Genet Sel Evol (1991) 23, 309-338

Article original

Approches statistiques de l'évaluation génétique des reproducteurs pour des caractères binaires à seuils

JL Foulley, E Manfredi

Institut national de la recherche agronomique, station de génétique quantitative et appliquée, équipe de génétique statistique, 78352 Jouy-en-Josas cedex, France

(Reçu le 7 novembre 1990; accepté le 29 mai 1991)

Résumé – Cet article présente 3 méthodes statistiques d'estimation des paramètres de position et de dispersion relatifs au modèle à seuils applicable à des caractères à variation phénotypique binaire en structure de modèle mixte des facteurs de variation. Ces méthodes concernent : l'approche linéaire de Grizzle, Starmer et Koch (1969) et son extension bayésienne au modèle mixte; l'approche du modèle linéaire généralisé et de la quasi-vraisemblance de Gilmour, Anderson et Rae (1985–87) et enfin la méthode bayésienne du mode conjoint *a posteriori* (MAP) de Gianola et Foulley (1983). Différents aspects comparatifs de ces 3 méthodes sont abordés en discussion.

évaluation des reproducteurs / variables discrètes / caractères à seuils / théorie asymptotique / modèle linéaire généralisé / quasi vraisemblance / inférence bayésienne / modèle mixte

Summary – Statistical approaches to genetic evaluation for threshold binary traits
This article describes 3 statistical methods of inference about location and dispersion parameters of the threshold model applied to binary traits under a mixed model structure of variation. These methods are : 1), the linear approach of Grizzle, Starmer and Koch (1969) and its Bayesian extension to a mixed model; 2), the quasi-likelihood approach to the generalized linear model as proposed by Gilmour et al *(1985–1987), and 3), the Bayesian method (joint mode* a posteriori-*MAP) of Gianola and Foulley (1983). Different aspects of comparison among these procedures are discussed.*

genetic evaluation / discrete variables / threshold traits / asymptotic theory / generalized linear model / quasi-likelihood / Bayesian inference / mixed model

POSITION DU PROBLÈME

L'évaluation génétique des reproducteurs repose actuellement sur le BLUP (Best linear unbiased prediction, Henderson, 1973) pour les paramètres de position et le REML (Restricted maximum likelihood, Patterson et Thompson, 1971) pour les

paramètres de dispersion. Ces méthodes statistiques se justifient pleinement dans le cadre du modèle linéaire gaussien.

Dans le cas de variables discrètes, l'application directe ou après aménagement du BLUP pose de sérieuses difficultés conceptuelles liées à la dépendance entre la fréquence et la variance des caractéristiques discrètes étudiées (Gianola, 1982; Foulley, 1987; Im *et al*, 1987; Foulley *et al*, 1990a). Quant aux algorithmes de calcul du REML, l'application de ceux-ci aux variables discrètes ne répond qu'à des motifs d'opportunité calculatoire. Pour plus de rigueur, on en est réduit à des estimateurs quadratiques tels que ceux proposés ou discutés notamment par Robertson et Lerner (1949), Landis et Koch (1977), Lavergne (1984) et Freycon (1989) pour un modèle à un seul facteur aléatoire ou par Beitler et Landis (1985) et Foulley (1987) pour un modèle mixte à 2 facteurs. Dans le même esprit de l'analyse de variance figurent les méthodes inférentielles de Taguchi qui sont très usitées dans l'industrie mais peu connues en sélection et qui s'avèrent en tout état de cause très critiquables d'un point de vue théorique (Hamada et Wu, 1990). Par ailleurs, l'analyse des données (Tukey, 1962; Benzecri, 1973), fournit toute une gamme d'outils intéressants pour le traitement statistique des données catégorielles, qui sont particulièrement adaptés à une approche statistique descriptive et exploratoire mais qui se révèlent plus difficiles à exploiter dans une optique inférentielle comme c'est le cas en génétique et sélection.

Le modèle «béta-binomial» des données ou son pendant «Dirichlet-multinomial» pour plusieurs catégories, offre un cadre conceptuel plus rigoureux et intéressant vis-à-vis de l'inférence statistique; il autorise en particulier le développement d'estimateurs du maximum de vraisemblance (Williams, 1975) ou bayésiens (Im, 1982; Foulley et Im, 1989) qui sont étroitement apparentés au BLUP (Quaas et Van Vleck, 1980; Foulley *et al*, 1990a). Malheureusement, ce modèle n'est pas généralisable à une situation plus complexe que celle d'un modèle aléatoire à un seul facteur (Im, 1982).

L'analyse génétique de tels caractères n'a eu de cesse de préoccuper les chercheurs depuis les origines de la génétique. L'expression discrète des phénotypes incline naturellement à une approche factorielle du déterminisme génétique avec, toutefois, de sérieuses difficultés d'ajustement du modèle aux observations (Manfredi, 1990) à moins d'un recours à des concepts *ad hoc* tels que celui de pénétrance et d'expressivité variable par exemple. De même, l'étude de la transmission du caractère d'une génération à la suivante ne peut plus s'appréhender simplement comme en présence d'un caractère continu, par les techniques classiques de régression et de corrélation. Il faut alors analyser des tables de contingence par des indices d'association spécifiques de telles structures (Haberman, 1982; Kendall et Stuart, 1961).

L'idée d'une susceptibilité normale sous-jacente à l'expression du caractère s'est fait jour et s'est développée peu à peu dans l'esprit des chercheurs pour pallier toutes ces difficultés. Pearson (1900, 1904) apparaît comme un pionnier dans ce domaine; fort de sa maîtrise de la distribution multinormale, il introduit le concept de corrélation tétrachorique entre variables discrètes pour quantifier les ressemblances entre apparentés en terme classique de corrélation (Fraser, 1980). Wright (1943a,b) introduit le modèle à seuils pour rendre compte de l'écart à des proportions mendéliennes monofactorielles dans l'analyse de l'hérédité du nombre de doigts du membre postérieur lors de croisements entre lignées de cobaye.

Le formalisme du modèle à seuils est en fait très simple, notamment pour un caractère tout-ou-rien comme le rappelle le développement suivant. Désignons par x la variable aléatoire relative au phénotype sous-jacent d'un individu d'une population donnée; on suppose que x est distribuée sur une échelle continue sous-jacente munie d'un seuil τ, suivant une loi normale $N(\mu, \sigma^2)$, de moyenne μ et de variance σ^2; dans ces conditions, la probabilité qu'un individu tiré au hasard dans la population présente un des phénotypes tout-ou-rien ($y = 1$ par exemple) est donnée par $\pi = \int_{\tau}^{+\infty} (2\pi)^{-1/2} \exp[-(t-\mu)^2/2\sigma^2] dt$; après le changement de variable $t^* = (t-\mu)/\sigma$, cette probabilité s'exprime à partir de la fonction de répartition $\Phi(.)$ de la loi normale par $\pi = \Phi[(\mu - \tau)/\sigma]$ avec pour argument, l'écart standardisé de la moyenne de la population au seuil.

La non linéarité de la relation entre expressions binaire et sous-jacente se manifeste également au niveau des valeurs génétiques définies sur ces 2 échelles. En effet, si l'on suppose un déterminisme génétique sous-jacent purement additif, on peut écrire $x = \mu + a + e$ où $a \sim N(0, \sigma_a^2)$ et $e \sim N(0, \sigma_e^2)$ désignent les effets génétiques et de milieu respectivement; la valeur génétique sur l'échelle binaire (g) correspond par définition au phénotype moyen des individus ayant tous la même valeur génétique (sous-jacente) soit $g = \Pr(x \geq \tau | \mu, a)$. Si l'on place l'origine au seuil ($\tau = 0$), g s'exprime par $g = \Phi[(\mu + a)/\sigma_e]$. On peut alors calculer aisément les moments de cette variable aléatoire, ce qui permet d'expliciter les relations entre paramètres génétiques sur les 2 échelles. Cette variable a pour espérance $\pi = E(g) = \Phi[\mu/(\sigma_a^2 + \sigma_e^2)^{1/2}]$ et pour variance $\sigma_g^2 = \Phi_2(\tilde{\mu}, \tilde{\mu}; h^2) - \Phi^2(\tilde{\mu})$ (Foulley et Im, 1989) où $\tilde{\mu} = \mu/(\sigma_a^2 + \sigma_e^2)^{1/2}$, $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ est l'héritabilité du caractère sur l'échelle sous-jacente et $\Phi_2(\alpha, \beta; \rho)$ est la fonction de répartition de la loi binormale réduite d'arguments α, β et de corrélation ρ. L'expression classique $\sigma_g^2 = h^2 \phi^2(\tilde{\mu})$ donnée par Robertson (1950) et Dempster et Lerner (1950) où $\phi(.)$ est la densité de la loi normale réduite, correspond en fait à une approximation au premier ordre de la formule précédente au voisinage de $h^2 = 0$ (Foulley et Im, 1989). L'utilisation de la formule de Robertson, Dempster et Lerner a été discutée notamment par Van Vleck (1972), Razungles (1977) et Hill et Smith (1977).

D'un point de vue génétique, l'hypothèse de normalité sur un continuum sous-jacent s'accorde bien avec celle d'un déterminisme polygénique classiquement adoptée dans l'étude des caractères quantitatifs. L'analyse génétique des caractères discrets à seuils s'intègre donc naturellement dans le cadre habituel de la génétique quantitative et de ses concepts. Il en résulte une cohérence de l'analyse, en particulier dans l'étude d'un mélange de caractères discrets et continus (Foulley *et al*, 1983; Siminianer et Schaeffer, 1990) et dans celle de caractères à hérédité mixte impliquant un gène majeur et des polygènes (Lalouel *et al* 1983; Foulley et Elsen, 1988; Elsen et Le Roy, 1990). Le caractère attractif de ce modèle s'est concrétisé par de nombreuses applications dans divers secteurs tels que par exemple les suivants :

– sensibilité aux maladies et anomalies congénitales chez l'homme (Falconer, 1965; Curnow et Smith, 1975) comme chez l'animal (*cf* par exemple Sellier et Ollivier, 1982 pour le syndrome dit «des pattes écartées» chez le porc);

– déterminisme génétique et environnemental du sexe (*cf* Bulmer et Bull, 1982 et Bull *et al*, 1982 pour une application chez certains poissons et la tortue);
– caractéristiques de reproduction et d'adaptation en zootechnie telles que la fertilité (Höschele et al, 1987), les difficultés de vêlage des bovins (Meijering, 1984, 1986; Djemali *et al*, 1987; Quaas *et al*, 1988; Hagger et Hoffer, 1990; Manfredi *et al*, 1991a, 1991b), la taille de portée et la survie des agneaux (Petersson et Danell, 1985; Bodin et Elsen, 1989), La gémellité des bovins (Manfredi *et al*, 1990, 1991c; Ron *et al*, 1990), la morphologie des pieds (Gilmour *et al*, 1987) et la qualité de la laine (Thompson *et al*, 1985) chez le mouton.

D'un point de vue statistique, le modèle à seuils est un cas particulier de la théorie des modèles linéaires généralisés (Nelder et Wedderburn, 1972; Mc Cullagh et Nelder, 1989) puisque dans son développement le plus simple d'une variable binaire, il s'explicite grâce à une fonction, de lien «probit» $\Phi^{-1}(\pi)$. De ce fait, le modèle à seuils pourra être abordé dans un cadre statistique très riche (Ducrocq, 1990) qui ouvre sur des applications dépassant largement le domaine de la génétique humaine et de la sélection animale pour s'étendre par exemple à la neurophysiologie et la séismologie (Brillinger, 1985), à la théorie des sondages (Grosbas, 1987), à la psychologie, aux sciences sociales (Hammerle, 1990) et à l'économétrie (Maddala, 1983; Judge *et al*, 1985).

Les modèles utilisés en sélection animale sont classiquement des modèles mixtes des facteurs de variation (Henderson, 1984) impliquant d'une part des effets fixes relatifs à des facteurs environnementaux (année, saison, élevage, type de conduite) et de niveau génétique des populations (effet «groupe») et, d'autre part, des effets aléatoires correspondant aux individus candidats à la sélection et retenus (effet «père» ou «animal» par exemple). De plus, à des fins de sélection, l'inférence statistique porte à la fois sur l'estimation de certains effets fixes et sur la prédiction d'effets aléatoires. Il y a là une originalité qui n'a pas toujours été prise en compte par la statistique générale et qui a motivé un intérêt et des développements statistiques de la part des généticiens quantitatifs.

Aussi cette revue a-t-elle pour but de faire le point sur les principales méthodes d'estimation statistique des paramètres de position et de dispersion intéressant le sélectionneur dans le cadre du modèle à seuils et d'une structure mixte des facteurs de variation. À cette fin, nous considérerons successivement : l'approche linéaire de Grizzle, Starmer et Koch (1969) et son extension bayésienne au modèle mixte; celle du modèle linéaire généralisé et de la quasi-vraisemblance telle que définie par Gilmour *et al* (1985) et enfin, l'approche bayésienne du mode conjoint *a posteriori* (MAP) développée notamment par Gianola et Foulley (1983).

Pour des raisons de simplicité d'exposition, nous limiterons cette présentation au modèle à seuils relatif à des réponses dichotomiques dit «threshold dichotomy distribution» dans la terminologie du généticien Wright (1968) ou «probit normal binomial distribution» dans celle du statisticien Williams (1988). Les facilités ou, au contraire les contraintes d'extension au cas polytomique seront abordés dans la discussion ainsi que d'autres aspects comparatifs de ces 3 méthodes.

