S}_\beta = \mathbf{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = \partial L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})/\partial \boldsymbol{\beta}$ 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 $\mathbf{J}_\beta = \partial^2 L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'$ 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

$$\boxed{U(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})|_{H_0} \underset{L}{\rightarrow} \chi^2_{\dim(B)-\dim(B_0)}}. \tag{55}$$

Un cas particulièrement intéressant est celui du test d'absence d'effets 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, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V})$. 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

$$U = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_1)'\tilde{V}^{-1}\mathbf{X}_2(\mathbf{X}'_2\tilde{\mathbf{V}}^{-1}\mathbf{X}_2)^-\mathbf{X}'_2\tilde{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1). \tag{56}$$

Si l'on pose $\underline{\mathbf{P}}_1 = \mathbf{V}^{-1}(\mathbf{I}_N - \mathbf{Q}_1)$ avec $\mathbf{Q}_1 = \mathbf{X}_1(\mathbf{X}'_1\mathbf{V}^{-1}\mathbf{X}_1)^-\mathbf{X}'_1\mathbf{V}^{-1}$, (56) peut s'écrire aussi

$$U = \mathbf{y}'\tilde{\underline{\mathbf{P}}}_1\mathbf{X}_2(\mathbf{X}'_2\tilde{\underline{\mathbf{P}}}_1\mathbf{X}_2)^-\mathbf{X}'_2\tilde{\underline{\mathbf{P}}}_1\mathbf{y} = \left[\hat{\boldsymbol{\beta}}_2(\tilde{\gamma})\right]'\mathbf{X}'_2\tilde{\underline{\mathbf{P}}}_1\mathbf{y}, \tag{57}$$

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

Il est intéressant de comparer cette statistique à celle de Wald appliquée au même test d'hypothèses. Par application de (51), il vient :

$$W = [\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}})]' (\mathbf{X}_2' \underset{\sim}{\tilde{\mathbf{P}}}_1 \mathbf{X}_2)^- \hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}}), \tag{58}$$

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

$$W = [\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}})]' \mathbf{X}_2' \underset{\sim}{\tilde{\mathbf{P}}}_1 \mathbf{y}. \tag{59}$$

À l'examen de (57) et de (59), 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, p. 487).

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 $\hat{\mathbf{V}}_i = (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})'$.

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.

3. Méthode dite REML

31. Exemple simple

311. 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 $y_i \sim_{\mathrm{iid}} \mathrm{N}(\mu, \sigma^2)$ 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 :

$$-2\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = N(\ln 2\pi + \ln\sigma^2) + \sum_{i=1}^{N} (y_i - \mu)^2/\sigma^2. \tag{60}$$

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

$$\sum_{i=1}^{N} (y_i - \mu)^2 = N\left[s^2 + (\overline{y} - \mu)^2\right], \tag{61}$$

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

$$-2\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = N\left[\ln 2\pi + \ln\sigma^2 + \frac{s^2 + (\overline{y} - \mu)^2}{\sigma^2}\right], \tag{62}$$

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

$$\partial(-2\mathrm{L}/N)/\partial\mu = -2(\overline{y} - \mu)/\sigma^2, \tag{63a}$$

$$\partial(-2\mathrm{L}/N)/\partial\sigma^2 = \frac{1}{\sigma^2} - \frac{s^2 + (\overline{y} - \mu)^2}{\sigma^4}. \tag{63b}$$

Par annulation de ces dérivées, on obtient :

$$\hat{\mu} = \overline{y}, \tag{64}$$

Et, pour $N \geqslant 2$,

$$\hat{\sigma}^2_{ML} = s^2 + (\overline{y} - \hat{\mu})^2 = s^2. \tag{65}$$

Or

$$\mathrm{E}(\hat{\sigma}^2_{ML}) = (N-1)\sigma^2/N \tag{66}$$

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).

312. 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.

3121. 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. À cet égard, on considère la transformation biunivoque suivante :

$$\mathbf{y}_{(N\times 1)} = \{y_i\} \leftrightarrow \mathbf{y}^*_{(N\times 1)} = (\mathbf{z}'_{N-1}, \overline{y})' \tag{67}$$

où $\mathbf{z}_{N-1} = \{z_i = y_i - \overline{y};\ i = 1, 2, \ldots, N-1\}$, 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 :

$$\mathrm{p}_Y(\mathbf{y}|\mu, \sigma^2) = \mathrm{p}_{Y^*}(\mathbf{y}^*|\mu, \sigma^2)|J| \tag{68}$$

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

Or, $\overline{y}$ 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 :

$$\mathrm{p}_{Y^*}(\mathbf{y}^*|\mu, \sigma^2) = \mathrm{p}_Z(\mathbf{z}_{N-1}|\sigma^2)\mathrm{p}_{\overline{Y}}(\overline{y}|\mu, \sigma^2), \tag{69}$$

Par ailleurs, eu égard à la définition de la transformation (67), la valeur du jacobien J ne dépend pas des paramètres ; on en déduit donc la décomposition suivante de la logvraisemblance $\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = \ln \mathrm{p}_Y(\mathbf{y}|\mu, \sigma^2)$:

$$\boxed{\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = \mathrm{L}_1(\sigma^2; \mathbf{z}_{N-1}) + \mathrm{L}_2(\mu, \sigma^2; \overline{y}) + cste}, \tag{70}$$

où $\mathrm{L}_1(\sigma^2; \mathbf{z}_{N-1}) = \ln \mathrm{p}_Z(\mathbf{z}_{N-1}|\sigma^2)$, $\mathrm{L}_2(\mu, \sigma^2; \overline{y}) = \ln \mathrm{p}_{\overline{Y}}(\overline{y}|\mu, \sigma^2)$, 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 $\mathbf{z}_{N-1} \sim \mathrm{N}(0, \mathbf{V}_Z)$ 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 (70), on montre que :

$$\boxed{-2\mathrm{L}_1(\sigma^2; \mathbf{z}_{N-1}) = (N-1)(\ln 2\pi + \ln\sigma^2) - \ln N + \Big[\sum_{i=1}^{N}(y_i - \overline{y})^2\Big]/\sigma^2}. \quad (71)$$

