Génét. Sél. Evol., 1988, **20** (2), 227-238

Posterior probability of the sire's genotype at a major locus based on progeny-test results for discrete characters

J.L. FOULLEY and J.M. ELSEN

Institut National de la Recherche Agronomique, Station de Génétique Quantitative et Appliquée,
Centre de Recherches de Jouy, 78350 Jouy-en-Josas, France
Institut National de la Recherche Agronomique, Station d'Amélioration Génétique des Animaux,
Centre de Recherches de Toulouse, BP 27, 31326 Castanet-Tolosan Cedex, France

Summary

This paper presents the expression of the posterior probability of the different possible genotypes of a sire at a major locus based on progeny performance for a discrete trait. Binomial, multinomial and Poisson distributions are considered fort this trait. Sires and dams are assumed unrelated. Polygenic as well a major gene effects are supposed to influence the trait recorded. An approximation to the computation of the posterior probability is suggested by expressing the latter, conditionally to estimated values of location and dispersion parameters (population, environmental and polygenic effects) which influence progeny performance.

Key words : major gene, Bayes' theorem, discrete characters.

Résumé

Probabilité a posteriori *du génotype paternel à un locus majeur à partir de l'observation des descendants : cas d'un caractère discret*

Cet article présente l'expression de la probabilité *a posteriori* des différents génotypes possibles d'un père à un locus majeur à partir des performances obtenues en contrôle de descendance pour un caractère discret. Des distributions binomiale, multinomiale et de Poisson sont considérées pour ce dernier. Les parents sont supposés non apparentés entre eux. Le caractère mesuré est supposé sous la dépendance d'un gène majeur et de polygènes. Un calcul approché de la probabilité est proposé basé sur l'expression de celle-ci, conditionnellement à des valeurs estimées des paramètres de position et de dispersion relatifs aux effets « population », environnementaux et polygéniques résiduels influençant les performances.

Mots clés : gène majeur, théorème de Bayes, caractère discret.

I. Introduction

For a trait with a mixed model of inheritance (a single locus and polygenes), rules to assign genotype at a major locus as a function of observed performance are often empirical. A good example for this is given by the *Booroola* gene with major effect on ovulation rate and litter size in sheep (Piper & Bindon, 1982). In that case, rules to assign genotype to females are usually those defined by Davis *et al.* (1982) ; *i.e.* : ewes are said carriers of the *F* gene if they have at least one observation where the ovulation rate is higher than 3. They are declared homozygote if one observation at least is higher than 5. These values 3 and 5 were derived from performance achieved in a relatively low prolific breed, the *Merino*. Obviously, these thresholds should be revised for more prolific breeds in which this gene is now introduced. Another example is muscular hypertrophy in cattle as affected by a major recessive gene (Vissac, 1972 ; Rollins *et al.*, 1972 ; Hanset & Michaux, 1985 *a* & *b*).

For such sex limited traits, inference about genotypes of males require information from related females, usually groups of paternal half sibs. In that situation, the former rule can be applied to female progeny individually and genotype inference based upon agreement between expected and observed numbers of the different genotypes in the female progeny of a given sire. Again, such criteria are open to criticism due to ignoring polygenic variation and for other methodological reasons (Foulley & Frebling, 1985).

To improve these empirical criteria, Elsen *et al.* (1988) dealing with progeny-testing of males, suggested to use computation of posterior probabilities. Classically, the trait observed in the progeny was assumed normally distributed. The purpose of this study is to extend this approach to discrete phenotypic distributions as often encountered in reproductive and productive performance.