MÉTHODE DE GRIZZLE, STARMER ET KOCH ET EXTENSION AU MODÈLE MIXTE (GSK-FI)

Estimation des paramètres de position

Estimation des paramètres en modèle à effets fixes

Si π_j, p_j et n_j désignent la probabilité, la fréquence de réponse $\left[p_j = \left(\sum_{r=1}^{n_j} y_{jr}\right)/n_j = y_{j+}/n_j\right]$ et le nombre d'observations pour la sous-population $j = 1, 2, \ldots, J$, la méthode de Grizzle, Starmer et Koch (1969) est basée sur le développement limité suivant d'une fonction quelconque g des observations (fonction «logit» ou «probit» par exemple dans le cas de variables binaires)

$$g(p_j) = g(\pi_j) + \left[\frac{\partial g(p)}{\partial p}\right]_{p=\pi_j} (p_j - \pi_j) + \varepsilon_j \qquad [1a]$$

En posant un modèle linéaire sur $g(\pi_j)$, soit $g(\pi_j) = \mathbf{x}_j'\boldsymbol{\beta}$ où $\boldsymbol{\beta}$ est le vecteur des effets fixes et $\mathbf{x}_j'$ est la j^e ligne de la matrice d'incidence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_J)'$ des effets $\boldsymbol{\beta}$, l'expression [1a] s'écrit sous la forme

$$g(p_j) = \mathbf{x}_j'\boldsymbol{\beta} + g'(\Pi_j)(p_j - \pi_j) + \varepsilon_j \qquad [1b]$$

ou, en notation matricielle complète

$$\mathbf{g} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}(\mathbf{p} - \boldsymbol{\Pi}) + \boldsymbol{\varepsilon} \qquad [1c]$$

avec

$$\mathbf{g} = \{g(p_j)\}, \quad \mathbf{p} = \{p_j\}, \quad \boldsymbol{\Pi} = \{\Pi_j\}, \quad \boldsymbol{\varepsilon} = \{\varepsilon_j\} \qquad [2a]$$

$$\mathbf{H} = \text{Diag}\left\{\left.\frac{\partial g(p)}{\partial p}\right|_{p=\pi_j}\right\} \qquad [2b]$$

Soit $\mathbf{V} = \text{Var}(\mathbf{p})$ la matrice de covariance des observations (ici la fréquence par classe).

En modèle à effets fixes, on a :

$$\mathbf{V} = \text{Diag}\{\pi_j(1 - \pi_j)/n_j\} \qquad [3a]$$

et, en posant

$$\mathbf{Q} = \mathbf{HVH} \qquad [3b]$$

on montre le résultat asymptotique suivant (Rao, 1973) :

$$\mathbf{Q}^{*-1/2}(\mathbf{g} - \mathbf{X}\boldsymbol{\beta}) \underset{L}{\longrightarrow} \mathrm{N}(\mathbf{0}, \mathbf{I}) \qquad [4]$$

où $\mathbf{Q}^* = \mathbf{Q}(\mathbf{p})$ est un estimateur asymptotiquement sans biais de $\mathbf{Q}$.
De [4], découle l'estimateur (dit «minimum g χ^2») de $\boldsymbol{\beta}$ minimisant

$$(\mathbf{g} - \mathbf{X}\boldsymbol{\beta})'\mathbf{Q}^{*-1}(\mathbf{g} - \mathbf{X}\boldsymbol{\beta}) \qquad [5a]$$

soit

$$\mathbf{X}'\mathbf{Q}^{*-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Q}^{*-1}\mathbf{g} \qquad [5b]$$

et le test de l'hypothèse H_0 : $\mathbf{k}'\boldsymbol{\beta} = \mathbf{m}$ à partir de la statistique :

$$(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m})'[\mathbf{k}'(\mathbf{X}'\mathbf{Q}^{*-1}\mathbf{X})^{-1}\mathbf{k}]^{-1}(\mathbf{k}'\hat{\boldsymbol{\beta}} - \mathbf{m}) \qquad [6]$$

distribuée asymptotiquement sous H_0 comme un $\chi^2_{r(k)}$. L'estimateur $\hat{\boldsymbol{\beta}}$ défini en [5b] est asymptotiquement sans biais et asymptotiquement normal (encore désigné par BAN : «Best Asymptotic Normal»). Dans le cas du modèle à seuils, la transformation g(.) est la fonction inverse de la fonction de répartition de la loi normale appelée habituellement «probit» ou quelquefois «normit» (Kotz et Jonhson, 1985) soit :

$$g(\pi_j) = \Phi^{-1}(\pi_j) \qquad [7a]$$

d'où

$$g'(\pi_j) = \left.\frac{\partial \Phi^{-1}(p)}{\partial p}\right|_{p=\pi_j} = \frac{1}{\phi[\Phi^{-1}(\pi_j)]} \qquad [7b]$$

où ϕ est la densité de la loi normale réduite. La matrice des pondérations $\mathbf{Q}^{*-1}$ de [5ab] s'écrit donc compte tenu des expressions [2b] et [3a],

$$\mathbf{Q}^{*-1} = \text{Diag}\left\{\frac{n_j\phi^2[\Phi^{-1}(p_j)]}{p_j(1-p_j)}\right\} \qquad [8]$$

Extension au modèle mixte

Au vecteur $\boldsymbol{\beta}$ des effets fixes, on substitue un vecteur

$$\boldsymbol{\theta}' = (\boldsymbol{\beta}, \mathbf{u}')' \qquad [9]$$

comportant un vecteur $\boldsymbol{\beta}$ d'effets fixes et un vecteur $\mathbf{u}$ d'effets aléatoires (valeurs génétiques notamment). Le résultat précédent [4] s'applique à la distribution conditionnelle de $g(\mathbf{p}|\boldsymbol{\theta})$, soit

$$\mathbf{Q}^{*-1/2}(\mathbf{g} - \mathbf{T}\boldsymbol{\theta}) \underset{L}{\longrightarrow} N(\mathbf{0}, \mathbf{I}) \qquad [10]$$

où $\mathbf{T} = (\mathbf{X}, \mathbf{Z})$ rassemble la matrice d'incidence $\mathbf{X}$ des effets $\boldsymbol{\beta}$ et celle $\mathbf{Z}$ des effets $\mathbf{u}$.

En supposant qu'*a priori*, $\boldsymbol{\theta}$ soit distribué normalement avec une espérance $\boldsymbol{\alpha}$ et une matrice de covariance $\boldsymbol{\Sigma}$

$$\boldsymbol{\theta} \sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma}) \qquad [11]$$

la distribution *a posteriori* de $\boldsymbol{\theta}$ résultant de [10] et [11] est donc aussi normale

$$\boldsymbol{\theta}|\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\Sigma} \sim \mathrm{N}(\widehat{\boldsymbol{\theta}}, \mathbf{C}) \qquad [12]$$

où $\widehat{\boldsymbol{\theta}}$, espérance et $\mathbf{C}$, variance *a posteriori* sont données par

$$(\mathbf{T}'\mathbf{Q}^{*-1}\mathbf{T} + \boldsymbol{\Sigma}^{-1})\widehat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{Q}^{*-1}\mathbf{g} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \qquad [13]$$

$$\mathbf{C} = (\mathbf{T}'\mathbf{Q}^{*-1}\mathbf{T} + \boldsymbol{\Sigma}^{-1})^{-1} \qquad [14]$$

Dans le cas du modèle mixte considéré en [9] et [10] $\boldsymbol{\Sigma}^{-1}$ a pour limite

$$\boldsymbol{\Sigma}^{-} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_u^{-1} \end{bmatrix} \qquad [15a]$$

et [13] se réduit à

$$(\mathbf{T}'\mathbf{Q}^{*-1}\mathbf{T} + \boldsymbol{\Sigma}^{-})\widehat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{Q}^{*-1}\mathbf{g} \qquad [15b]$$

L'expression [15b] revient à écrire les équations BLUP sur les données $\mathrm{g} = \{\Phi^{-1}(p_j)\}$ en prenant pour matrice de covariance résiduelle $\mathbf{Q}^* = \mathrm{Diag}\{p_j(1-p_j)/n_j\phi^2[\Phi^{-1}(p_j)]\}$

Ce raisonnement bayésien formulé par Foulley et Im (1989) conduit à une généralisation de la méthode de Grizzle *et al* (1969) au modèle mixte. Une telle extension avait été suggérée pour les logits par Gianola (1980a) en utilisant, non pas une approche bayésienne comme ici, mais des arguments asymptotiques sur les solutions du modèle mixte.

Une des difficultés avec cette méthode réside dans le traitement des cellules j, pour lesquelles la fréquence de la réponse est extrême ($p_j = 0$ ou 1). Dans ce cas, l'élément diagonal de $\mathbf{Q}^{*-1}$ tend vers 0 puisque $\phi^2/p(1-p)$ tend vers 0 quand p tend vers 0 ou 1. De plus $\mathrm{g}(p)$ tend vers plus ou moins l'infini. Pour éviter cela, on peut suggérer de remplacer les fréquences p_j pour les valeurs extrêmes par $(y_{j+} + 1/2)/(n_j + 1)$ (Mc Cullagh et Nelder, 1989) ou par $(y_{j+} + 1)/(n_j + 2)$. Une autre approche de type «bayésien empirique» consiste dans le calcul d'un estimateur $\widehat{p}_j$ de ce type de cellules remplaçant la fréquence, basé sur les cellules où $p_j \neq 0$ et 1 et qu'on utilisera dans $\mathbf{g}$ et $\mathbf{Q}^*$ du système [15b]. Judge *et al* (1985) suggèrent un estimateur des moindres carrés non pondérés, basé, soit sur un modèle linéaire, soit sur un modèle «normit» des probabilités.

Estimation des paramètres de dispersion

Considérons le vecteur $\mathbf{u}$ des effets aléatoires en [9] distribué suivant une loi normale $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$ et désignons par $\boldsymbol{\gamma}_u$ le vecteur des paramètres dont dépend la matrice de covariance $\boldsymbol{\Sigma}_u$. Par exemple, dans le cas d'un modèle à un seul facteur aléatoire tel le père, $\boldsymbol{\Sigma}_u$ s'écrit $\mathbf{A}\sigma_s^2$ où $\mathbf{A}$ est une matrice égale à 2 fois la matrice de parenté (selon Malécot) entre les pères et $\boldsymbol{\gamma}_u = \sigma_s^2$, est la composante «père» de la variance égale à un quart de la variance génétique additive.

Foulley *et al* (1987a, 1989a) ont montré que le maximum de vraisemblance marginale (REML dans le cas de normalité des observations) pouvait être obtenu par l'algorithme itératif suivant (de l'itération r à $r+1$) :

$$\boldsymbol{\gamma}_u^{[r+1]} = \text{Arg Max}_{\gamma_u \in \Gamma_u} \{E_c^{[r]}[\ln p(\mathbf{u}|\boldsymbol{\gamma}_u)]\} \quad [16]$$

où $E_c^{[r]}(.)$ indique une espérance prise par rapport à la distribution de $\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma}_u^{[r]}$ donc conditionnement au vecteur des données $\mathbf{y}$ et Γ_u représente l'espace paramétrique du vecteur $\boldsymbol{\gamma}_u$.

Nous considérerons un modèle type comportant un vecteur $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_k', \ldots, \mathbf{u}_K')'$ formé de K sous-vecteurs $\mathbf{u}_k$ indépendants et de dimensions différentes tels que $\mathbf{u}_k \sim N(\mathbf{0}, \mathbf{A}_k \sigma_{u_k}^2)$ où $\mathbf{A}_k$ est une matrice $(q_k \times q_k)$ connue, définie-positive et $\sigma_{u_k}^2$ est la composante de variance relative au k^e facteur aléatoire. Pour une telle structure, la densité de $\mathbf{u}$ s'écrit :

$$p(\mathbf{u}|\boldsymbol{\gamma}_u) = \prod_{k=1}^{K} p(\mathbf{u}_k|\sigma_{u_k}^2) \quad [17]$$

$$p(\mathbf{u}_k|\sigma_{u_k}^2) \propto (\sigma_{u_k}^2)^{-q_k/2} \exp(-\mathbf{u}_k' \mathbf{A}_k^{-1} \mathbf{u}_k / 2\sigma_{u_k}^2) \quad [18]$$

Il découle de [17] que la maximisation en [16] revient à résoudre les K équations suivantes :

$$\sigma_{u_k}^{2[r+1]} = \text{Arg Max}_{\sigma_{u_k}^2} \{E_c^{[r]}[\ln p(\mathbf{u}_k|\sigma_{u_k}^2)]\} \quad [19]$$

où $E_c^{[r]}(.)$ indique une espérance prise par rapport à la distribution de $\mathbf{u}_k|\mathbf{y}, \boldsymbol{\gamma}_u^{[r]}$ avec $\boldsymbol{\gamma}_u^{[r]} = \{\sigma_{u_k}^{2[r]}\} : k = 1, 2, \ldots, K$

Or, compte-tenu de [18],

$$E_c[\ln p(\mathbf{u}_k|\sigma_{u_k}^2)] = \text{const.} - (q_k/2)\ln(\sigma_{u_k}^2) - E_c(\mathbf{u}_k' \mathbf{A}_k^{-1} \mathbf{u}_k)/2\sigma_{u_k}^2$$