Il s'en suit que :

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

et, par annulation :

$$\hat{\sigma}^2 = Ns^2/(N-1);\ N \geqslant 2 \qquad \textbf{(72)}$$

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

3122. 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 (65) : $\hat{\sigma}^2 = s^2 + (\overline{y} - \mu)^2$ 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 $\mathrm{E}\big[(\overline{y} - \mu)^2 \big| \mathbf{y}, \sigma^2\big]$ à l'instar de ce qui est fait avec l'algorithme EM (Foulley, 1993).

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

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

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 (72) 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 $\overline{y}$.

32. Cas général

321. 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 $f(\mathbf{u}, \mathbf{v}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = f(\mathbf{v}|\boldsymbol{\beta}, \gamma)g(\mathbf{u}|\mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ 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 [2] 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).

322. Application au modèle linéaire mixte gaussien

Dans le cadre du modèle $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, 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 » [3] dans la littérature anglo-saxonne. Par définition, la variable $\mathbf{v}$ est ancillaire de $\boldsymbol{\beta}$ puisque $\mathbf{S}\mathbf{X} = \mathbf{0}$ 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 $\tilde{\mathbf{v}} = \tilde{\mathbf{S}}\mathbf{y} = (\mathbf{I}_N - \tilde{\mathbf{H}})\mathbf{y}$ 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

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

3. Car $\underline{\mathbf{H}}\mathbf{y} = \hat{\mathbf{y}}$.

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, p. 251) de prendre $\mathbf{K}'$ sous la forme $\mathbf{WS}$ où $\mathbf{W}$ est une matrice $[N - \mathrm{r}(\mathbf{X})] \times N$ 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 $N \times [N - \mathrm{r}(\mathbf{X})]$ 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 :

$$-2\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}) = \big[N - \mathrm{r}(\mathbf{X})\big]\ln 2\pi + \ln|\mathbf{K}'\mathbf{V}\mathbf{K}| + \mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}. \quad (73)$$

Cette expression va grandement se simplifier du fait des relations suivantes (Searle, 1979, p. 2.14 à 2.17 ; Quaas, 1992 ; Rao et Kleffe, 1988, p. 247) :

$$|\mathbf{K}'\mathbf{V}\mathbf{K}| = |\mathbf{V}||\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}||\underline{\mathbf{X}}'\underline{\mathbf{X}}|^{-1}|\mathbf{K}'\mathbf{K}|, \quad (74a)$$

$$\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}' = \underline{\mathbf{P}} \quad (74b)$$

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 $\underline{\mathbf{P}} = \mathbf{V}^{-1}(\mathbf{I}_N - \mathbf{Q})$ avec $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$.

En insérant (74ab) dans (73), et en isolant la constante C, on obtient l'expression suivante de la logvraisemblance :

$$\boxed{-2\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}) = C + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y}} \quad (75a)$$

avec

$$C = \big[N - \mathrm{r}(\mathbf{X})\big]\ln 2\pi - \ln|\underline{\mathbf{X}}'\underline{\mathbf{X}}| + \ln|\mathbf{K}'\mathbf{K}|. \quad (75b)$$

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

$$C' = \big[N - \mathrm{r}(\mathbf{X})\big]\ln 2\pi - \ln|\underline{\mathbf{X}}'\underline{\mathbf{X}}|, \quad (76a)$$

$$C'' = \big[N - \mathrm{r}(\mathbf{X})\big]\ln 2\pi. \quad (76b)$$

La première (76a) (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 (75a) par rapport à γ_k, il vient :

$$\frac{\partial\big[(-2\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y})\big]}{\partial\gamma_k} = \frac{\partial \ln|\mathbf{V}|}{\partial\gamma_k} + \frac{\partial \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|}{\partial\gamma_k} + \mathbf{y}'\frac{\partial\underline{\mathbf{P}}}{\partial\gamma_k}\mathbf{y} \quad (77)$$

Par manipulation algébrique, on montre que :

$$\frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} + \frac{\partial \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|}{\partial \gamma_k} = \operatorname{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right). \tag{78}$$

De même, à partir de $\underline{\mathbf{P}} = \mathbf{V}^{-1}(\mathbf{I}_N - \mathbf{Q})$, 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

$$\frac{\partial\big[-2\mathrm{L}(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y})\big]}{\partial \gamma_k} = \operatorname{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - \mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\mathbf{y} \tag{79}$$

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 $\operatorname{tr}(\underline{\mathbf{P}}\mathbf{V}_k) = \sum_{l=0}^{K} \operatorname{tr}(\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{V}_l)\gamma_l$ et le système des équations REML s'écrit :

$$\boxed{\sum_{l=0}^{K} \operatorname{tr}(\hat{\underline{\mathbf{P}}}\mathbf{V}_k\hat{\underline{\mathbf{P}}}\mathbf{V}_l)\hat{\gamma}_l = \mathbf{y}'\hat{\underline{\mathbf{P}}}\mathbf{V}_k\hat{\underline{\mathbf{P}}}\mathbf{y}}. \tag{80}$$

En posant

$$\tilde{\mathbf{F}} = \{\tilde{f}_{kl}\} = \{\operatorname{tr}(\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{V}_l)\} \tag{81a}$$

$$\mathbf{G} = \{g_k\} = \{\mathbf{y}'\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{y}\}. \tag{81b}$$

Le système en (80) peut être résolu numériquement par un algorithme itératif ayant la forme d'un système linéaire en $\boldsymbol{\gamma}$:

$$\boxed{\tilde{\mathbf{F}}(\boldsymbol{\gamma}^{[n]})\boldsymbol{\gamma}^{[n+1]} = \mathbf{g}(\boldsymbol{\gamma}^{[n]})}, \tag{82}$$

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, $\tilde{f}_{kl} = \sum_{ij}\{\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{Z}_l\}_{ij}^2$.