II. Methodology

The following design is considered. Each sire whose genotype at a major locus is investigated, is randomly mated to several females having a given genotype (usually but not necessarily so), and progeny performance for a trait influenced by the major locus is recorded. Let us define the following symbols :

G_i : a random variable corresponding to the genotype of individual i,

g_i : a realized value of this random variable among r possible values coded by the integers 1, 2, ..., r ; G(g) will be used for sires and G*(g*) for progeny,

G and **g** : column vectors made up of former components,

y : a data vector with elements y_{ij} where i = 1, 2, ..., q refers to sire and j = 1, 2, ..., n_i to progeny within sire such as

$$\mathbf{y} = (\mathbf{y}'_1, \mathbf{y}'_2, \ldots, \mathbf{y}'_q)' \quad \text{with } \mathbf{y}_i = \{y_{ij}\} \text{ for } j = 1, 2, \ldots, n_i$$

$\mathbf{Y}$, $\mathbf{Y}_{ij}$, $\mathbf{Y}_i$ symbolize the corresponding random variables.

By a straightforward application of Bayes' theorem, one has :

$$\Pr(\mathbf{G} = \mathbf{g} \mid \mathbf{y}) \propto \Pr(\mathbf{G} = \mathbf{g})\ p(\mathbf{y}|\mathbf{G} = \mathbf{g}) \tag{1}$$

where $\Pr(\cdot)$ and $p(\cdot)$ refer to a probability and a density function respectively.

If sires and dams are unrelated, the random variables $(\mathbf{G}_i, \mathbf{Y}'_i)$ are independent and (1) reduces just to :

$$\Pr(\mathbf{G} = \mathbf{g}|\mathbf{y}) = \prod_i \Pr(G_i = g_i|\mathbf{y}_i) \tag{2}$$

and letting $\mathbf{G}^*_i = \{g^*_{ij}\}$ for $j = 1, 2, \ldots, n_i$, it can be shown that (ELSEN *et al.*, 1988)

$$\Pr(G_i = k|\mathbf{y}_i) \propto \Pr(G_i = k) \sum_{\mathbf{g}^*_i} \Pr(\mathbf{G}^*_i = \mathbf{g}^*_i|G_i = k)\ p(\mathbf{y}_i|\mathbf{g}^*_i) \tag{3}$$

with

$$\Pr(\mathbf{G}^*_i = \mathbf{g}^*_i|G_i = k) = \prod_{j=1}^{n_i} \Pr(G^*_{ij} = g^*_{ij}|G_i = k) \tag{4}$$

$$p(\mathbf{y}_i|\mathbf{g}^*_i) = \int_{R\alpha} p(\mathbf{y}_i|\mathbf{g}^*_i, \boldsymbol{\alpha})\ dF(\boldsymbol{\alpha}) \tag{5a}$$

$$\int_{R\alpha}[\prod_j p(y_{ij}|g^*_{ij}, \boldsymbol{\alpha})]dF(\boldsymbol{\alpha}) \tag{5b}$$

where $F(\boldsymbol{\alpha})$ is the cumulative density function pertaining to the prior distribution of a vector $\boldsymbol{\alpha}$ of parameters ($\boldsymbol{\alpha} \in R_\alpha$) involved in environmental and residual polygenic influences and which will be discussed more in detail in the next chapter.

Using (4) and (5b) in (3) and changing the order of summation and product operators, one obtains the general expression of the posterior probability that sire i has genotype k knowing its progeny performance $\mathbf{y}_i$ (ELSEN *et al.*, 1988) :

$$\Pr(G_i = k|\mathbf{y}_i)\ \alpha\ \Pr(G_i = k) \times \int_{R\alpha}[\prod_{j=1}^{n_i} \sum_{\ell=1}^{r} \Pr(G^*_{ij} = \ell|G_i = k)$$

$$p(y_{ij}|G^*_{ij} = \ell, \boldsymbol{\alpha})]dF(\boldsymbol{\alpha}). \tag{6}$$

Notice that (6) involves the following elements :

i) $\Pr(G_i = k)$ is the prior probability that sire i has genotype k. These probabilities depend on the mating structure out of which tested males are born. For instance, when introducing the *Booroola* gene into a foreign population by successive backcrossings, one will usually have heterozygote (*F*+) and homozygote normal (++) rams in equal proportions at each generation. In human genetics, these prior probabilities are generally assumed to be those of a Hardy-Weinberg population.

ii) $\Pr(G^*_{ij} = \ell|G_i = k)$ are elements of transition matrix such as the T matrix developed by GEPPERT & KOLLER (1938) and LI & SACKS (1954) when dams are chosen at random in a Hardy-Weinberg population.

iii) $p(y_{ij}|G^*_{ij} = \ell, \boldsymbol{\alpha})$ is the likelihood function for the observed performance of progeny j born out of sire i conditionally to its genotype and which will be discussed more in detail in the next section.