Cette fonction admet pour dérivée : $(-1/2\sigma_{u_k}^2)[q_k - E_c(\mathbf{u}_k' \mathbf{A}_k^{-1} \mathbf{u}_k)/\sigma_{u_k}^2]$ et est maximum par rapport à $\sigma_{u_k}^2$ pour $\sigma_{u_k}^2 = E_c(\mathbf{u}_k' \mathbf{A}_k^{-1} \mathbf{u}_k)/q_k$, d'où l'algorithme itératif :

$$\sigma_{u_k}^{2[r+1]} = E(\mathbf{u}_k' \mathbf{A}_k^{-1} \mathbf{u}_k | \mathbf{y}, \boldsymbol{\gamma}_u^{[r]})/q_k \quad [20]$$

On a vu en [12] que la distribution de $\mathbf{u}_k|\mathbf{y}, \boldsymbol{\gamma}_u$ était normale avec une espérance et variance données en [13], [14] et [15b]. L'algorithme présenté en [20] peut donc s'expliciter en :

$$\sigma_{u_k}^{2[r+1]} = [\hat{\mathbf{u}}_k' \mathbf{A}^{-1} \hat{\mathbf{u}}_k + \text{tr}(\mathbf{A}^{-1} \mathbf{C}_{uu,kk})]^{[r]}/q_k \quad [21a]$$

où

$$\hat{\mathbf{u}}_k = E(\mathbf{u}_k|\mathbf{y}, \boldsymbol{\gamma}_u); \quad \mathbf{C}_{uu,kk} = \text{Var}(\mathbf{u}_k|\mathbf{y}, \boldsymbol{\gamma}_u) \quad [21b]$$

Ce raisonnement est très général et peut être étendu aisément au cas de 2 vecteurs corrélés $\mathbf{u}_k$ et $\mathbf{u}_l$ de même dimension $q_k = q_l = q$ tels que :

$$\text{Cov}(\mathbf{u}_k, \mathbf{u}_l') = \begin{bmatrix} \sigma_{u_k}^2 & \sigma_{u_k u_l} \\ \sigma_{u_k u_l} & \sigma_{u_l}^2 \end{bmatrix} \otimes \mathbf{A}$$

L'algorithme à appliquer pour obtenir le maximum de vraisemblance marginale de $\sigma_{u_{kl}}$ est une simple extension de [21a] (*cf* démonstration dans Foulley *et al*, 1987a) et s'écrit :

$$\sigma^{2[r+1]}_{u_{kl}} = [\hat{\mathbf{u}}'_k \mathbf{A}^{-1} \hat{\mathbf{u}}_l + \mathrm{tr}\,(\mathbf{A}^{-1} \mathbf{C}_{uu,kl})]^{[r]}/q \qquad [22a]$$

où $\hat{\mathbf{u}}_k$, $\hat{\mathbf{u}}_l$ ont la même définition qu'en [21b] et

$$\mathbf{C}_{uu,kl} = \mathrm{Cov}\,(\mathbf{u}_k, \mathbf{u}'_l | \mathbf{y}, \boldsymbol{\gamma}_u) \qquad [22b]$$

Ce type de situation se rencontre par exemple en sélection animale avec le modèle «père (**s**), grand-père maternel (**t**)». On a alors :

$$\mathbf{u} = (\mathbf{s}', \mathbf{t}')'; \mathrm{Var}(\mathbf{u}) = \begin{bmatrix} \sigma_s^2 & \sigma_{st} \\ \sigma_{st} & \sigma_t^2 \end{bmatrix} \otimes \mathbf{A}$$

où **A** est une matrice égale à 2 fois la matrice de parenté (selon Malécot) entre les mâles; σ_s^2, σ_t^2, σ_{st} sont les composantes de variance et de covariance «père» et «grand-père maternel» interprétables en termes de variances et covariances d'effets génétiques directs et maternels. On rencontre également cette paramétrisation dans une structure de modèle multicaractères impliquant des variances et covariances génétiques entre caractères.

Cet algorithme correspond précisément à l'algorithme EM (initiales de «Expectation-Maximization», *cf* Dempster *et al*, 1977) appliqué à l'estimation des composantes de la variance et de la covariance. On peut, selon les mêmes principes (Foulley *et al*, 1989ab), développer un algorithme au second ordre de type Newton-Raphson ne faisant aussi intervenir que les éléments des équations d'Henderson [15b] (*cf* Annexe A).

MÉTHODE DE LA QUASI-VRAISEMBLANCE ET MODÈLE LINÉAIRE GÉNÉRALISÉ (GAR)

L'approche du modèle linéaire généralisé (Mc Cullagh et Nelder, 1989) pour l'analyse de données tout ou rien a été utilisée dans le cadre du modèle mixte par plusieurs auteurs dont Williams (1982), Gilmour *et al* (1985), Zeger et Liang (1986) Zeger *et al* (1988). Des revues critiques de ce type d'application du modèle linéaire généralisé ont été également effectées par Thompson (1990), Knuiman et Laird (1990) et Ducrocq (1990). Nous nous restreindrons ici à la présentation de la méthode de Gilmour, Anderson et Rae (1985) (en abrégé «GAR») telle qu'elle fut proposée par ces auteurs puis réexaminée par Foulley (1987), Höschele et Gianola (1989) et Foulley *et al* (1990a).

Estimation des paramètres de position

Conformément à la théorie du modèle linéaire généralisé (Mc Cullagh et Nelder, 1989), la probabilité de réponse π_j d'une observation de la classe j est transformée

par une fonction de lien (normit dans le modèle à seuil) qui rend le «prédicteur» linéaire vis-à-vis des variables explicatives. On écrit donc, sachant $\mathbf{u}$:

$$\eta_j = \Phi^{-1}(\pi_j) = \mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u}; \quad j = 1, 2, \ldots, J \qquad [23]$$

où $\boldsymbol{\beta}$ et $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \sigma_u^2)$ sont des vecteurs d'effets fixes et aléatoires comme précédemment.

La formulation donnée en [23] est une extension au modèle mixte de la version originale restreinte au départ aux seuls effets fixés.

Estimation des effets fixes

Si la distribution conditionnelle des réponses binaires ($y_{jr} = 0, 1$) sachant $\boldsymbol{\beta}$ et $\mathbf{u}$ est bien un processus de Bernouilli, $y_{jr}|\boldsymbol{\beta}, \mathbf{u} \sim \mathrm{B}(1, \pi_j)$, la distribution marginale des $y_{jr}|\boldsymbol{\beta}$ (après intégration de $\mathbf{u}$) n'est plus accessible simplement eu égard aux corrélations induites par le vecteur $\mathbf{u}$. On a alors affaire à un processus à variation extrabinomiale ou, de façon plus générale et selon la terminologie anglo-saxonne (Williams, 1988), à un modèle «surdispersé». Dans une telle situation, il s'avère commode d'avoir recours à la théorie de la quasi-vraisemblance (Wedderburn, 1974) pour estimer les effets fixes. La mise en application de cette théorie est particulièrement simple puisqu'elle ne requiert que les expressions de l'espérance ($\boldsymbol{\mu}$) et de la variance (matrice $\mathbf{V}$) des observations ($\mathbf{y}$) en fonction des variables explicatives ($\boldsymbol{\beta}$).

Si l'on définit par commodité les observations comme les fréquences de réponses observées dans la classe j : $p_j = \left(\sum_{r=1}^{n_j} y_{jr}\right)/n_j = y_{j+}/n_j$, on a :

$$\widetilde{\pi}_j = \mathrm{E}(p_j) = \mathrm{E}_u[\Phi(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u})]$$

En utilisant la formule de Curnow (1984), cette espérance s'exprime alors par $\widetilde{\pi}_j = \Phi\{\mathbf{x}_j'\boldsymbol{\beta}/[1 + \sigma_j^2]^{1/2}\}$ avec $\sigma_j^2 = \mathbf{z}_j'\boldsymbol{\Sigma}_u\mathbf{z}_j$. En général, les modèles utilisés en sélection animale conduisent, en l'absence de consanguinité, à des variances homogènes ($\forall j, \sigma_j^2 = \sigma^2$). Ainsi, par exemple, dans un modèle comportant le seul facteur aléatoire père ($\mathbf{s}$), on a $\sigma^2 = \sigma_s^2$, variance entre pères. Dès lors, il est commode de changer la paramétrisation sur les effets fixés, en posant :

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/(1 + \sigma^2)^{1/2} \qquad [24]$$

d'où la formulation simple suivante, identique à la présentation classique du modèle généralisé à effets purement fixés :

$$\widetilde{\pi}_j = \Phi(\widetilde{\eta}_j) \qquad [25a]$$

$$\widetilde{\eta}_j = \mathbf{x}_j'\tilde{\boldsymbol{\beta}} \qquad [25b]$$

De même, en ce qui concerne les éléments de $\mathbf{V}$ matrice de covariance des observations p_j, on montre que (*cf* Annexe B) :

$$\mathrm{Var}\,(p_j) = \{\widetilde{\pi}_j(1 - \widetilde{\pi}_j) + (n_j - 1)[\Phi_2(\widetilde{\eta}_j, \widetilde{\eta}_j; t) - \widetilde{\pi}_j^2]\}/n_j \qquad [26a]$$

$$\mathrm{Cov}\,(p_j, p_{j'}) = \Phi_2[\widetilde{\eta}_j, \widetilde{\eta}_{j'}; (\mathbf{z}_j'\boldsymbol{\Sigma}_u\mathbf{z}_{j'})/(1 + \sigma^2)] - \widetilde{\pi}_j\widetilde{\pi}_{j'} \qquad [26b]$$

où $\Phi_2(a, b; , r)$ est la fonction de répartition de la loi binormale réduite de corrélation r et d'arguments a, b. et $t = \sigma^2/(1 + \sigma^2)$. Dans l'exemple du seul facteur aléatoire «père», $t = \sigma_s^2/(1 + \sigma_s^2) = h^2/4$.

La fonction de quasi-vraisemblance $L(\tilde{\boldsymbol{\beta}}; \mathbf{p})$ est définie par l'équation différentielle :

$$\frac{\partial L(\tilde{\boldsymbol{\beta}}; \mathbf{p})}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{K}(\tilde{\boldsymbol{\beta}})\mathbf{V}^{-1}(\tilde{\boldsymbol{\beta}})[\mathbf{p} - \tilde{\boldsymbol{\Pi}}(\tilde{\boldsymbol{\beta}})] \qquad [27]$$

où

$$\mathbf{K}(\tilde{\boldsymbol{\beta}}) = \left\{ \frac{\partial \tilde{\boldsymbol{\Pi}}'}{\partial \tilde{\boldsymbol{\beta}}} \right\}; \quad \tilde{\boldsymbol{\Pi}} = \{\tilde{\pi}_j\}; \quad \mathbf{p} = \{p_j\}$$

$\mathbf{V}_{(J \times J)} = \mathrm{Var}(\mathbf{p})$ est la matrice de covariance des fréquences observées dont les éléments sont donnés en [26a,b].

La maximisation de $L(\tilde{\boldsymbol{\beta}}; \mathbf{p})$ par rapport $\tilde{\boldsymbol{\beta}}$ s'effectue par résolution d'un algorithme itératif de second ordre (scores par exemple) qui s'écrit, de l'itération r à l'itération $r + 1$:

$$\mathbf{X}'\mathbf{W}^{[r]}\mathbf{X}\tilde{\boldsymbol{\beta}}^{[r+1]} = \mathbf{X}'\mathbf{W}^{[r]}\boldsymbol{\zeta}^{[r]} \qquad [28]$$

où $\mathbf{W}_{[J \times J]}$ est une matrice de pondération définie par

$$\mathbf{W} = \mathbf{D}\mathbf{V}^{-1}\mathbf{D} \qquad [29a]$$

$\boldsymbol{\zeta}_{[J \times 1]}$ est une variable de travail telle que

$$\boldsymbol{\zeta} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{D}^{-1}(\mathbf{p} - \tilde{\boldsymbol{\Pi}}) \qquad [29b]$$

$\mathbf{D}_{[J \times J]}$ est une matrice diagonale de [29a,b] définie par $\mathbf{D} = \mathrm{Diag}\{\partial\tilde{\pi}_j/\partial\tilde{\eta}_j\}$ soit, compte-tenu de [25a]

$$\mathbf{D} = \mathrm{Diag}\{\phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})\}; \quad j = 1, 2, \ldots, J \qquad [29c]$$

En fait, Gilmour *et al* (1985) ne résolvent pas le système [28]. Ils proposent une approximation des éléments de $\mathbf{V}$ basée sur un développement limité de $\Phi_2(a, b; r)$ au voisinage de $r = 0$, qui s'écrit $\Phi_2(a, b; r) \doteq \Phi(a)\Phi(b) + r\phi(a)\phi(b)$ (Tallis, 1962) d'où, l'écriture suivante des éléments de $\mathbf{v}$:

$$\mathrm{Var}(p_j) \doteq \{\tilde{\pi}_j(1 - \tilde{\pi}_j) + (n_j - 1)t\phi^2(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})\}/n_j \qquad [30a]$$

$$\mathrm{Cov}(p_j, p_{j'}) \doteq (\mathbf{z}_j'\boldsymbol{\Sigma}_u\mathbf{z}_{j'})\phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})\phi(\mathbf{x}_{j'}'\tilde{\boldsymbol{\beta}})/(1 + \sigma^2) \qquad [30b]$$