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 $\dfrac{\partial\big[\mathrm{L}(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y})\big]}{\partial\gamma_k} = \dfrac{1}{2}\big[\mathbf{y}'\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{y} - \operatorname{tr}(\underline{\mathbf{P}}\mathbf{V}_k)\big]$ des équations REML est par définition nulle, alors que celle du score relatif à la vraisemblance profilée $\dfrac{\partial\big[\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y})\big]}{\partial\gamma_k} = \dfrac{1}{2}\big[\mathbf{y}'\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{y} - \operatorname{tr}(\mathbf{V}^{-1}\mathbf{V}_k)\big]$ 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 (82) 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, p. 6.7; Rao et Kleffe, 1988, p. 236).

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) :

$$\begin{aligned} -\frac{\partial^2 \mathrm{L}}{\partial\gamma_k\partial\gamma_l} = \frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial^2\mathbf{V}}{\partial\gamma_k\partial\gamma_l}\right) - \frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\right) \\ -\frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\left(\frac{\partial^2\mathbf{V}}{\partial\gamma_k\partial\gamma_l} - 2\frac{\partial\mathbf{V}}{\partial\gamma_l}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\right)\underline{\mathbf{P}}\mathbf{y} \end{aligned}, \tag{83}$$

$$\boxed{\mathrm{E}\left(-\frac{\partial^2 \mathrm{L}}{\partial\gamma_k\partial\gamma_l}\right) = \frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\right)}. \tag{84}$$

Comme pour ML, on montre aisément que le système des équations des scores de Fisher équivaut dans le cas linéaire au système (82) au coefficient 1/2 près.

La complémentarité des formules (83) et (84) 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

$$AI_{kl} = \frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\underline{\mathbf{P}}\mathbf{y}. \tag{85}$$

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

323. 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 :

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}) = \int \mathrm{p}(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\beta}, \tag{86}$$

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

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

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}) = \int \mathrm{p}(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\beta}, \tag{87}$$

où $p(\mathbf{y}|\boldsymbol{\beta},\gamma)$ est la densité usuelle des observables sachant les paramètres et $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ est la densité a priori de $\boldsymbol{\beta} \in R^p$ 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 $\mathbf{y} \sim N(\underline{\mathbf{X}}\boldsymbol{\beta}, \mathbf{V})$, la densité $p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})$ s'écrit :

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

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) :

$$(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta}) = (\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), \tag{88}$$

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

$$p(\mathbf{y}|\boldsymbol{\gamma}) = (2\pi)^{-N/2}|\mathbf{V}|^{-1/2} \exp\left[-(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})/2\right] \int \exp\left[-(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})/2\right] d\boldsymbol{\beta} \tag{89}$$

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

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

L'intégrale en (89) est donc égale à $(2\pi)^{p/2}|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|^{-1/2}$ d'où l'expression de la densité marginale :

$$\boxed{p(\mathbf{y}|\boldsymbol{\gamma}) = (2\pi)^{-(N-p)/2}|\mathbf{V}|^{-1/2}|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|^{-1/2} \exp\left[-(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})/2\right]}, \tag{90}$$

dont moins deux fois le logarithme est bien identique à (75a) 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 :

$$\hat{\boldsymbol{\gamma}}_{REML} = \text{argmax}_{\gamma\in\Gamma} \ln p(\mathbf{y}|\boldsymbol{\gamma}), \tag{91a}$$

$$\hat{\boldsymbol{\gamma}}_{REML} = \text{argmax}_{\gamma\in\Gamma} \ln \pi(\boldsymbol{\gamma}|\mathbf{y}). \tag{91b}$$

33. Aspects calculatoires

331. 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 :

$$\sigma_k^{2[t+1]} = \{\hat{\mathbf{u}}_k'(\boldsymbol{\eta}^{[t]})\hat{\mathbf{u}}_k(\boldsymbol{\eta}^{[t]}) + \mathrm{tr}[\mathbf{C}_{kk}(\boldsymbol{\eta}^{[t]})]\sigma_0^{2[t]}\}/q_k, \tag{92a}$$

$$\sigma_0^{2[t+1]} = [\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\boldsymbol{\eta}^{[t]})\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'(\boldsymbol{\eta}^{[t]})\mathbf{Z}'\mathbf{y}]/[N - \mathrm{r}(\mathbf{X})] \tag{92b}$$

et, pour l'algorithme d'Harville :

$$\sigma_k^{2[t+1]} = [\hat{\mathbf{u}}_k'(\boldsymbol{\eta}^{[t]})\hat{\mathbf{u}}_k(\boldsymbol{\eta}^{[t]})]/\{q_k - \mathrm{tr}[\mathbf{C}_{kk}(\boldsymbol{\eta}^{[t]})]/\eta_k^{[t]}\}, \tag{92c}$$

où $\boldsymbol{\eta}^{[t]} = \{\sigma_k^{2[t]}/\sigma_0^{2[t]}\}$ est, comme précédemment, le vecteur des rapports de variance des K facteurs aléatoires à la variance résiduelle à l'itération n, $\hat{\mathbf{u}}_k(\boldsymbol{\eta}^{[t]})$ 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 (92ac) restent inchangées. Il en est de même pour la variance résiduelle à la nuance importante près que, pour REML, $[N - \mathrm{r}(\mathbf{X})]$ se substitue à N au dénominateur de (92b).