III. Likelihood functions

Several cases may be considered according to which distribution is hypothesized for the performance trait recorded.

A. *Normal distribution*

This situation has been described thoroughly by Elsen *et al.* (1988) and will just be summarized here so as to introduce the parameterization. Let us now designate the jth progeny of the ith sire by the subscript m, the conditional distribution of the progeny performance may be written as :

$$Y_m | (G^*_{m} = l), \boldsymbol{\alpha} \sim N(\mathbf{w}'_{ml}\, \boldsymbol{\theta}, \sigma^2_{e\ell}) \tag{7}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\theta}', \boldsymbol{\gamma}')$ is a concatenation of location ($\boldsymbol{\theta}$) and dispersion ($\boldsymbol{\gamma}$) parameters defined as follows :

$$\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}')'. \tag{8}$$

The $\boldsymbol{\beta}$ component usually represents « population » and systematic environmental effects such as herd × year × season, age of dam, etc.

To make the model flexible, these effects will be allowed to vary from one genotype to another. Therefore, we will write

$$\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_\ell, \ldots, \boldsymbol{\beta}'_r) \tag{9a}$$

and

$$\boldsymbol{\beta}'_\ell = (v_\ell, \mathbf{b}'_{\ell 1}, \mathbf{b}'_{\ell 2}, \ldots,) \tag{9b}$$

where v_ℓ is the progeny mean for genotype ℓ and $\mathbf{b}_{\ell 1}$, $\mathbf{b}_{\ell 2}$, ... are effects of factors 1, 2, ... expressed in deviation from this mean.

Polygenic variations will be represented by the vector $\mathbf{u}$ which may include different kinds of effects (Foulley *et al.*, 1987 *c*) *e.g.* sire and/or maternal grand sire transmitting abilities, additive genetic values and/or permanent influences. Again, these effects are formally allowed to vary according to the genotype of the progeny in which the trait is expressed

$$\mathbf{u}' = (\mathbf{u}'_1, \mathbf{u}'_2, \ldots, \mathbf{u}'_\ell, \ldots, \mathbf{u}'_r). \tag{10}$$

The vector $\boldsymbol{\gamma}$ of dispersion parameters is decomposed into :

$$\boldsymbol{\gamma}' = (\boldsymbol{\gamma}'_u, \boldsymbol{\gamma}'_e). \tag{11}$$

The vector $\boldsymbol{\gamma}_u$ component is formed by variances and covariances among elements in (10), say r variances $\rho^2_{u\ell}$ and r(r − 1)/2 covariances $\sigma_{u\ell\ell'}$, when just one factor (*e.g.* additive genetic values or sire) is considered.

Conditionally to $G^*_m = \ell$, $\boldsymbol{\beta}$ and $\mathbf{u}$, the distribution of progeny performance has a mean which is a specific linear combination of effects in $\boldsymbol{\theta}$, say $\mathbf{w}'_{m\ell}\, \boldsymbol{\theta}$, where $\mathbf{w}'_{m\ell}$ is a known row incidence matrix for progeny m knowing its genotype is ℓ and has a variance which is the usual residual variance $\sigma^2_{e_\ell}$. This leads to

$$\gamma_e' = (\sigma^2_{e_1}, \sigma^2_{e_2}, \ldots, \sigma^2_{e_\ell}, \ldots, \sigma^2_{e_r}) \tag{12}$$

when allowance is made for possible heterogeneity of residual variances among progeny genotypes.