Avec ces approximations [30a,b], la matrice $\mathbf{V}$ se met alors sous la forme

$$\mathbf{V} = \mathbf{V_0} + \mathbf{DZGZ'D} \qquad [31]$$

avec

$$\mathbf{V}_0 = \text{Diag}\left\{[\tilde{\pi}_j(1-\tilde{\pi}_j) - t\phi^2(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})]/n_j\right\} \quad [32a]$$

$$\mathbf{G} = \boldsymbol{\Sigma}_u/(1+\sigma^2) \quad [32b]$$

On retrouve aussi [32b] en faisant sur $\mathbf{u}$ le même changement de variable $\tilde{\mathbf{u}} = \mathbf{u}/(1+\sigma^2)^{1/2}$ que celui fait sur $\boldsymbol{\beta}$ en [24]. L'inverse de $\mathbf{W}$ en [28] s'écrit alors :

$$\mathbf{W}^{-1} = \mathbf{D}^{-1}\mathbf{V}_0\mathbf{D}^{-1} + \mathbf{ZGZ}' \quad [33a]$$

ou encore, en posant $\mathbf{R} = \mathbf{D}^{-1}\mathbf{V}_0\mathbf{D}^{-1}$

$$\mathbf{W}^{-1} = \mathbf{R} + \mathbf{ZGZ}' \quad [33b]$$

On reconnaît en [33b] la forme classique d'une matrice de covariance d'observations de modèle linéaire qui permet à Gilmour *et al* (1985) de résoudre de façon approchée le système [28] en $\tilde{\boldsymbol{\beta}}$ à partir des équations du modèle mixte d'Henderson, ici :

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1^{[r]}}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1^{[r]}}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1^{[r]}}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1^{[r]}}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\tilde{\boldsymbol{\beta}}}^{[r+1]} \\ \hat{\tilde{\mathbf{u}}}^{[r+1]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1^{[r]}}\boldsymbol{\zeta}^{[r]} \\ \mathbf{Z}'\mathbf{R}^{-1^{[r]}}\boldsymbol{\zeta}^{[r]} \end{bmatrix} \quad [34a]$$

ou, en posant $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}'\tilde{\mathbf{u}}')'$; $\mathbf{T} = (\mathbf{X}, \mathbf{Z})$

$$(\mathbf{T}'\mathbf{R}^{-1^{[r]}}\mathbf{T} + \boldsymbol{\Sigma}^-)\hat{\tilde{\boldsymbol{\theta}}}^{[r+1]} = \mathbf{T}'\mathbf{R}^{-1}\boldsymbol{\zeta}^{[r]} \quad [34b]$$

où $\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}$ est un cas particulier de [15a].

En [34ab], la variable de travail $\boldsymbol{\zeta}$ définie en [29b] peut aussi se mettre sous la forme :

$$\boldsymbol{\zeta} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{R}\tilde{\boldsymbol{\nu}} \quad [35a]$$

où

$$\tilde{\boldsymbol{\nu}}_{[J\times 1]} = \left\{ \phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}}) \, \frac{y_{j+} - n_j\Phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})}{\Phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})[1-\Phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})] - t\phi^2(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})} \right\} \quad [35b]$$

$$\mathbf{R}_{[j\times j]} = \text{Diag}\left\{ \frac{\Phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})[1-\Phi(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})] - t\phi^2(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})}{n_j\phi^2(\mathbf{x}_j'\tilde{\boldsymbol{\beta}})} \right\} \quad [35c]$$

Prédiction des variables aléatoires

Dans la procédure développée par Gilmour (1983) et Gilmour *et al* (1985), les variables aléatoires sont prédites à partir des solutions $\hat{\tilde{\mathbf{u}}}$ du système itératif [34a].

Il est difficile de trouver une justification à cette méthode de prédiction des **u** (Knuimann et Laird, 1990) hormis celle de calquer la procédure du système des équations du modèle mixte d'Henderson.

Estimation des paramètres de dispersion

Pour une structure comportant un vecteur **u** formé de K sous-vecteurs $\tilde{\mathbf{u}}_k$ indépendants tels que $\tilde{\mathbf{u}}_k \sim \mathrm{N}[\mathbf{0}, \mathbf{A}_k \sigeq{u_k}^2]$ avec, compte-tenu du changement d'échelle en [24] et [34a] $\sigeq{u_k}^2 = \sigma^2_{u_k}/(1+\sigma^2)$. Gilmour *et al* (1985) arguent de l'utilisation de $\mathrm{E}(\tilde{\mathbf{u}}'_k \mathbf{A}_k^{-1} \tilde{\mathbf{u}}_k)$ pour justifier un estimateur de $\sigeq{u_k}^2$ obtenu par une formule itérative de type EM, similaire à [22a],

$$\sigeq{u_k}^{[r+1]} = [\hat{\tilde{\mathbf{u}}}'_k \mathbf{A}_k^{-1} \hat{\tilde{\mathbf{u}}}_k + \mathrm{tr}\,(\mathbf{A}_k^{-1} \tilde{\mathbf{C}}_{uu,kk})]^{[r]}/q_k \qquad [36]$$

où $\hat{\tilde{\mathbf{u}}}_k$ et $\tilde{\mathbf{C}}_{uu,kk}$ sont respectivement la solution en $\tilde{\mathbf{u}}_k$ et le bloc relatif à ce même vecteur dans l'inverse de la matrice des coefficients du système [34a].

Une autre approche préconisée notamment par Knuimann et Laird (1990) réside dans l'estimation des composantes de la variance par le maximum de vraisemblance. La vraisemblance $\mathrm{L}(\tilde{\boldsymbol{\beta}}, \sigeq{u_1}^2, \sigeq{u_2}^2, \ldots, \sigeq{u_k}^2, \ldots, \sigeq{u_K}^2; \mathbf{p})$ nécessite l'intégration des $\tilde{\mathbf{u}}$, soit :

$$\mathrm{L}(\tilde{\boldsymbol{\beta}}, \sigeq{u_1}^2, \sigeq{u_2}^2, \ldots, \sigeq{u_k}^2, \ldots, \sigeq{u_K}^2; \mathbf{p}) = \ln \int \mathrm{p}(\mathbf{p}|\, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}) \mathrm{p}(\tilde{\mathbf{u}}|\, \boldsymbol{\gamma}_{\tilde{u}}) \mathrm{d}\tilde{\mathbf{u}} \qquad [37]$$

Thompson (1990) préconise une approximation de cette intégrale par une quadrature de Gauss. Dans un cadre similaire, mais avec des variables discrètes de Poisson, Im et Foulley (1990) proposent une méthode approchée de maximisation de [37] qui évite l'intégration explicite des $\tilde{\mathbf{u}}$.

MÉTHODE DU MODE CONJOINT A POSTERIORI -MAP (GF-HM)

Cette méthode a été développée à l'origine par Gianola et Foulley (1983). Des résultats identiques ont été obtenus simultanément par Harville et Mee (1984), ces derniers utilisant toutefois un raisonnement classique de modèle mixte. Elle sera désignée en abrégé par les initiales «GF-HM». Des procédures similaires ont été développées par Stiratelli *et al* (1984) ainsi que par Zellner et Rossi (1984) avec pour ce dernier une fonction de lien logistique.

Le modèle est le même au départ que celui présenté en [1b] et en [23] à la différence près que d'un point de vue bayésien, une distinction entre effets fixés et aléatoires n'a pas lieu d'être. On écrit donc :

$$\eta_j = \Phi^{-1}(\pi_j) = \mathbf{t}'_j \boldsymbol{\theta} \qquad [38]$$

L'intérêt des techniques bayésiennes en sélection animale a été mis en avant au cours des deux dernières décennies par plusieurs auteurs dont Rönningen (1971), Dempfle (1977), Lefort (1980) et Gianola et Fernando (1986).

Estimation des paramètres de position

L'inférence sur $\boldsymbol{\theta}$ en statistique bayésienne passe par l'obtention de la distribution *a posteriori*, qui s'écrit compte-tenu du théorème de Bayes :

$$p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\Sigma}) \quad [39]$$

L'expression [39] est le produit de la distribution *a priori* de $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\Sigma})$ sachant les hyperparamètres $(\boldsymbol{\alpha},\boldsymbol{\Sigma})$ par la distribution conditionnelle des données (ici $\mathbf{y} = \{y_{j+}\}$) sachant $\boldsymbol{\theta}$ ou vraisemblance en $\boldsymbol{\theta}$. Sachant $\boldsymbol{\theta}$, les données y_{j+} sont indépendantes entre elles et ont une distribution binomiale $B(n_j, \pi_j)$ de paramètres n_j (effectif de la classe j) et π_j. La vraisemblance s'écrit donc :

$$p(\mathbf{y}|\boldsymbol{\theta}) \propto \prod_{j=1}^{J} [\Phi(\mathbf{t}_j'\boldsymbol{\theta})]^{y_{j+}} [1 - \Phi(\mathbf{t}_j'\boldsymbol{\theta})]^{n_j - y_{j+}} \quad [40]$$

Le choix du modèle à seuils de Wright conduit à supposer que les distributions marginale et conditionnelle sachant $\boldsymbol{\theta}$ $[x_{jr}|\boldsymbol{\theta} \sim N(\mathbf{t}_j'\boldsymbol{\theta}; 1)]$ de la variable phénotypique x_{jr} associée à la r^e observation de la population j sont normales sur une échelle sous-jacente. Cette hypothèse implique donc le choix d'une distribution *a priori* de $\boldsymbol{\theta}$ qui soit aussi normale, d'où

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\Sigma}) \propto \exp[-(\boldsymbol{\theta} - \boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\alpha})/2] \quad [41]$$

et

$$p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \propto \prod_{j=1}^{J} [\Phi(\mathbf{t}_j'\boldsymbol{\theta})]^{y_{j+}} [1 - \Phi(\mathbf{t}_j'\boldsymbol{\theta})]^{n_j - y_{j+}} \exp[-(\boldsymbol{\theta} - \boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\alpha})/2] \quad [42]$$

Gianola et Foulley (1983) ont proposé comme estimateur ponctuel de $\boldsymbol{\theta}$ le mode *a posteriori* (MAP) de [42]. Celui-ci peut s'obtenir par résolution d'un système itératif du second ordre tel que :

$$(\mathbf{T}'\mathbf{S}^{-1[r]}\mathbf{T} + \boldsymbol{\Sigma}^{-1})\boldsymbol{\theta}^{*[r+1]} = \mathbf{T}'\mathbf{S}^{-1[r]}\boldsymbol{\Psi}^{[r]} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \quad [43]$$

où

$$\boldsymbol{\Psi} = \mathbf{T}\boldsymbol{\theta} + \mathbf{S}\mathbf{v} \quad [44]$$

avec

$$\mathbf{v}_{[j\times 1]} = \left[\phi(\mathbf{t}_j'\boldsymbol{\theta}) \, \frac{y_{j+} - n_j\Phi(\mathbf{t}_j'\boldsymbol{\theta})}{\Phi(\mathbf{t}_j'\boldsymbol{\theta})[1 - \Phi(\mathbf{t}_j'\boldsymbol{\theta})]} \right] \quad [45a]$$

et $\mathbf{S}_{[J\times J]}$ matrice diagonale définie par exemple avec la méthode des scores par :

$$\mathbf{S} = \text{Diag}\left\{ s_j = \frac{\Phi(\mathbf{t}_j'\boldsymbol{\theta})[1 - \Phi(\mathbf{t}_j'\boldsymbol{\theta})]}{n_j\phi^2(\mathbf{t}_j'\boldsymbol{\theta})} \right\} \quad [45b]$$

On sera amené à appréhender la distribution *a posteriori* par sa forme asymptotique normale (Berger, 1985; Fouley, 1987) :

$$\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\alpha},\boldsymbol{\Sigma} \sim N[\boldsymbol{\theta}^*(\boldsymbol{\alpha},\boldsymbol{\Sigma}), \overline{\mathbf{C}}(\boldsymbol{\alpha},\boldsymbol{\Sigma})] \qquad [46a]$$

où $\boldsymbol{\theta}^*$ est la solution MAP de [43] et

$$\overline{\mathbf{C}}(\boldsymbol{\alpha},\boldsymbol{\Sigma}) = [\mathbf{T}'\mathbf{S}^{-1}(\boldsymbol{\alpha},\boldsymbol{\Sigma})\mathbf{T} + \boldsymbol{\Sigma}^{-1}]^{-1} \qquad [46b]$$

Si l'on s'intéresse à une structure de modèle mixte avec des effets fixes $\boldsymbol{\beta}$ et aléatoires $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$, celle-ci pourra être traitée comme un cas bayésien dégénéré en faisant, dans les formules [43], [44] et [45ab] :

$$\boldsymbol{\theta}' = (\boldsymbol{\beta}', \mathbf{u}'); \quad \boldsymbol{\Sigma}^{-1} \to \boldsymbol{\Sigma}^- = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_u^{-1} \end{bmatrix}; \quad \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \to \mathbf{0} \qquad [47]$$

Un autre algorithme du MAP qui mérite attention, a été proposé par Mee (1982) et Zhao (1987) en utilisant le raisonnement de l'algorithme EM (Dempster *et al*, 1977).