332. Algorithme EM

Nous placerons pour simplifier la présentation dans le cadre du modèle linéaire mixte à un seul facteur aléatoire $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{Z}'\sigma_u^2 + \mathbf{I}_N\sigma_0^2)$, les formules se généralisant ensuite aisément au cas de plusieurs facteurs indépendants.

Les données observables (ou données incomplètes dans la terminologie EM) sont constituées du vecteur $\mathbf{y}$ et on peut prendre pour données complètes le vecteur $\mathbf{x} = (\boldsymbol{\beta}', \mathbf{u}', \mathbf{e}')'$. Ici, à l'instar de Dempster, Laird et Rubin (1977), $\boldsymbol{\beta}$ n'est pas considéré comme un paramètre mais comme une variable aléatoire parasite dont la variance tend vers une valeur limite infinie (*cf.* l'interprétation bayésienne de REML au §323). Vis-à- vis du vecteur des données complètes supposé gaussien, les statistiques exhaustives de σ_u^2 et σ_0^2 sont les formes quadratiques $\mathbf{u}'\mathbf{u}$ et $\mathbf{e}'\mathbf{e}$. Examinons le cas de σ_u^2. Les phases E (« Expectation ») et M (« Maximization ») se formulent ainsi :

– phase E : expression de l'espérance conditionnelle de σ_u^2 sachant $\mathbf{y}$ et les valeurs courantes des paramètres $\boldsymbol{\gamma}^{[t]} = (\sigma_u^{2[t]}, \sigma_0^{2[t]})'$, c'est-à-dire

$$\nu^{[t]} = \mathrm{E}(\mathbf{u}'\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma}^{[t]}). \tag{93}$$

– phase M : obtention de $\sigma_u^{2[t+1]}$ en égalant l'espérance *a priori* de $\mathbf{u}'\mathbf{u}$ à $\nu^{[t]}$, soit $\mathrm{E}(\mathbf{u}'\mathbf{u}|\boldsymbol{\gamma}^{[t]}) = \nu^{[t]}$ ce qui conduit à :

$$q\sigma_u^{2[n+1]} = \nu^{[n]} \tag{94}$$

où $q = \dim(\mathbf{u})$.

Explicitons donc $\nu^{[t]}$. Par définition

$$\mathrm{E}_c^{[t]}(\mathbf{u}'\mathbf{u}) = \mathrm{E}(\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{[t]})'\mathrm{E}(\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{[t]}) + \mathrm{tr}\big[\mathrm{var}(\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{[t]})\big]. \tag{95}$$

Or,

$$\mathrm{E}(\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{[t]}) = \hat{\mathbf{u}}(\boldsymbol{\gamma}^{[t]}) \tag{96}$$

est le BLUP de $\mathbf{u}$ basé sur $\boldsymbol{\gamma}^{[t]} = (\sigma_0^{2[t]}, \sigma_1^{2[t]})'$ soit la solution du système suivant :

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}^{[t]}) \\ \hat{\mathbf{u}}(\boldsymbol{\gamma}^{[t]}) \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

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

Par ailleurs

$$\mathrm{var}(\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{[t]}) = \mathrm{var}(\hat{\mathbf{u}}^{[t]} - \mathbf{u}) = \mathbf{C}_{uu}(\boldsymbol{\gamma}^{[t]})\sigma_0^{2[t]}, \tag{97}$$

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

$$\mathbf{C}(\boldsymbol{\gamma}^{[t]}) = \begin{bmatrix} \mathbf{C}_{\beta\beta}(\boldsymbol{\gamma}^{[t]}) & \mathbf{C}_{\beta u}(\boldsymbol{\gamma}^{[t]}) \\ \mathbf{C}_{u\beta}(\gamma^{[t]}) & \mathbf{C}_{uu}(\gamma^{[t]}) \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda^{[t]}\mathbf{I}_q \end{bmatrix}^{-}.$$

En reportant (96) et (97) dans (95), on obtient une expression identique à celle de l'algorithme de type « Henderson »

$$\boxed{\sigma_u^{2[t+1]} = \big\{\hat{\mathbf{u}}'(\boldsymbol{\gamma}^{[t]})\hat{\mathbf{u}}(\boldsymbol{\gamma}^{[t]}) + \mathrm{tr}\big[\mathbf{C}_{uu}(\boldsymbol{\gamma}^{[t]})\big]\sigma_0^{2[t]}\big\}/q}, \tag{98}$$

qui se généralise à plusieurs facteurs comme indiqué en (92a).

Le même raisonnement s'applique à la variance résiduelle σ_0^2 qui aboutit à $N\sigma_0^{2[t+1]} = \mathrm{E}(\mathbf{e}'\mathbf{e}|\boldsymbol{\gamma}^{[t]})$. Par manipulation algébrique, cette formule s'explicite en

$$\boxed{\begin{aligned} \sigma_0^{2[t+1]} = \big\{&\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}'(\boldsymbol{\gamma}^{[t]})\mathbf{T}'\mathbf{y} - \lambda^{[t]}\hat{\mathbf{u}}'(\boldsymbol{\gamma}^{[t]})\hat{\mathbf{u}}(\boldsymbol{\gamma}^{[t]}) \\ &+\big[\mathrm{r}(\mathbf{X}) + q - \lambda^{[t]}\mathrm{tr}\big[\mathbf{C}_{uu}(\boldsymbol{\gamma}^{[t]})\big]\sigma_0^{2[t]}\big]\big\}/N \end{aligned}} \tag{99}$$

où $\mathbf{T} = (\mathbf{X}, \mathbf{Z})$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}')'$ et $\lambda^{[t]} = \sigma_0^{2[t]}/\sigma_u^{2[t]}$.