B. *Binomial and multinomial distributions*

When the trait considered is an all-or-none trait, such as twinning in cattle for which a major gene effect has been suggested (Morris & Day, 1986), one will use the « liability » concept originally developed by Wright (1934). This model postulates an underlying normal distribution rendered dichotomous *via* an abrupt threshold.

Keeping the same notation as in the previous section, but with vector $\boldsymbol{\theta}$ defined now as a vector of effects in the underlying scale, we will take for a binary variate, $y_m = 0$ or 1 for categories coded [0] to the left and [1] to the right of the threshold respectively.

$$\Pr[Y_m = y_m | (G_m^* = \ell), \mu_{m\ell}] = [\Phi(\mu_{m\ell})]^{y_m}[1 - \Phi(\mu_{m\ell})]^{1 - y_m} \tag{13a}$$

with

$$\mu_{m\ell} = \mathbf{w}_{m\ell}' \, \boldsymbol{\theta} \tag{13b}$$

where $\Phi(\cdot)$ is the standardized normal cumulative density function and $\mu_{m\ell}$ the mean of the distribution of progeny performance conditionally to genotype ℓ, the origin being taken at the threshold. It has been assumed in (13 a & b) that the residual variance in the underlying scale $\sigma^2_{e\ell}$ is constant. This is not a very limiting constraint since the heterogeneity in residual variances in the underlying continuum results usually from a scale effect.

This approach can also be applied to ordered polytomies. In that case, the probability that progeny m responds in category q (q = 1, 2, ..., c + 1) can be written as

$$\Pr[Y_{mq} = 1 | (G_m^* = \ell), \eta_{m\ell}] = \Phi(t_q - \eta_{m\ell}) - \Phi(t_{q-1} - \eta_{m\ell}) \tag{14a}$$

where $t_1, t_2, \ldots, t_q, \ldots, t_c$ are parameters locating a reference population from the c thresholds (Gianola & Foulley, 1983) with $t_0 = -\infty$ and $t_{c+1} = +\infty$

and

$$\eta_{m\ell} = \mathbf{w}_{m\ell}' \, \boldsymbol{\theta} \tag{14b}$$

is a location parameter similar to (13b). Hence, the likelihood is product binomial or multinomial according to the distribution considered.

C. *Poisson distribution*

Some traits such as ovulation rate or litter size might be better described with a Poisson than a multinomial distribution. Then, following Foulley *et al.*, (1987 *c*), the conditional distribution of progeny performance is given by

$$\Pr[Y_m = y_m | (G_m^* = \ell), \lambda_{m\ell}] = \lambda_{m\ell}^{y_m} \exp(-\lambda_{m\ell})/y_m\,! \tag{15a}$$

with

$$\lambda_{m\ell} = \exp(\mu_{m\ell}) \tag{15b}$$

$$\mu_{m\ell} = \mathbf{w}'_{m\ell}\ \boldsymbol{\theta} \tag{15c}$$

As pointed out by these authors, one may also envision truncated Poisson distributions. For zero excluded, (15a) must be replaced by

$$\Pr[Y_m = y_m > 0 \mid (G^*_m = \ell),\ \lambda_{m\ell}] = \lambda^{Y_m}_{m\ell} / \{y_m\ !\ [\exp(\lambda_{m\ell}) - 1]\}. \tag{16}$$

IV. Computations

Computing the exact posterior probability that sire i has genotype ℓ according to formula (6) is formidable task which requires integrating out both location ($\boldsymbol{\theta}$) and dispersion ($\boldsymbol{\gamma}$) parameters. Even when $\boldsymbol{\gamma}$ is known, this integration involves the calculation or Σr^{n_i} terms which can be expressed analytically only in the normal case (ELSEN *et al.*, 1988).