Supposons qu'on puisse observer les variables continues sous-jacentes $\mathbf{x} = \{x_j\}$ (j indice d'une cellule élémentaire d'effectif unité). Dans le modèle à seuils, on a les distributions suivantes :

$$\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\Gamma} \sim \mathrm{N}(\mathbf{T}\boldsymbol{\theta},\boldsymbol{\Gamma}) \qquad [48a]$$

$$\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\alpha},\boldsymbol{\Sigma}) \qquad [48b]$$

où $\boldsymbol{\Gamma}$ représente la variance phénotypique résiduelle sur la sous-jacente sachant le vecteur de paramètres $\boldsymbol{\theta}$. Pour un modèle binaire univariate, $\boldsymbol{\Gamma}$ est la matrice $\mathbf{I}\sigma_e^2$. L'espérance de la distribution conditionnelle de $\mathbf{x}$, soit $\mathbf{T}\boldsymbol{\theta}$ a la même signification qu'en [38]; de même la distribution en [48b] est identique à celle en [41], d'où

$$\mathrm{L}(\boldsymbol{\theta};\mathbf{x},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \ln \mathrm{p}(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\Gamma}) + \ln \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\Sigma}) \qquad [49]$$

Dans l'EM généralisé applicable au mode *a posteriori* (Dempster *et al*, 1977), on considère la fonction :

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \mathrm{E}[\mathrm{L}(\boldsymbol{\theta};\mathbf{x},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\Sigma})|\mathbf{y},\widehat{\boldsymbol{\theta}}] \qquad [50a]$$

où $\mathbf{y} = \{y_j\}$ est le vecteur des données binaires observées.

En [50a], tout se passe comme si l'ensemble «complet» des données (selon la terminologie de Dempster *et al*, 1977) se restreignait aux seules variables sous-jacentes $\mathbf{x}$ puisque l'information en $\mathbf{x}$ contient celle sur les données binaires $\mathbf{y}$, autrement dit :

$$\mathrm{L}(\boldsymbol{\theta};\mathbf{x},\mathbf{y},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \mathrm{L}(\boldsymbol{\theta};\mathbf{x},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\Sigma})$$

Compte-tenu des distributions en [48ab], l'expression en [50a] s'écrit :

$$\mathrm{Q}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = (-1/2)[\mathrm{E}(\mathbf{x}|\mathbf{y},\widehat{\boldsymbol{\theta}}) - \mathbf{T}\boldsymbol{\theta}]'\boldsymbol{\Gamma}^{-1}[\mathrm{E}(\mathbf{x}|\mathbf{y},\widehat{\boldsymbol{\theta}}) - \mathbf{T}\boldsymbol{\theta}] \qquad [50b]$$
$$(-1/2)(\boldsymbol{\theta}-\boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\alpha}) + \mathit{Const}$$

En fait, dans la phase E, on remplace $E(\mathbf{x}|\mathbf{y},\widehat{\boldsymbol{\theta}})$ par :

$$\mathbf{x}^{*[r]} = E(\mathbf{x}|\mathbf{y},\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{[r]}) \tag{51}$$

puis, dans la phase M, on maximise [50b] par rapport à $\boldsymbol{\theta}$ soit :

$$\mathbf{T}'\boldsymbol{\Gamma}^{-1}(\mathbf{x}^{*[r]} - \mathbf{T}\boldsymbol{\theta}^{*[r+1]}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}^{*[r+1]} - \boldsymbol{\alpha}) = \mathbf{0}$$

d'où le système itétatif

$$(\mathbf{T}'\boldsymbol{\Gamma}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-1})\boldsymbol{\theta}^{*[r+1]} = \mathbf{T}'\boldsymbol{\Gamma}^{-1}\mathbf{x}^{*[r]} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \tag{52}$$

L'écriture en [52] offre une bonne illustration de la façon dont procède l'algorithme EM. Si $\mathbf{x}$ était observable, le système en $\boldsymbol{\theta}$ serait le système linéaire classique des équations généralisées du modèle mixte $(\mathbf{T}'\boldsymbol{\Gamma}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-1})\widehat{\boldsymbol{\theta}} = \mathbf{T}'\boldsymbol{\Gamma}^{-1}\mathbf{x} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}$. Comme $\mathbf{x}$ n'est pas observable, tout se passe en [52] comme si on le remplaçait dans ce système par son espérance conditionnelle $\mathbf{x}^*$ sachant les données discrètes observées $\mathbf{y}$ et la valeur du paramètre inconnu $\boldsymbol{\theta}$ à l'itération précédente.

Dans le cas d'un modèle binaire univariate, l'égalité $\boldsymbol{\Gamma} = \mathbf{I}\sigma_e^2$ entraîne une importante simplification en [52]. Comme pour $j \neq j'$, $p(x_j|y_{j'},\boldsymbol{\theta}) = p(x_j|\boldsymbol{\theta})$, x_j et $y_{j'}$ étant conditionnellement indépendants sachant $\boldsymbol{\theta}$, les éléments x_j^* de $\mathbf{x}^*$ correspondent à $x_j^* = E(x_j|y_j,\boldsymbol{\theta})$ ce qui s'écrit aussi (Im et Gianola, 1988) :

$$x_j^* = E(x_j|y_j = 0,\boldsymbol{\theta}) + y_j[E(x_j|y_j = 1,\boldsymbol{\theta}) - E(x_j|y_j = 0,\boldsymbol{\theta})] \tag{53}$$

et, on montre aisément que :

$$E(x_j|y_j = 1,\boldsymbol{\theta}) = \mathbf{t}_j'\boldsymbol{\theta} + [\phi(\mathbf{t}_j'\boldsymbol{\theta})/\Phi(\mathbf{t}_j'\boldsymbol{\theta})] \tag{54a}$$

$$E(x_j|y_j = 0,\boldsymbol{\theta}) = \mathbf{t}_j'\boldsymbol{\theta} - \{\phi(\mathbf{t}_j'\boldsymbol{\theta})/[1 - \Phi(\mathbf{t}_j'\boldsymbol{\theta})]\} \tag{54b}$$

Les expressions en [54ab] correspondent aux scores normaux tels que présentés par Gianola et Foulley (1983). Compte-tenu de l'expression de ν_j en [45a] appliquée au cas d'une cellule élémentaire $n_j = 1$, il apparaît que x_j^* peut se mettre aussi sous la forme $x_j^* = \mathbf{t}_j'\boldsymbol{\theta} + \nu_j$. Si de plus, comme [47], on pose $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \mathbf{u}')$; $\boldsymbol{\Sigma}^{-1} \to \boldsymbol{\Sigma}^- = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_u^{-1} \end{bmatrix}$; $\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \to \mathbf{0}$ le système [52] se réduit alors à

$$(\mathbf{T}'\mathbf{T} + \boldsymbol{\Sigma}^-)\boldsymbol{\theta}^{[r+1]} = \mathbf{T}'\mathbf{x}^{*[r]} \tag{55a}$$

avec

$$\mathbf{x}^* = \mathbf{T}\boldsymbol{\theta} + \mathbf{v} \tag{55b}$$

La matrice des coefficients est identique à celle des équations du modèle mixte d'Henderson ; les itérations ne modifient que le second membre. Cet algorithme rappelle la suggestion faite par Aisbett (1983) visant à simplifier les calculs. Toutefois, s'agissant d'une procédure EM, cet algorithme est un algorithme de

premier ordre comme le montre bien l'écriture de [55a] pour $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, valeur de convergence

$$(\mathbf{T}'\mathbf{T} + \boldsymbol{\Sigma}^-)\boldsymbol{\theta}^* = \mathbf{T}'\mathbf{T}\boldsymbol{\theta}^* + \mathbf{T}'\mathbf{v}$$

soit

$$\mathbf{T}'\mathbf{v} - \boldsymbol{\Sigma}^-\boldsymbol{\theta}^* = \mathbf{0} \qquad [56]$$

Cette dernière expression [56] représente la condition d'annulation de la dérivée première du logarithme de la densité *a posteriori* définie en [42] pour un *a priori* uniforme sur la composante $\boldsymbol{\beta}$ (*cf* Gianola et Foulley, 1983). S'agissant d'un algorithme de premier ordre, il faut donc s'attendre à ce que cet algorithme converge plus lentement que l'algorithme des scores donné en [43].

Estimation des paramètres de dispersion

Partant d'une distribution normale *a priori* des $\mathbf{u}$ telle que $\mathbf{u} \sim \mathrm{N}[\mathbf{0}, \boldsymbol{\Sigma}_u(\boldsymbol{\gamma}_u)]$, les résultats généraux présentés en [16], [21 ab] et [22 ab] s'appliquent ici. La difficulté réside alors dans l'obtention de l'espérance et de la variance de la distribution *a posteriori* de $\mathbf{u}$. La plupart des auteurs (Harville et Mee, 1984; Stiratelli *et al*, 1984; Foulley *et al*, 1987a) ont proposé de les approcher par les paramètres homologues de la distribution asymptotique définis en [46 ab], d'où pour la même structure qu'en [17], l'algorithme :

$$\sigma^{2[r+1]}_{u_k} = \left\{ [\widehat{\mathbf{u}}^{*'}_k \mathbf{A}^{-1} \widehat{\mathbf{u}}^*_k]^{[r]} + \mathrm{tr}\,[\mathbf{A}^{-1}\overline{\mathbf{C}}_{uu,kk}(\boldsymbol{\gamma}^{[r]}_u)] \right\} / q_k \qquad [57]$$

Sur la base de cette même approximation asymptotique, on peut développer un algorithme de second ordre (*cf* Annexe A) formellement similaire à celui décrit dans la première partie pour la méthode GSK-FI. Par ailleurs, on peut avoir recours à une méthode approchée (Van Raden et Yung, 1988) adaptée aux caractères à seuils par Manfredi (1990) lorsque la taille des fichiers rend les méthodes précédentes difficilement applicables.

DISCUSSION

Les 3 méthodes statistiques ainsi que les algorithmes correspondants décrits dans ce chapitre soit en abrégé (GSK-FI; GAR; GF-HM) permettent de traiter le modèle à seuil de Wright à des fins d'évaluation génétique. L'objet de cette discussion est d'effectuer une appréciation critique comparative de ces procédures.

Comparaison des algorithmes

Foulley (1987) a souligné les similitudes et les différences existant entre l'algorithme de calcul des $\boldsymbol{\beta}$ et $\mathbf{u}$ décrit en [43] pour la méthode GF-HM et celui proposé par GAR. En particulier, la matrice des pondérations $\mathbf{S}$ et la variable de travail $\boldsymbol{\Psi}$ de GF-HM bien que de formes très similaires à leurs homologues $\mathbf{R}$ (*cf* [33 ab]; [35c]) et $\boldsymbol{\zeta}$ (*cf* [35 abc]) en diffèrent parce qu'elles font intervenir explicitement $\boldsymbol{\beta}$ et $\mathbf{u}$ et non $\boldsymbol{\beta}$ et Var($\mathbf{u}$) comme chez GAR.

Quant aux méthodes GSK-FI et GF-HM, leurs matrices de pondération $\mathbf{Q}^*$ (*cf* [8]) et $\mathbf{S}$ (*cf* [45b]) ont exactement la même forme : elles ne diffèrent que par l'argument utilisé dans la fonction : p_j pour GSK-FI et π_j pour GF-HM. Quant aux variables de travail du second membre $\mathbf{g}$ et $\boldsymbol{\Psi}$ respectivement pour GSK-FI et GF-HM, la seconde s'interprète en fait comme un développement limité au 1er ordre de la première au voisinage de $p_j = \pi_j$. En effet :

$$\begin{aligned} g(p_j) &\doteq g(\pi_j) + \left.\frac{\partial g(p)}{\partial p}\right|_{p=\pi_j} (p_j - \pi_j) \\ &\doteq \mathbf{t}_j'\boldsymbol{\theta} + (p_j - \pi_j)/\phi(\mathbf{t}_j'\boldsymbol{\theta}) \\ &\doteq \mathbf{t}_j'\boldsymbol{\theta} + s_j\nu_j \end{aligned} \qquad [58]$$

Enfin, il faut noter que l'algorithme de l'estimateur MAP est itératif contrairement à celui de GSK-FI.

Estimation de $\boldsymbol{\beta}$ *et* $\mathbf{u}$ *avec des variances inconnues*

Les 3 méthodes d'estimation de $\boldsymbol{\beta}$ et $\mathbf{u}$ développées ici supposent la connaissance de la matrice de covariance $\boldsymbol{\Sigma}_u$ de $\mathbf{u}$ connue donc celle du vecteur $\boldsymbol{\gamma}_u$ des paramètres dont elle dépend. Quand $\boldsymbol{\gamma}_u$ n'est pas connu on peut adopter, à l'instar de Foulley *et al* (1987a, 1989b) et Gianola *et al* (1986), une approche bayésienne empirique fondée sur la distribution *a posteriori* conditionnellement à $\boldsymbol{\gamma}_u$ égalé au mode de la distribution marginale *a posteriori* de $\boldsymbol{\theta}$ qui se réduit au maximum de vraisemblance marginale avec une distribution *a priori* uniforme sur $\boldsymbol{\gamma}_u$. C'est précisément l'estimateur de $\boldsymbol{\gamma}_u$ développé en [21 ab] [22 ab] et [57]. L'extension bayésienne empirique des méthodes GSK-FI et GF-HM est donc immédiate. Au contraire, si l'on peut, dans l'algorithme [34a], remplacer les composantes de $\boldsymbol{\gamma}_u$ par leurs estimations telles que proposées par GAR en [36], ce procédé ne résulte pas d'une justification théorique claire.

Choix de l'estimateur ponctuel $\boldsymbol{\theta}$

Du point de vue de l'efficacité de la sélection en une génération, le critère qui maximise l'espérance de la valeur génétique des individus sélectionnées à nombres de candidats et retenus fixes, est l'espérance de la distribution *a posteriori* des valeurs génétiques (Goffinet et Elsen, 1984; Fernando et Gianola, 1986).