L'expression de l'équation de récurrence EM pour la variance résiduelle s'avère plus complexe que celle de l'algorithme de type Henderson. En fait, comme l'ont montré Foulley and van Dyk (2000), l'algorithme EM-REML

d'Henderson peut s'interpréter comme une variante de l'algorithme EM dit ECME («Expectation Conditional Maximization Either»).

Il s'agit là de la présentation de l'algorithme classique qui malheureusement peut s'avérer assez lent dans certaines situations. On peut améliorer très significativement ses performances par standardisation des effets aléatoires (Foulley et Quaas, 1995; Meng et van Dyk, 1998) et, plus généralement, grâce à la technique d'expansion paramétrique (Liu, Rubin et Wu, 1998; van Dyk, 2000; Foulley et van Dyk, 2000) qui apparaît très prometteuse y compris dans ses prolongements stochastiques (Liu et Wu ,1999; van Dyk et Meng, 2001).

333. Calcul de $-2RL$

Reprenons l'expression (75ab) de la logvraisemblance résiduelle soit, en reprenant la notation de Welham et Thompson :

$$-2\mathrm{RL} = [N - \mathrm{r}(X)]\ln 2\pi + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \qquad (100)$$

On a déjà montré (*cf.* (40)) 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 $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}')'$ est la solution des équations dites du modèle mixte

$$(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \Sigma^{-})\hat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{R}^{-1}\mathbf{y} \text{ avec } \mathbf{T} = (\underline{\mathbf{X}}, \mathbf{Z}) \text{ et } \Sigma^{-} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} \end{bmatrix}.$$

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

$$|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \Sigma^{-}| = |\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}|\,|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|. \qquad (101)$$

On a aussi montré (41) que :

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

d'où

$$|\mathbf{V}||\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}| = |\mathbf{R}||\mathbf{G}||\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \Sigma^{-}|. \qquad (102)$$

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

$$\boxed{\begin{aligned} -2\mathrm{RL} = {} & [N - \mathrm{r}(X)]\ln 2\pi + \ln|\mathbf{R}| + \ln|\mathbf{G}| + \ln|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \Sigma^{-}| \\ & + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y} \end{aligned}} \qquad (103)$$

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 (103), $\mathbf{R}$ 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 $\ln|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \Sigma^-|$. Celles-ci se résorbent en partie en ayant recours à une transformation de Cholesky $\mathbf{E}\mathbf{E}' = \mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \Sigma^-$ de la matrice des coefficients, si bien que $\ln|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \Sigma^-| = 2\sum_{j=1}^{rg(\mathbf{E})} \ln e_{jj}$ où les e_{jj} sont les termes diagonaux de $\mathbf{E}$.

34. Vraisemblance résiduelle et tests

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

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\mathbf{X}'\big[\mathbf{V}(\hat{\boldsymbol{\gamma}})\big]^{-1}\mathbf{y}. \tag{104}$$

et sa variance d'échantillonnage :

$$\mathrm{var}\big[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}})\big] = \mathrm{var}\big[\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\big] + \mathrm{E}\big\{\big[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\big]\big[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\beta}}(\gamma)\big]'\big\} \tag{105}$$

Cette formule montre clairement que l'estimateur usuel $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) = \big[\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\gamma}})\mathbf{X}\big]^{-1}$ pose problème puisqu'à la fois $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ diffère du premier terme $\mathrm{var}\big[\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\big] = \boldsymbol{\Phi}(\boldsymbol{\gamma})$ (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 $\boldsymbol{\Phi}_A(\boldsymbol{\gamma}) = \boldsymbol{\Phi}(\boldsymbol{\gamma}) + \Lambda(\boldsymbol{\gamma})$, 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 $\hat{\Lambda}$ de la correction Λ. Comme $\mathbf{E}\big[\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\big] \neq \boldsymbol{\Phi}(\boldsymbol{\gamma})$, il faut faire également un ajustement pour le biais $\mathbf{B} = \mathbf{E}\big[\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\big] - \boldsymbol{\Phi}(\boldsymbol{\gamma})$. 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 $\mathbf{W} = \mathrm{Var}(\hat{\boldsymbol{\gamma}})$ et,

$$\frac{\partial^2 \boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} = \boldsymbol{\Phi}(\mathbf{P}_k \boldsymbol{\Phi} \mathbf{P}_l + \mathbf{P}_l \boldsymbol{\Phi} \mathbf{P}_k - \mathbf{Q}_{kl} - \mathbf{Q}_{lk} + \mathbf{R}_{kl})\boldsymbol{\Phi} \tag{106}$$

avec

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

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

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

$$\Lambda \approx \boldsymbol{\Phi}\left[\sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}(\mathbf{Q}_{kl} - \mathbf{P}_k \boldsymbol{\Phi} \mathbf{P}_l)\right]\boldsymbol{\Phi} \tag{107}$$

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 $\mathbf{B} = -\Lambda$ ce qui implique $\hat{\boldsymbol{\Phi}} = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) - \hat{\mathbf{B}}$; comme $\hat{\boldsymbol{\Phi}}_A = \hat{\boldsymbol{\Phi}} + \hat{\Lambda}$, on aboutit en définitive à :

$$\boxed{\hat{\boldsymbol{\Phi}}_A = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) + 2\hat{\Lambda}}. \tag{108}$$