Therefore some approximations are necessary especially with discrete distributions. As suggested by ELSEN *et al.*, (1988) with normal traits and GIANOLA *et al.* (1986) and FOULLEY *et al.* (1987 *a, b*) in genetic evaluation problems, one can evaluate (1) conditionally to $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ values estimated from the data *i.e.* compute

$$\Pr(\mathbf{G} = \mathbf{g}|\mathbf{y},\ \boldsymbol{\theta}^*\ (\boldsymbol{\gamma}^*)] \tag{17}$$

$$\boldsymbol{\theta}^*(\boldsymbol{\gamma}) :\ \mathrm{Max}_{\boldsymbol{\theta}}\ \log\ p(\boldsymbol{\theta}|\mathbf{y},\ \boldsymbol{\gamma}) \tag{18a}$$

$$\boldsymbol{\gamma}^* :\ :\ \mathrm{Max}_{\boldsymbol{\gamma}}\ \log\ p(\boldsymbol{\gamma}|\mathbf{y}) \tag{18b}$$

where

$\boldsymbol{\theta}^*$ ($\boldsymbol{\gamma}$) is the mode of the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\gamma}$

$\boldsymbol{\gamma}^*$ is the mode of the marginal posterior distribution of $\boldsymbol{\gamma}$ which corresponds to the marginal maximum likelihood (ML) estimator when a flat prior for $\boldsymbol{\gamma}$ is used.

A. *Estimation of location parameters*

In consideration of the threshold-liability model, we will take as prior density (GIANOLA & FOULLEY, 1983 ; FOULLEY *et al.*, 1987 *c*)

$$p(\boldsymbol{\theta}|\boldsymbol{\Gamma},\ \boldsymbol{\theta}_0) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\ \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right]. \tag{19}$$

Usually, but no necessarily so, one takes $\boldsymbol{\theta}'_0 = (\boldsymbol{\delta}', \mathbf{0})$ and the $\boldsymbol{\beta}$ component of $\boldsymbol{\Gamma}$, say $\boldsymbol{\Gamma}_\beta \rightarrow \infty$, so as to mimic a mixed model structure with $\boldsymbol{\Gamma}$ depending on γ_u only. Now, with the same assumptions as before, the likelihood can be written as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{q}\ \prod_{j=1}^{n_i}\ \sum_{\ell=1}^{r}\ \Pr(G^*_{ij} = \ell)p\ [y_{ij}|(G^*_{ij} = \ell),\ \boldsymbol{\theta}] \tag{20}$$

$$\Pr(G^*_{ij} = \ell) = \sum_{k=1}^{r}\ \Pr(G_i = k)\ \Pr(G^*_{ij} = \ell|G_i = k). \tag{21}$$

Hence, the posterior distribution is

$$p(\boldsymbol{\theta}|y, \boldsymbol{\Gamma}, \boldsymbol{\theta}_0) \propto (19) \times (20). \tag{22}$$

Maximizing (22) with respect to $\boldsymbol{\theta}$ involves solving a nonlinear system of equations, using for instance the Newton-Raphson algorithm (FOULLEY *et al.*, 1987 *b*). Let $L(\boldsymbol{\theta})$ be the logarithm of the posterior density defined in (22) and, again m being a single subscript for the combination ij, the first and second partial derivatives of L with respect to $\boldsymbol{\theta}$ are :

$$\dot{L}(\boldsymbol{\theta}) = -\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \sum_m \sum_\ell q_{m\ell}\, v_{m\ell}\, \mathbf{w}_{m\ell} \tag{23a}$$

$$\ddot{L}(\boldsymbol{\theta}) = -\Gamma^{-1} - \sum_m \sum_\ell r_{m\ell}\, \mathbf{w}_{m\ell}\, \mathbf{w}'_{m\ell}. \tag{23b}$$