L'estimateur $\widehat{\boldsymbol{\theta}}$ défini en [13] pour la méthode GSK-FI est bien une espérance mais la distribution considérée en [12] est une approximation fondée sur un résultat asymptotique de la loi conditionnelle du normit des observations. L'estimateur $\boldsymbol{\theta}^*$ proposé en [43] pour la méthode GF-HM est le mode de la distribution *a posteriori*, qui se distinguera d'autant plus de l'espérance que la distribution sera plus asymétrique (cas d'un faible effectif par niveau de facteur). L'estimateur modal n'en reste pas moins potentiellement intéressant eu égard à sa justification en théorie de la décision et il a l'avantage sur l'espérance d'être moins sensible aux «queues» des distributions. Dans le cas d'une distinction nette entre paramètres

d'intérêt et paramètres de nuisance, on peut l'améliorer en considérant le mode de la distribution marginale après intégration des paramètres de nuisance tels que les effets parasites de milieu en génétique animale. Cette opération n'est malheureusement pas possible analytiquement dans le cas des données binaires analysées selon GF-HM. Cette idée de marginalisation est en fait sous-jacente à la procédure GAR en vue de l'estimation des effets fixes $\boldsymbol{\beta}$.

Comparaison des estimateurs

À notre connaissance, l'estimateur de $\boldsymbol{\theta}$ en modèle mixte selon GSK-FI n'a fait l'objet d'aucune application excepté celle mentionnée par Foulley et Im (1989) à propos de la précision d'une évaluation génétique sur descendance.

Par construction, on s'attend à de meilleures propriétés de l'estimateur de $\boldsymbol{\beta}$ obtenu par la méthode GAR que par celle de GF-HM. Une comparaison rigoureuse de ces méthodes a été effectuée par Höschele et Gianola (1989) grâce aux techniques de simulation. Celle-ci concerne une variable binaire suivant un modèle de susceptibilité sous-jacente comportant les facteurs «troupeau × année × saison» et «groupe de pères» analysés comme fixes et le facteur «pères» intra groupe considéré comme aléatoire. Le dispositif comprenait les descendants de 50 pères originaires de 4 groupes. En une première étape, 2000 descendants étaient ainsi simulés répartis en 135 cellules «troupeau × année × saison». En un deuxième stade, 1000 descendants étaient répartis entre 10 pères ayant obtenu les meilleurs évaluations génétiques au 1er stade selon l'une ou l'autre méthode.

Les méthodes ont été comparées pour l'estimation de l'héritabilité et des effets «groupe» sur la base de statistiques d'échantillonnage (biais, variance, erreur quadratique moyenne) empiriques observées sur 45 répétitions. Les estimateurs de GF-HM surpassent ceux de GAR en termes de biais et d'erreur quadratique aussi bien pour l'héritabilité que pour les effets «groupe» et cela sur les échantillons du stade 1 comme sur ceux du stade 2 avec sélection. Quand on considère la valeur génétique transmise vraie de taureaux sélectionnés à des pressions de sélection de 10, 20 et 40% à chaque stade, la différence est négligeable entre les 2 méthodes. Cette étude conforte donc le bien fondé de l'approche GF-HM en matière d'évaluation génétique, y compris vis-à-vis de l'estimation génétique de certaines composantes fixes pour lesquelles on aurait pu espérer une supériorité de la méthode GAR. Cette étude met par ailleurs clairement en évidence un biais non négligeable dans l'estimation de l'héritabilité (moyenne de 0,30 et 0,34 avec les méthodes GF-HM et GAR respectivement pour une valeur vraie de 0,25), ce qui, dans le cas de la méthode bayésienne, pose la question de la validité de l'approximation de l'espérance par le mode.

Estimation des probabilités de réponse

Un des problèmes délicats avec les modèles opérant des transformations de données ou de paramètres est le retour à une estimation des paramètres d'origine. Avec le modèle linéaire généralisé et les méthodes basées sur la vraisemblance (et la quasi-vraisemblance), la propriété d'invariance permet aisément d'obtenir l'estimation

homologue de la probabilité de réponse π_j, dans la cellule j par la fonction inverse de la fonction de lien soit, avec les notations de [25 ab] $\widehat{\widehat{\pi}}_j = \Phi(\mathbf{x}_j\widehat{\widehat{\boldsymbol{\beta}}})$.

En fait, cette propriété ne répond pas au besoin du sélectionneur qui est intéressé par la probabilité de réponse dans une sous-population définie par combinaison, non seulement des niveaux des facteurs fixes, mais aussi de niveaux de facteurs considérés comme aléatoires. Par exemple, dans une étude sur les difficultés de vêlage, on se demandera quelle est la probabilité d'un vêlage dystocique pour une fille née du père X (effet aléatoire) mettant bas dans la saison Y (effet fixe). La méthode GAR ne nous permet pas de répondre simplement à ce problème. Il faudrait la généraliser en faisant appel, par exemple, au concept de vraisemblance prédictive (Bjornstad, 1990).

Au contraire, le concept de densité prédictive (Zellner, 1971) qui apparaît tout naturellement en statistique bayésienne, permet de bien répondre au problème concret que se pose le sélectionneur. Ainsi, Foulley *et al* (1989b) considèrent la probabilité de réponse pour une future observation f de la sous-population j sachant les données observées (vecteur y_0). Celle-ci s'écrit :

$$\begin{aligned}\Pr(y_{if}=1|\mathbf{y}_0) &= \int_{\infty}^{+\infty} \Pr(y_{if}=1|\eta_j,\boldsymbol{\alpha},\boldsymbol{\Sigma})p(\eta_j|\mathbf{y}_0,\boldsymbol{\alpha},\boldsymbol{\Sigma})\mathrm{d}\eta_j \\ &= \int_{-\infty}^{+\infty} \Phi(\eta_j)p(\eta_j|\mathbf{y}_0,\boldsymbol{\alpha},\boldsymbol{\Sigma})\mathrm{d}\eta_j \qquad [59]\end{aligned}$$

Cette probabilité apparaît comme une moyenne pondérée de $\Phi(\eta_j)$ par la densité *a posteriori* de η_j, donc par un indicateur du degré de connaissance qu'on a sur ce paramètre. Cette densité *a posteriori* est décrite par des approximations normales données en [12] et [46 ab] pour les méthodes GSK-FI et GF-HM respectivement. Nous écrirons donc :

$$\eta_j|\mathbf{y}_0,\boldsymbol{\alpha},\boldsymbol{\Sigma} \dot{\sim} \mathrm{N}(\widehat{\eta}_j, \sigma^2_{\eta_j|\mathbf{y}_0}) \qquad [60]$$

Partant de [60], l'intégrale de [59] se calcule aisément à la formule de Curnow (1984) ; on obtient alors :

$$\Pr(y_{jf}=1|\mathbf{y}_0) = \Phi[\widehat{\eta}_j/(1+\sigma^2_{\eta_j|\mathbf{y}_0})^{1/2}] \qquad [61]$$

La formule [61] indique bien l'incidence de l'incertitude sur l'estimation de η_j qui tend, par réduction de la valeur absolue de l'argument de l'intégrale [59] à «régresser» la prédiction de la probabilité de réponse sur 1/2. Il convient toutefois de noter la difficulté importante qu'il y aura à calculer $\sigma^2_{\eta_j|\mathbf{y}_0}$.

Cette approche permet d'aborder ensuite rigoureusement le problème de l'évaluation génétique des pères sur l'échelle des probabilités. Soit η_{ik} le paramètre de l'échelle sous-jacente correspondant à la cellule «père i par combinaison élémentaire k» des autres facteurs. Le père i sera évalué sur l'échelle des probabilités par le paramètre

$$P_i = \sum_{k\in E_i} \delta_{ik}\Phi(\eta_{ik}) \qquad [62]$$

où E_i représente l'ensemble des indices des cellules où figure potentiellement un descendant de i et δ_{ik} est la probabilité (fixée par le sélectionneur) que le père i ait un descendant dans la cellule indicée par ik.

Généralement, pour une comparaison équitable, le sélectionneur se fixera des probabilités δ_{ik} identiques d'un père à l'autre. La formule [62] montre clairement qu'on ne peut pas, en modèle non linéaire, évaluer des pères sans préciser les variantes dans lesquelles on veut les caractériser.

Extension à d'autres situations

L'approche GSK-FI peut se généraliser au cas de plusieurs catégories par une approche de type BLUP multidimensionnel (Gianola, 1980b). Cette approche ne prend pas en compte toutefois l'ordre des catégories de réponse; elle suppose également de se restreindre à une fonction de lien logistique (*cf* multinomial logit model; Maddala, 1983), soit pour la catégorie c de la sous-population j

$$\pi_{jc}/\pi_{jC} = \exp(\mathbf{t}'_j\boldsymbol{\theta}_c)$$

pour $c = 1, 2, \ldots, C-1$, la dernière catégorie n'étant pas explicitée.

Les méthodes GAR et GF-HM, au contraire, s'appliquent très bien à des polytomies ordonnées (*cf* Gilmour *et al*, 1985; Gianola et Foulley, 1983). Pour les C catégories ordonnées délimitées par les seuils $(\tau_1, \tau_2, \ldots, \tau_c, \ldots, \tau_{C-1})$, on peut écrire sachant le paramètre $\boldsymbol{\theta}$ et avec la convention $(\tau_0 = -\infty\,;\ \tau_C = +\infty)$

$$\begin{cases} \pi_{jc} = \Phi(\tau_c - \eta_j) - \Phi(\tau_{c-1} - \eta_j) \\ \eta_j = \mathbf{t}'_j\boldsymbol{\theta} \end{cases}$$

Maintes extensions ont été effectuées dans le cadre de l'approche bayésienne, notamment dans les situations suivantes : réponses binaires multiples complètes (Foulley et Gianola, 1984; Höschele *et al*, 1986; Foulley *et al*, 1987a) ou avec information manquante (Foulley et Gianola, 1986); mélange de variables binaires et continues (Foulley *et al*, 1983; Simianer et Schaeffer, 1989; Janss, 1990); mélanges de variables binaires et de Poisson (Foulley *et al*, 1987b). L'approche a été également étendue à des situations d'assignation incertaine des observations à certains facteurs de variation tels que le génotype majeur (Foulley et Elsen, 1988) ou la paternité (Foulley *et al*, 1987b; Foulley *et al*, 1990b).

REMERCIEMENTS

Les auteurs tiennent à remercier les responsables de l'UPRA Maine-Anjou pour le constant encouragement qu'ils ont accordé aux auteurs dans l'application expérimentale de certaines des méthodes décrites dans ce texte – notamment sur les difficultés de vêlage et la gémellité – ainsi que pour le soutien financier apporté à la réalisation des travaux du second auteur. Les auteurs adressent également leurs remerciements à V Ducrocq, M San Cristobal et à un lecteur anonyme de la revue qui, par leurs critiques et suggestions, ont permis d'améliorer la présentation du manuscrit.

RÉFÉRENCES

Aisbett CW (1983) *Maximum likelihood estimation with ordered categorical data and a threshold model.* Technical report, AGBU, University of New England, Armidale

Beitler P, Landis JR (1985) A mixed effects model for categorical data. *Biometrics* 41, 991-1000

Benzecri JP (1973) *L'analyse des données. I- La taxonomie II- L'analyse des correspondances.* Dunod, Paris

Berger JO (1985) *Statistical decision theory and bayesian analysis.* 2nd edn Springer-Verlag, New York

Bjornstad JF (1990) Predictive likelihood : a review. *Statist Sci* 5, 242-265

Bodin L, Elsen JM (1989) Variability of litter size of French sheep breeds following natural and induced ovulation. *Anim Prod* 48, 535-541

Bull JJ, Vogt RC, Bulmer MG (1982) Heritability of sex ratio in turtles with environmental sex determination. *Evolution* 36, 333-341

Bulmer MG, Bull JJ (182) Models of polygenic sex determination and sex ratio control. *Evolution* 36, 13-26

Brillinger DR (1985) *What do seismology and neurophysiology have in common ? – statistics.* Technical report 50, University of California, Berkeley

Curnow R (1984) Progeny testing for all-or-none traits when a multifactorial model applies. *Biometrics* 40, 375-382

Curnow R, Smith C (1975) Multifactorial models for familial diseases in man. *J R Statist Soc A* 138, 131-169

Dempfle L (1977) Relation entre le BLUP et estimateurs bayésiens. *Ann Génét Sél Anim* 9, 27-32

Dempster ER, Lerner IM (1950) Heritability of threshold characters. *Genetics*, 35, 212-236

Dempster A, Laird N, Rubin R (1977) Maximum likelihood estimation from incomplete data *via* the EM algorithm. *J R Statist Soc B* 39, 1-20

Djemali M, Berger P, Freeman A (1987) Ordered categorical sire evaluation for dystocia in Holstein. *J Dairy Sci* 70, 2374-2384

Ducrocq V (1990) Estimation of genetic parameters arising in nonlinear models. *In : 4th world congress on genetics applied to livestock production, Edinburgh, 23-27 july 1990*, vol 13 (Hill WG, Thompson R, Woolliams JA, eds) 419-428.