Rappelons que $\mathbf{W}$ peut être approché par l'inverse $\mathbf{J}^{-1}(\boldsymbol{\gamma})$ de la matrice d'information de Fisher soit avec REML : $\mathbf{J}(\boldsymbol{\gamma}) = 1/2\tilde{\mathbf{F}}(\boldsymbol{\gamma}) = \{1/2\mathrm{tr}(\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{V}_l)\}$. Mais on peut également utiliser la matrice d'information observée (83) ou d'information moyenne (85). 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 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

$$\boxed{F^* = \lambda F} \tag{109}$$

où

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

– λ un facteur d'échelle ($0 < \lambda \leqslant 1$) de la forme $\lambda = m/(m + r - 1)$ 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 $\mathrm{F}(r, m)$; ils s'imposent de surcroît que ce soit une distribution exacte $\mathrm{F}(r, m)$ 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 H_0 : $\mathbf{k}'\boldsymbol{\mu} = \mathbf{0}$ sous le modèle multidimensionnel $\mathbf{Y}_i \sim_{\mathrm{iid}} \mathrm{N}_m(\boldsymbol{\mu}, \Sigma)$; $i = 1, 2, \ldots, N$, (Rao, 1973, p. 564-565). La statistique de Hotelling s'écrit alors :

$$T^2 = \min_{H_0}(\overline{\mathbf{Y}} - \boldsymbol{\mu})'(\mathbf{S}/N)^{-1}(\overline{\mathbf{Y}} - \boldsymbol{\mu}), \tag{110}$$

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

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

$$m = 4 + \frac{r+2}{r\rho - 1} \text{ et } \lambda = \frac{m}{E^*(m-2)} \tag{111}$$

où $\rho = V^*/2E^{*2}$ (112a)

avec $E^* = (1 - A_2/r)^{-1}$; $V^* = \dfrac{2}{r}\dfrac{1 + c_1 B}{(1 - c_2 B)^2(1 - c_3 B)}$ (112b)

$$c_1 = g/\left[3r + 2(1-g)\right]; c_2 = (r-g)/\left[3r + 2(1-g)\right]$$
$$c_3 = (r + 2 - g)/\left[3r + 2(1-g)\right] \tag{112c}$$

pour $g = \left[(r+1)A_1 - (r+4)A_2\right]/\left[(r+2)A_2\right]$. (112d)

$$B = (A_1 + 6A_2)/2r \tag{112e}$$

$$A_1 = \sum_{k=1}^K\sum_{l=1}^K W_{kl}\mathrm{tr}(\boldsymbol{\theta}\boldsymbol{\Phi}\mathbf{P}_k\boldsymbol{\Phi})\mathrm{tr}(\boldsymbol{\theta}\boldsymbol{\Phi}\mathbf{P}_l\boldsymbol{\Phi}) \tag{112f}$$

$$A_2 = \sum_{k=1}^K\sum_{l=1}^K W_{kl}\mathrm{tr}(\boldsymbol{\theta}\boldsymbol{\Phi}\mathbf{P}_k\boldsymbol{\Phi}\boldsymbol{\theta}\boldsymbol{\Phi}\mathbf{P}_l\boldsymbol{\Phi}) \tag{112g}$$

sachant que

$$\boldsymbol{\theta} = \mathbf{k}(\mathbf{k}'\boldsymbol{\Phi}\mathbf{k})^{-1}\mathbf{k}'. \tag{112h}$$

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 Procmixed de SAS (version 8).

342. 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 $-2\mathrm{RL}_m$ 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 $\mathbf{S}\mathbf{y}$ 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 :

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

et

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

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.* 75ab) qu'on peut écrire sous la forme :

$$-2\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}) = C(\underline{\mathbf{X}}_0) + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0| + (\mathbf{y} - \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0)'\mathbf{V}^{-1}(\mathbf{y} - \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0), \qquad (113)$$

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 $\mathbf{K}_0(\mathbf{K}_0'\mathbf{V}\mathbf{K}_0)^{-1}\mathbf{K}_0' = \underline{\mathbf{P}}_0$ où $\underline{\mathbf{P}}_0 = \mathbf{V}^{-1}(\mathbf{I} - \mathbf{Q}_0)$, d'où

$$-2\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}) = C(\underline{\mathbf{X}}_0) + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0| + (\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1)'\underline{\mathbf{P}}_0(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1). \qquad (114)$$

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

$$-2\mathrm{L}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}] = C(\underline{\mathbf{X}}_0) + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0| + \min_{\beta_1}(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1)'\underline{\mathbf{P}}_0(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1)$$

Or, on peut montrer par manipulation matricielle que :

$$\min_{\beta_1}(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1)'\underline{\mathbf{P}}_0(\mathbf{y} - \underline{\mathbf{X}}_1\boldsymbol{\beta}_1) = \min_{\beta}(\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta}) = (\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}), \qquad (115)$$

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 :

$$-2\mathrm{L}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}] = C(\underline{\mathbf{X}}_0) + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0| + (\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}). \qquad (116)$$

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 :

$$A = -2\max_{\gamma}\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}) + 2\max_{\gamma}\mathrm{L}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}]. \qquad (117a)$$

ou, encore

$$A = -2\mathrm{L}(\hat{\boldsymbol{\gamma}}; \mathbf{K}_0'\mathbf{y}) + 2\mathrm{L}[\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\gamma}}), \tilde{\boldsymbol{\gamma}}; \mathbf{K}_0'\mathbf{y}], \qquad (117b)$$

où $\tilde{\boldsymbol{\gamma}} = \arg\max_{\gamma}\mathrm{L}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}]$.
Si, à l'instar de Welham et Thompson, on introduit la notation suivante :

$$-2\mathrm{RL}[\mathbf{y}, \underline{\mathbf{X}}_j\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{S}(\underline{\mathbf{X}}_i)] = C(\underline{\mathbf{X}}_i) + \ln|\underline{\mathbf{X}}_i'\mathbf{V}^{-1}\underline{\mathbf{X}}_i| + \ln|\mathbf{V}| + (\mathbf{y} - \underline{\mathbf{X}}_j\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \underline{\mathbf{X}}_j\boldsymbol{\beta}) \qquad (118)$$