Elsen JM, Le Roy P (1990) Detection of major genes and determination of genotypes : application to discrete variables. *In : 4th world congress on genetics applied to livestock production, Edinburgh, 23-27 july 1990*, vol 15 (Hill WG, Thompson R, Woolliams JA, eds) 37-49

Falconer DS (1965) The inheritance of liability to certain diseases estimated from the incidence among relatives. *Ann Hum Genet* 29, 51-76

Fernando RL, Gianola D (1984) Optimal property of the conditional mean as a selection criterion. *Theor Appl Genet 72*, 822-825

Foulley JL (1987) Méthodes d'évaluation des reproducteurs pour des caractères discrets à déterminisme polygénique en sélection animale. Thèse d'Etat, Université de Paris-Sud-Orsay

Foulley JL, Gianola D (1984) Estimation of genetic merit from bivariate "all-or-none" responses. *Génét Sél Évol* 16, 285-306

Foulley JL, Gianola D (1986) Sire evaluation for multiple binary responses when information is missing on some traits. *J Dairy Sci* 69, 2681-2695

Foulley JL, Elsen JM (1988) Posterior probability of the sire's genotype at a major locus based on progeny test results for discrete characters. *Génét Sél Évol* 20, 227-238

Foulley JL, Im S (1989) Probability statements about the transmitting ability of progeny-tested sires for all-or-none trait with an application to twinning in cattle. *Génét Sél Évol 21*, 359-376

Foulley JL, Gianola D, Thompson R (1983) Prediction of genetic merit from data on categorical and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. *Génét Sél Évol* 15, 401-424

Foulley JL, Im S, Gianola D, Höschele I (1987a) Empirical Bayes estimation of parameters for n polygenic binary traits. *Génét Sél Évol* 19, 197-224

Foulley JL, Gianola D, Planchenault D (1987b) Sire evaluation with uncertain paternity. *Génét Sél Évol* 19, 83-102

Foulley JL, Gianola D, Im S (1989a) A simple algorithm for computing marginal maximum likelihood estimates of variance components and its relation to EM. *In : 47e session de l'ISI, Paris, 29 août-6 sept 1989*, Vol I, 337-338

Foulley JL, Gianola D, Im S, Misztal I (1989b) Une approche bayésienne de l'analyse génétique de caractères discrets. *In : Biométrie et données discrètes.* (Asselain B, Duby C, Masson JP, Tranchefort J, eds) Ensar, Rennes, Vol 7, 6-35

Foulley JL, Gianola D, Im S (1990a) Genetic evaluation polygenic traits in animal breeding. *In : Advances in statistical methods for genetic improvement of livestock* (Gianola D, Hammond K, eds) Springer-Verlag, Heidelberg, 361-396

Foulley JL, Thompson R, Gianola D (1990b) On sire evaluation with uncertain paternity. *Genet Sel Evol* 23, 373-376

Fraser FC (1980) The William Allan memorial award address : evolution of a palatable multifactorial model. *Amer J Hum Genet* 32, 796-813

Freycon V (1989) Estimation quadratique des paramètres d'un modèle à un facteur aléatoire sur une variable binaire. *In : Biométrie et données discrètes.* (Asselain B, Duby C, Masson JP, Tranchefort J, eds) Ensar, Rennes, Vol 7, 36-48

Gianola D (1980a) A method of sire evaluation for dichotomies. *J Anim Sci* 51, 1266-1271

Gianola D (1980b) Genetic evaluation of animals for traits with categorical responses. *J Anim Sci* 51, 1272-1276

Gianola D (1982) Theory and analysis of threshold characters. *J Anim Sci* 54, 1079-1096

Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical data with a threshold model. *Génét Sél Évol* 15, 201-224

Gianola D, Fernando R (1986) Bayesian methods in animal breeding theory. *J Anim Sci* 63, 217-244

Gianola D, Foulley JL, Fernando R (1986) Prediction of breeding values when variances are not known. *Génét Sél Évol* 18, 485-498

Gilmour A (1983) The estimation of genetic parameters from categorical data. PhD thesis, Massey University, Palmerston North, NZ

Gimour A, Anderson RD, Rae A (1985) The analysis of binomial data by a generalized linear mixed model. *Biometrika* 72, 593-599

Gilmour A, Anderson RD, Rae A (1987) Variance components on an underlying scale for ordered multiple threshold categorical data using a generalized linear model. *J Anim Breed Genet* 104, 149-155

Goffinet B, Elsen JM (1984) Critère optimal de sélection : quelques résultats généraux. *Génét Sél Évol* 16, 307-318

Grizzle A, Starmer C, Koch GG (1969) Analysis of categorical data by linear models. *Biometrics* 25, 489-504

Grosbas JM (1987) Les données manquantes. *In : Les sondages* (Droesbeke JJ, Fichet B, Tassi P, eds), Economica, Paris, 173-195

Haberman SJ (1982) Measures of association. *In : Encyclopedia of statistical sciences* (Kotz S, Johnson NL, eds) John Wiley and Sons, New York, Vol 1, 130-137

Hagger C, Hoffer A (1990) Genetic analysis of calving traits in the Swiss Black and White, Braunvieh and Simmental breeds by Reml and Mapp procedures. *Livest Prod Sci* 24, 93-107

Hamada M, Wu CFJ (1990) A critical look at accumulation analysis. *Technometrics* 32, 119-162

Hammerle A (1990) Latent variable models for categorical longitudinal data. *In : 15th international biometric conference, Budapest, july 2–6, 1990.* Invited papers, 227-233

Harville DA, Mee RW (1984) A mixed model procedure for analysing ordered categorical data. *Biometrics* 40, 393-408

Henderson CR (1973) Sire evaluation and genetic trend. *In : Proceedings of the animal breeding and genetic symposium in honor of Dr JL Lush.* American Society of Animal Science and American Dairy Science Association, Champaign, Illinois, 10-41

Henderson CR (1984) *Applications of linear models in animal breeding.* University of Guelph, Guelph

Hill WG, Smith C (1977) Estimating heritability of a dichotomous trait. *Biometrics* 33, 231-236

Höschele I, Foulley JL, Colleau JJ, Gianola D (1986) Genetic evaluation for multiple binary responses. *Génét Sél Évol* 18, 299-321

Höschele I, Gianola D, Foulley JL (1987) Estimation of variance components with quasi-continuous data using Bayesian methods. *J Anim Breed Genet* 104, 334-349

Höschele I, Gianola D (1989) Bayesian *versus* maximum quasi-likelihood methods for sire evaluation with categorical data. *J Dairy Sci* 72, 1569-1577

Im S (1982) Contribution à l'étude des tables de contingence à paramètres aléatoires : utilisation en biométrie. Thèse de 3^{e} cycle, Université Paul Sabatier, Toulouse

Im S, Gianola D (1988) Offspring parent regression for a binary trait. *Theor Appl Genet* 75, 720-722

Im S, Foulley JL (1990) Likelihood procedures for estimating fixed effects in a mixed model for Poisson variables. *In : 41st annual meeting of the EAAP, Toulouse, july 9–12, 1990*, Vol 1, 114 (abstr)

Im S, Foulley JL, Gianola D (1987) A linear model for genetic evaluation on categorical traits. *In : 82nd annual meeting of the american dairy science association, Columbia, Missouri, june 21-24, 1987. J Dairy Sci* 70 suppl 1, 124, (abstr)
Janss L (1990) *Evaluation of beef bulls for direct effects on calving difficulties.* Technical report, University of Wageningen
Judge GG, Griffiths WE, Carter Hill R, Lütkepohl H, Lee TC (1985) *The theory and practice of econometrics.* John Wiley and Sons, New York, 3rd ed
Kendall MG, Stuart A (1961) *The advanced theory of statistics* Vol 2, Hafner, New York
Knuiman M, Laird N (1990) Parameter estimation in variance component models for binary response data. *In : Advances in Statistical Methods for Genetic Improvement of Livestock* (Gianola D, Hammond K, eds) Springer-Verlag, Heidelberg, 194-207
Kotz S, Johnson NL (1985) *Encyclopedia of statistical sciences.* John Wiley and Sons, New York, vol 6, 359
Lalouel JM, Rao DC, Morton NE, Elston RC (1983) A unified model. *J Hum Genet* 35, 816-826
Landis JR, Koch GG (1977) A one-way components of variance models for categorical data. *Biometrics* 33, 671-679
Lavergne C (1984) Contribution à l'étude des modèles à effets aléatoires dans l'analyse des données qualitatives. Thèse de 3^{e} cycle, Université Paul Sabatier, Toulouse
Lefort G (1980) Le modèle de base de la sélection : justifications et limites. *In : Biométrie et Génétique* (Legay J, Masson JP, Tomassone R, eds) INRA, département de biométrie, 4, 1-14
Madalla GS (1983) *Limited dependent and qualitative variables in econometrics.* Cambridge University Press, New York
Manfredi E (1990) Analyse génétique des conditions de naissance chez les bovins par le modèle à seuils. Thèse de Docteur en Sciences, Université de Paris Sud, Orsay
Manfredi E, San Cristobal M, Foulley JL, Gillard P, Valais A (1990) Genetic analysis of twinning in the Maine Anjou breed. *In : 41th Annual Meeting of the EAAP*, Toulouse, France, July 9-12, 1990, vol 1, 114 (abstr)
Manfredi E, Ducrocq V, Foulley JL (1991a). Genetic analysis of dystocia in cattle. *J Dairy Sci* 74, 1715-1723
Manfredi E, San Cristobal M, Foulley JL (1991b) Some factors affecting the estimation of genetic parameters for cattle dystocia under a threshold model. *Anim Prod* (in press)
Manfredi E, Foulley JL, San Cristobal M, Gillard P (1991c) Genetic parameters for twinning in the Maine-Anjou breed. *Genet Sel Evol.* (submitted)
Mc Cullagh P, Nelder J (1989) *Generalized linear models.* Chapman and Hall, London, 2nd edn
Mee RW (1982) Analysis of ordered categorical responses assuming an underlying variable. PhD thesis, Iowa State University, Ames, IA
Meijering A (1984) Dystocia and stillbirth in cattle-a review of causes, relations and implications. *Livest Prod Sci* 11, 143-177
Meijering A (1986) Dystocia in dairy cattle breeding. PhD thesis, Wageningen University

Misztal I, Gianola D, Foulley JL (1989) Computing aspects of a non linear method of sire evaluation for categorical data. *J Dairy Sci* 72, 1557-1568

Nelder JA, Wedderburn RWM (1972) Generalized linear models. *J R Statist Soc A* 135, 370-384

Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. *Biometrika* 58, 545-558

Pearson K (1990) Mathematical contributions to the theory of evolution. VIII. On the inheritance of characters not capable of exact quantitative measurement. *Phil Trans Roy Soc A*, 195, 79

Pearsson K (1904) *Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation.* Drapers Co Memoirs, vol 1, London

Petersson CJ, Danell O (1985) Factors influencing lamb survival in four Swedish sheep breeds. *Acta Agric Scand* 35, 217-232

Quaas RL, Van Vleck LD (1980) Categorical trait sire evaluation by best linear unbasied prediction of future progeny categories frequencies. *Biometrics* 36, 117-122

Quaas RL, Zhao Y, Pollack EJ (1988) Describing interactions in dystocia scores with a threshold model. *J Anim Sci* 66, 396-399

Rao CR (1973) *Linear statistical inferences and its applications.* J Wiley and Sons, New York, 2nd edn

Razungles J (1977) Héritabilité des caractères discrets : étude bibliographique critique. *Ann Génét Anim* 9, 43-61

Robertson A (1950) Proof that the additive heritability on the p scale is given by the expression $\bar{z}^2 h_x^2/\overline{pq}$. *Genetics* 35, 234-236

Robertson A, Lerner IM (1949) The heritability of all-or-none traits : viability of poultry. *Genetics* 34, 395-411

Ron M, Ezra E, Weller JI (1990) Genetics analysis of twinning rate in Israeli Holstein cattle. *Genet Sel Evol* 22, 349-359

Rönningen K (1971) Some properties of the selection index derived by "Henderson mixed model method". *Z Tierz Zuchtungsbiol* 88, 186-193

Searle SR (1971) *Linear models.* J Wiley and Sons, New York

Sellier P, Ollivier L (1982) Étude génétique du syndrome de l'abduction des membres (splayleg) chez le porcelet nouveau-né. *Ann Génét Sél Anim* 14, 77-92

Simianer H, Schaeffer LR (1989) Estimation of covariance components between continuous and one binary trait. *Genet Sel Evol* 21, 303-315

Stiratelli R, Laird N, Ware J (1984) Random effects model for serial observations with binary response. *Biometrics* 40, 961-971

Tallis G (1962) The maximum likelihood estimation of correlation from contingency tables. *Biometrics* 18, 342-353

Thompson R (1990) Generalized linear models and applications to animal breeding. *In : Advances in statistical methods for genetic improvement of livestock* (Gianola D, Hammond K, eds) Springer-Verlag, Heidelberg, 341-358

Thompson R, Mc Guirk BJ, Gilmour AR (1985) Estimating the heritability of all-or-none and categorical traits by offspring-parent regression. *Z Tierz Zuchtungsbiol* 102, 342-354

Tukey J (1962) The future of data analysis. *Ann Math Statist* 33, 1-67
Van Raden PM, Yung YC (1988) A general purpose approximation to restricted maximum likelihood : the tilde-hat approach. *J Dairy Sci* 71, 187-194
Van Vleck LD (1972) Estimation of heritability of threshold characters. *J Dairy Sci* 55, 218-225
Wedderburn RWM (1974) Quasi-likelihood and generalized linear models. *Biometrika* 61, 439-447
Williams DA (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenecity. *Biometrics* 31, 949-952
Williams DA (1982) Extra-binomial variation in logistic linear models. *Applied Statist* 31, 144-148
Williams DA (1988) Extra-binomial variation in toxicology : *In : 14th international biometric conference : invited papers.* Société Adolphe Quételet, Gembloux, 301-313
Wright S (1934a) An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* 19, 506-536
Wright S (1934b) The results of crosses between inbred strains of guinea pigs differing in number of digits. *Genetics* 19, 537-551
Wright S (1968) *Evolution and the genetics of population. 1. Genetic and biometric foundations.* The University of Chigaco Press, Chicago
Zeger SL, Liang KY (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 1-10
Zeger SL, Liang KY, Albert PS (1988) Models for longitudinal data : a generalized estimating equation approach. *Biometrics* 44, 1049-1066
Zellner A (1971) *An introduction to Bayesian inference in econometrics.* J Wiley and Sons, New York
Zellner A, Rossi PE (1984) Bayesian analysis of dichotomous quantal response models. *J Econom* 25, 365-393
Zhao Y (19876) Estimation of parameters in a mixed threshold model : its application to dystocia and birth weight in Simmental cattle. PhD thesis, Cornell University

ANNEXE A

Algorithme de Newton-Raphson appliqué à l'estimation des composantes u *de la variance par maximum de vraisemblance marginale*

Foulley *et al* (1989ab) ont montré qu'on pouvait obtenir le maximum de vraisemblance marginale pour un vecteur de paramètres $\boldsymbol{\gamma}$ sans intégration explicite des paramètres de nuisance $\boldsymbol{\theta}$ par un algorithme de Newton-Raphson basé sur l'expression des dérivées premières et seconde du logarithme $L(\boldsymbol{\gamma};\mathbf{y}) = \ln \int p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\gamma})d\boldsymbol{\theta}$ de la vraisemblance marginale :

$$\frac{\partial}{\partial \boldsymbol{\gamma}} L(\boldsymbol{\gamma};\mathbf{y}) = E_c \left\{ \frac{\partial}{\partial \boldsymbol{\gamma}} \ln p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\gamma}) \right\} \qquad [A.1]$$

$$\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} L(\boldsymbol{\gamma};\mathbf{y}) = E_c\left\{\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} \ln p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\gamma})\right\} + \text{Var}\left\{\frac{\partial}{\partial\boldsymbol{\gamma}} \ln p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\gamma})\right\} \quad [A.2]$$

où $E_c(.)$ et $\text{Var}_c(.)$ représentent l'espérance et la variance prises par rapport à la distribution *a posteriori* de $\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma}$.
Par définition $p(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\gamma}) = p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\gamma})p(\boldsymbol{\theta}|\boldsymbol{\gamma})$. On considérera le cas où $\boldsymbol{\theta}$ représente comme dans l'article les paramètres de position sur la sous-jacente et $\boldsymbol{\gamma} = \boldsymbol{\gamma}_u = \{\sigma^2_{u_k}\}$; $k = 1,2,\ldots,K$ représente le vecteur des composantes u de la variance pour une structure comportant comme en [17] un vecteur $\mathbf{u}$ formé de K sous-vecteurs $\mathbf{u}_k$ tel que $\mathbf{u} = (\mathbf{u}'_1,\mathbf{u}'_2,\ldots,\mathbf{u}'_k,\ldots,\mathbf{u}'_k)'$ avec $\mathbf{u}_k \sim N(\mathbf{0},\mathbf{A}_k\sigma^2_{u_k})$ et $\text{Cov}(\mathbf{u}_k,\mathbf{u}'_k) = \mathbf{0}$. Dans ce cas

$$p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\gamma}) \propto p(\mathbf{y}|\boldsymbol{\theta})$$
$$p(\boldsymbol{\theta}|\boldsymbol{\gamma}) \propto p(\mathbf{u}|\boldsymbol{\gamma}_u) = \prod_{k=1}^{K} p(\mathbf{u}_k|\sigma^2_{u_k}) \quad [A.3]$$

Le problème consiste alors à calculer les dérivées première et seconde de $\ln p(\mathbf{u}_k|\sigma^2_{u_k})$ par rapport à $\sigma^2_{u_k}$ puis à prendre l'espérance et la variance de la dérivée première ainsi que l'espérance de la dérivée seconde par rapport à la distribution de $\mathbf{u}_k|\mathbf{y},\boldsymbol{\gamma}_u$. Comme $\mathbf{u}_k|\sigma^2_{u_k} \sim N(\mathbf{0},\mathbf{A}_k\sigma^2_{u_k})$, ces dérivées s'écrivent :

$$\frac{\partial}{\partial\sigma^2_{u_k}} \ln p(\mathbf{u}_k|\sigma^2_{u_k}) = [(\mathbf{u}'_k\mathbf{A}_k^{-1}\mathbf{u}_k/\sigma^2_{u_k}) - q_k]/2\sigma^2_{u_k} \quad [A.4a]$$

$$\frac{\partial^2}{\partial(\sigma^2_{u_k})^2} \ln p(\mathbf{u}_k|\sigma^2_{u_k}) = [q_k - (2\mathbf{u}'_k\mathbf{A}_k^{-1}\mathbf{u}_k/\sigma^2_{u_k})]/2\sigma^4_u \quad [A.4b]$$

L'espérance et la variance sont donc celles de la forme quadratique $\mathbf{u}'_k\mathbf{A}_k^{-1}\mathbf{u}_k$ soit (Searle, 1971),

$$E_c(\mathbf{u}'_k\mathbf{A}_u^{-1}\mathbf{u}_k) = \hat{\mathbf{u}}'_k\mathbf{A}_k^{-1}\hat{\mathbf{u}}_k + \text{tr}\,(\mathbf{A}_k^{-1}\mathbf{C}_{uu,kk}) \quad [A.5a]$$

$$\text{Var}_c(\mathbf{u}'_k\mathbf{A}_u^{-1}\mathbf{u}_k) = 4\hat{\mathbf{u}}'_k\mathbf{A}_k^{-1}\mathbf{C}_{uu,kk}\mathbf{A}_k^{-1}\hat{\mathbf{u}}_k + 2\text{tr}\,(\mathbf{A}_k^{-1}\mathbf{C}_{uu,kk}\mathbf{A}_k^{-1}\mathbf{C}_{uu,kk}) \quad [A.5b]$$

où $\hat{\mathbf{u}}_k = E(\mathbf{u}_k|\mathbf{y},\boldsymbol{\gamma}_u)$; $\text{Var}(\mathbf{u}_k|\mathbf{y},\boldsymbol{\gamma}_u) = \mathbf{C}_{uu,kk}$

En reportant [A.4ab] dans [A.1 et 2] et en utilisant les expressions des moments donnés en [A.5b], l'algorithme de Newton Raphson s'écrit, de l'itération $[r]$ à $[r+1]$:

$$\sigma^{2[r+1]}_{u_k} = \sigma^{2[r]}_{u_k} + \frac{q_k\sigma^{6[r]}_{u_k} - E_c^{[r]}(\mathbf{u}'_k\mathbf{A}_k^{-1}\mathbf{u}_k)\sigma^{4[r]}_k}{q_k\sigma^{4[r]}_{u_k} - 2E_c^{[r]}(\mathbf{u}'_k\mathbf{A}_k^{-1}\mathbf{u}_k)\sigma^{2[r]}_{u_k} + (1/2)\text{Var}_c^{[r]}(\mathbf{u}'_k\mathbf{A}_k^{-1}\mathbf{u}_k)} \quad [A.6]$$

Une formule de ce type est mentionnée sans démonstration dans Höschele *et al* (1986). Il est à noter que le calcul de [A.6] nécessite les mêmes arguments que ceux utilisés dans l'algorithme de premier ordre, c'est-à-dire l'espérance et la variance de la distribution *a posteriori* des $\mathbf{u}$ obtenus à partir des équations de type de celles d'Henderson ou leurs approximations asymptotiques dans le cas de GF-HM.

ANNEXE B

Expression des termes de la matrice de covariance des fréquences p_j

Soit

$$p_j = \left(\sum_{r=1}^{n_j} y_{jr} \right) / n_j \qquad [B.1]$$

La fréquence de réponse dans la sous-population j où y_{jr} est une indicatrice binaire indicée par $r = 1, 2, \ldots, n_j$ distribuée conditionnellement sachant $\boldsymbol{\beta}$ et $\mathbf{u}$ comme une variable de Bernouilli de probabilité $\pi_j = \Phi(\eta_j) = \Phi(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u})$ (*cf* [23ab]).
Par définition de p_j en [B.1], l'espérance de y_{jr} est égale à celle de p_j soit à $\mathrm{E}(y_{jr}) = \mathrm{E}_{\mathbf{u}}[\Phi(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u})]$ ou avec les notations de [25 ab]

$$\mathrm{E}(y_{jr}) = \tilde{\pi}_j = \Phi(\tilde{\eta}_j) \qquad [B.2]$$

Enfin, on peut faire correspondre à y_{jr} une variable sous-jacente

$$x_{jr} = \mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u} + \varepsilon_{jr} \qquad [B.3]$$

telle que :

$$\varepsilon_{jr} = x_{jr} | \boldsymbol{\beta}, \mathbf{u} \sim \mathrm{N}(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u}; 1) \qquad [B.4a]$$

$$x_{jr} \sim \mathrm{N}(\mathbf{x}_j'\boldsymbol{\beta}; 1 + \sigma^2) \qquad [B.4b]$$

Eu égard à [B.1], la variance de p_j se décompose en :

$$\mathrm{Var}(p_j) = \left[\sum_r \mathrm{Var}(y_{jr}) + \sum_{r,r'} \mathrm{Cov}(y_{jr}, y_{jr'}) \right] / n_j^2 \qquad [B.5]$$

Comme y_{jr} est une variable binaire, elle vérifie la propriété $\mathrm{E}(y_{jr}) = \mathrm{E}(y_{jr}^2)$, d'où $\mathrm{Var}(y_{jr}) = \mathrm{E}(y_{jr})[1 - \mathrm{E}(y_{jr})]$ et en utilisant l'expression de $\mathrm{E}(y_{jr})$ en [B.2], on obtient (*cf* Foulley, 1987) :

$$\mathrm{Var}(y_{jr}) = \tilde{\pi}_j(1 - \tilde{\pi}_j) \qquad [B.6]$$

Par définition :

$$\mathrm{Cov}(y_{jr}, y_{jr'}) = \mathrm{E}(y_{jr}y_{jr'}) - \mathrm{E}(y_{jr})\mathrm{E}(y_{jr'})$$

Or, pour une variable binaire, $\mathrm{E}(y_{jr}y_{jr'}) = \Pr[(y_{jr} = 1) \cap (y_{jr'} = 1)]$. Vis-à-vis de la sous-jacente, cette probabilité conjointe équivaut à $\Pr[(x_{jr} > \tau) \cap (x_{jr} > \tau)]$, ($\tau$ étant le seuil), ou encore à $\Pr\left[\left(\frac{x_{jr} - \eta_j}{\sqrt{1+\sigma^2}} > \frac{\tau - \eta_j}{\sqrt{1+\sigma^2}}\right) \cap \left(\frac{x_{jr'} - \eta_j}{\sqrt{1+\sigma^2}} > \frac{\tau - \eta_j}{\sqrt{1+\sigma^2}}\right)\right]$.
L'origine étant au seuil ($\tau = 0$) et, en notant que $z_{jr} = \frac{x_{jr} - \eta_j}{\sqrt{1+\sigma^2}}$ est une variable normale réduite, la probabilité recherchée s'exprime comme une fonction

de répartition d'une loi binormale réduite, à savoir $\Pr[(z_{jr} < \widetilde{\eta}_j) \cap (z_{jr'} < \widetilde{\eta}_j)]$. Les variables z_{jr} et $z_{jr'}$ étant en corrélation de $t = \sigma^2/(1+\sigma^2)$, il vient

$$\Pr[(y_{jr} = 1) \cap (y_{jr'} = 1)] = \Phi_2(\widetilde{\eta}_j, \widetilde{\eta}_j; t)$$

où $\Phi_2(a, b; r)$ est la fonction de répartition de loi binormale réduite de corrélation r et d'arguments a, b.
Sachant [B.2], on a :

$$\mathrm{Cov}(y_{jr}, y_{jr'}) = \Phi_2(\widetilde{\eta}_j, \widetilde{\eta}_j; t) - \Phi^2(\widetilde{\eta}_j) \qquad [B.7]$$

et, en reportant [B.6] et [B.7] dans [B.5], on obtient en définitive :

$$\mathrm{Var}(p_j) = \{\widetilde{\pi}_j(1-\widetilde{\pi}_j) + (n_j - 1)[\Phi_2(\widetilde{\eta}_j, \widetilde{\eta}_j; t) - \widetilde{\pi}_j^2]\}/n_j \qquad [B.8]$$

Le raisonnement fait sur la variance s'applique de la même façon à la covariance avec, pour $j \neq j'$

$$\mathrm{Cov}(p_j, p_{j'}) = \mathrm{Cov}(y_{jr}, y_{j'r'}) \qquad [B.9a]$$
$$\mathrm{Cov}(y_{jr}, y_{j'r'}) = \mathrm{E}(y_{jr} y_{j'r'}) - \mathrm{E}(y_{jr})\mathrm{E}(y_{j'r'}) \qquad [B.9b]$$

À y_{jr} et $y_{j'r'}$, on peut associer comme précédemment les variables sous-jacentes normales réduites z_{jr} et $z_{j'r'}$ en corrélation de $\mathbf{z}_j' \boldsymbol{\Sigma}_u \mathbf{z}_{j'}/(1+\sigma^2)$, d'où

$$\mathrm{Cov}(p_j, p_{j'}) = \Phi_2[\widetilde{\eta}_j, \widetilde{\eta}_{j'}; \mathbf{z}_j' \boldsymbol{\Sigma}_u \mathbf{z}_{j'}/(1+\sigma^2)] - \Phi(\widetilde{\eta}_j)\Phi(\widetilde{\eta}_{j'}) \qquad [B.10]$$