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

$$\boxed{A = -2\mathrm{RL}[\mathbf{y}, \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0(\hat{\boldsymbol{\gamma}}), \hat{\boldsymbol{\gamma}}, \mathbf{S}(\underline{\mathbf{X}}_0)] + 2\mathrm{RL}[\mathbf{y}, \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\gamma}}), \tilde{\boldsymbol{\gamma}}, \mathbf{S}(\underline{\mathbf{X}}_0)]}. \qquad (119)$$

À 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

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

puis, en passant au modèle réduit sur la base de l'expression (118) correspondante :

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

conduisant à la statistique

$$\boxed{D = -2\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}_0\tilde{\tilde{\boldsymbol{\beta}}}_0(\tilde{\tilde{\boldsymbol{\gamma}}}), \tilde{\tilde{\boldsymbol{\gamma}}}, \mathbf{S}(\underline{\mathbf{X}})\right] + 2\mathrm{RL}\left[\mathbf{y}, \underline{\mathbf{X}}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}), \hat{\boldsymbol{\gamma}}, \mathbf{S}(\underline{\mathbf{X}})\right]}, \qquad (120)$$

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.

343. 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 $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{Z}'\sigma_u^2 + \mathbf{I}_N\sigma_e^2)$ 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 $L_C = \mathrm{Max}_{\sigma_u^2\geqslant 0, \sigma_e^2>0}\mathrm{RL}(\sigma_u^2, \sigma_e^2; \mathbf{y})$ 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 ($C : \sigma_u^2 \geqslant 0, \sigma_e^2 > 0$), 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 :

$$\boxed{\lambda \underset{L}{\rightarrow} 1/2D_0 + 1/2\chi_1^2}. \tag{121}$$

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

$$\Pr(\chi_1^2 \geqslant s) = 2\alpha \tag{122}$$

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 $H_0 : \Sigma = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & 0 \end{pmatrix}$ vs $H_1 : \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$, 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 $\text{var}(\mathbf{u}_1', \mathbf{u}_2')' = \Sigma \otimes \mathbf{I}_q$. Ce modèle se rencontre dans l'analyse de données longitudinales (Laird et Ware, 1982; Diggle *et al.*, 1994). Si l'on contraint Σ sous H_1 à être définie semi-positive, alors, (Stram et Lee , 1994)

$$\lambda \underset{L}{\rightarrow} 1/2\chi_1^2 + 1/2\chi_2^2. \tag{123}$$

De la même façon, on généralise ensuite au cas du test $H_0 : \Sigma = \begin{pmatrix} \Sigma_{11(q\times q)} & 0 \\ 0 & 0 \end{pmatrix}$ vs $H_1 : \Sigma_{(q+1)\times(q+1)} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \sigma_{22} \end{pmatrix}$ définie semi-positive pour lequel $\lambda \underset{L}{\rightarrow} 1/2\chi_q^2 + 1/2\chi_{q+1}^2$.

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. À 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.

Remerciements

J.-L. Foulley tient à remercier les responsables et les étudiants de la section « Biostatistiques » de l'ENSAI de Rennes et ceux du DEA de Génétique multifactorielle pour lui avoir confié et avoir suivi un enseignement sur le

modèle linéaire mixte et l'estimation des composantes de la variance sans lequel cet article de synthèse n'aurait pas vu le jour.

Les trois auteurs expriment leur gratitude à Mrs Bernard Bonaïti, Jean-Jacques Colleau, Jorge Colaco et les lecteurs mandatés par la revue pour leur examen critique du manuscrit.

RÉFÉRENCES

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.

DEMPSTER A., LAIRD N., RUBIN R. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* B, 39,1-20.

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., 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.

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.

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 J.S.J. (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.

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.

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 Statistique*, 143, 1-2, 155-165.

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 S.R. (1982), *Matrix algebra useful for statistics*, J. 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.

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.

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.

YATES F. (1934), The analysis of multiple classifications with unequal numbers in the different classes, *Journal of the American Statistical Association*, 29, 51-66.

ANNEXE 1

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 :

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

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 $\lambda_i \geqslant 0$ pour $i = 1$ à p et des $\mu_j \in \mathbb{R}$ pour $j = 1$ à q tels que :

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

Ce sont les conditions de Karush, Kühn, Tucker qui sont 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 $a \geqslant 0$. 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 :

$$\sup_{\alpha \in (C_\Gamma - \alpha_0)} [-(Z-\alpha)^T I(\alpha_0)(Z-\alpha)] - \sup_{\alpha \in (C_{\Gamma_0} - \alpha_0)} [-(Z-\alpha)^T I(\alpha_0)(Z-\alpha)] \tag{I.3}$$

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 :

$$\inf_{\alpha \in \tilde{C}_0} ||\tilde{Z} - \alpha||^2 - \inf_{\alpha \in \tilde{C}} ||\tilde{Z} - \alpha||^2 \tag{I.4}$$

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

$$Y = X\beta + Z_1 u_1 + Z_2 u_2 + e \tag{I.5}$$

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 e est le vecteur des erreurs résiduelles. On suppose que $(u_1, u_2)^T$ est un vecteur gaussien centré de variance

$$Var \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \sigma_{11}^2 I_q & \sigma_{12} I_q \\ \sigma_{12} I_q & \sigma_{22}^2 I_q \end{pmatrix} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix} \bigotimes I_q = \Sigma \bigotimes I_q \tag{I.6}$$

On suppose que e est gaussien centré de variance $\sigma_e^2 I_N$ 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 $\sigma_{11}^2 \sigma_{22}^2 - \sigma_{12}^2 \geqslant 0$ 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 $\sigma_{11}^2 \sigma_{22}^2 - \sigma_{12}^2 \geqslant 0$. 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 (Figure 1).

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 $\alpha_1 \geqslant 0$. On obtient alors :

$$\inf_{\alpha \in \tilde{C}_0} ||\tilde{Z} - \alpha||^2 - \inf_{\alpha \in \tilde{C}} ||\tilde{Z} - \alpha||^2 = \tilde{Z}_1^2 I_{\tilde{Z}_1 > 0} + \tilde{Z}_2^2 \tag{I.7}$$

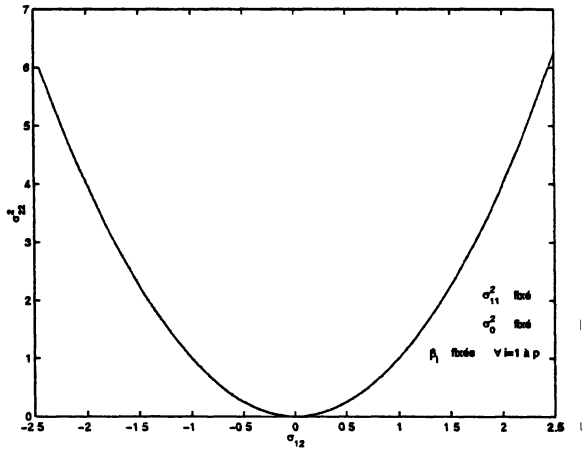

FIG. 1. – Représentation de l'espace paramétrique pour σ_{11}, σ_0 et β_i, $\forall i = 1 \cdots p$, fixés.

qui suit une loi $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$. Ce résultat se généralise au test de q contre $q+1$ effets aléatoires pour lequel la loi asymptotique de la statistique de test est $\frac{1}{2}\chi^2_q + \frac{1}{2}\chi^2_{q+1}$ 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

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = N\ln(2\pi) + \ln|\mathbf{V}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{II.1}$$

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 :

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \tag{II.2}$$

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \tag{II.3}$$

On en déduit l'expression des dérivées partielles secondes

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \tag{II.4}$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}\partial \gamma_k} = 2\mathbf{X}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \tag{II.5}$$

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} &= \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l}\right) - \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right) \\ &\quad - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\left(\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} - 2\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right)\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\end{aligned} \tag{II.6}$$

En divisant par deux, ces formules fournissent les termes qui permettent de calculer la matrice d'information dite observée $\mathbf{I}(\hat{\boldsymbol{\alpha}};\mathbf{y}) = -\dfrac{\partial^2 \mathrm{L}(\boldsymbol{\alpha};\mathbf{y})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}\Big|_{\alpha=\hat{\alpha}}$ 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 :

$$\mathbf{J}_{\beta\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \tag{II.7}$$

$$\mathbf{J}_{\beta\gamma} = \mathbf{0}, \tag{II.8}$$

$$(\mathbf{J}_{\gamma\gamma})_{kl} = \frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \tag{II.9}$$

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

$$r(\boldsymbol{\gamma}) = \left[N - \mathrm{r}(\mathbf{X})\right]\ln 2\pi + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \qquad \text{(II.10)}$$

où $r(\boldsymbol{\gamma}) = -2\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y})$.
En différenciant par rapport à γ_k, on obtient :

$$\frac{\partial r(\boldsymbol{\gamma})}{\partial \gamma_k} = \frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} + \frac{\partial \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|}{\partial \gamma_k} + \mathbf{y}'\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k}\mathbf{y} \qquad \text{(II.11)}$$

Or,

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

$$\frac{\partial \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|}{\partial \gamma_k} = -\mathrm{tr}\left[(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}})^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\underline{\mathbf{X}}\right]$$

$$= -\mathrm{tr}\left[\mathbf{V}^{-1}\underline{\mathbf{X}}(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}})^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right],$$

ce qui permet de faire apparaître et de factoriser la matrice $\underline{\mathbf{P}}$, soit

$$\frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} + \frac{\partial \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|}{\partial \gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) \qquad \text{(II.12)}$$

Il reste à expliciter $\dfrac{\partial \underline{\mathbf{P}}}{\partial \gamma_k}$. Par définition, $\mathbf{V}\underline{\mathbf{P}} = (\mathbf{I} - \mathbf{Q})$ avec $\mathbf{Q} = \underline{\mathbf{X}}(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}})^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}$. 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), $\dfrac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\mathbf{V}^{-1}(\mathbf{I} - \mathbf{Q})\dfrac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}$, 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 :

$$\begin{aligned}\frac{\partial^2 r(\boldsymbol{\gamma};\mathbf{y})}{\partial \gamma_k \partial \gamma_l} &= \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l}\right) - \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right) \\ &\quad -\mathbf{y}'\underline{\mathbf{P}}\left(\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} - 2\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right)\underline{\mathbf{P}}\mathbf{y}\end{aligned} \tag{II.16}$$

La matrice d'information de Fisher s'en déduit immédiatement

$$(\mathbf{J}_{\gamma\gamma})_{kl} = \frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \tag{II.17}$$

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) :

$$|\mathbf{A}| = |\mathbf{A}_{11}|\,|\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|$$

Ici, $|\mathbf{A}| = |\mathbf{R}^{-1}|\,|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{Z}| = 1/|\mathbf{R}|\,|\mathbf{G}|$.
De même, par symétrie

$$|\mathbf{A}| = |\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}|\,|\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}|$$
$$|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}|/|\mathbf{V}|$$

d'où $\boxed{|\mathbf{V}| = |\mathbf{R}|\,|\mathbf{G}|\,|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}|}$.