Hence, putting

$$\mathbf{s}_\ell = \{q_{m\ell}\ v_{m\ell}\} \text{ for } m = 1, 2, \ldots, N \tag{24a}$$

$$\mathbf{R}_\ell = \text{Diag } \{r_{m\ell}\} \text{ for } m = 1, 2, \ldots, N \tag{24b}$$

$$\mathbf{W}_\ell = (\mathbf{w}_{1\ell}, \mathbf{w}_{2\ell}, \ldots, \mathbf{w}_{m\ell}, \ldots, \mathbf{w}_{N\ell})' \tag{24c}$$

$$\mathbf{z}_\ell = \mathbf{W}_\ell\boldsymbol{\theta} + \mathbf{R}_\ell^{-1}\, \mathbf{s}_\ell \tag{24d}$$

the Newton-Raphson algorithm consists in iterating from round t to t + 1 with

$$[(\sum_\ell \mathbf{W}'_\ell\, \mathbf{R}_\ell^{[t]}\, \mathbf{W}_\ell) + \Gamma^{-1}]\, \boldsymbol{\theta}^{[t+1]} = \sum_\ell \mathbf{W}'_\ell\, \mathbf{R}_\ell^{[t]}\, \mathbf{z}_\ell^{[t]} + \boldsymbol{\Gamma}^{-1}\, \boldsymbol{\theta}_0 \tag{25}$$

where

$$q_{m\ell} = \frac{\Pr(G^*_m = \ell)\ p[y_m|(G^*_m = \ell), \boldsymbol{\theta}, \Gamma]}{\sum_\ell \Pr(G^*_m = \ell)\ p[y_m|(G^*_m = \ell), \boldsymbol{\theta}, \Gamma]} \tag{26a}$$

$$v_{m\ell} = \partial \log p[y_m|(G^*_m = \ell), \boldsymbol{\theta}, \Gamma]/\partial\mu_{m\ell} \tag{26b}$$

$$r_{m\ell} = -q_{m\ell}(\partial v_{m\ell}/\partial\mu_{m\ell}) - q_{m\ell}(1 - q_{m\ell})v^2_{m\ell} \tag{26c}$$

Analytical expressions of these coefficients can be derived explicitly for the different discrete distributions considered previously, *i.e.*

i) for a Bernouilli variate

$$v_{m\ell} = \varphi(\mu_{m\ell})\ [y_m - \Phi(\mu_{m\ell})]/\{\Phi(\mu_{m\ell})\ [1 - \Phi(\mu_{m\ell})]\} \tag{27a}$$

$$r_{m\ell} = q_{m\ell}v_{m\ell}(q_{m\ell}v_{m\ell} + \mu_{m\ell}) \tag{27b}$$

where φ is the standardized normal density function. For several ordered categories, explicit formulae for $v_{m\ell}$ and $E(\partial v_{m\ell}/\partial\eta_{m\ell})$ are shown in GIANOLA & FOULLEY (1983) ; moreover, the system in (25) must be augmented by sectors pertaining to the thresholds.

ii) for a Poisson variate

As shown by FOULLEY *et al.* (1987 *c*)

$$v_{m\ell} = y_m - \lambda_{m\ell} \tag{28a}$$

$$r_{m\ell} = q_{m\ell}\ \lambda_{m\ell} - q_{m\ell}\ (1 - q_{m\ell})\ (y_m - \lambda_{m\ell})^2. \tag{28b}$$

If the Poisson model truncated at zero is employed, $v_{m\ell}$ becomes

$$v_{m\ell} = y_m - \{\lambda_{m\ell}/[1 - \exp(-\lambda_{m\ell})]\} \tag{29a}$$

and $r_{m\ell}$ is calculated according to (26c) with

$$\partial v_{m\ell}/\partial \mu_{m\ell} = -\frac{\lambda_{m\ell}}{1-\exp(-\lambda_{m\ell})}\left[1-\frac{\lambda_{m\ell}\exp(-\lambda_{m\ell})}{1-\exp(-\lambda_{m\ell})}\right]. \qquad (29b)$$

The system in (25) with formulae (26a, b & c) are similar to those given by Foulley *et al.* (1987 *a*) for genetic evaluation with uncertain parternity. The $q_{m\ell}$ coefficient gives rise to an interesting interpretation since (26a) can be viewed as the posterior probability that progeny m has genotype ℓ. This expression would also occur naturally if the problem was set up in terms of incomplete data and solve accordingly *via* the EM algorithm (Dempster *et al.*, 1977, p. 16). As a matter of fact, letting $q_{m\ell} = 1$ in the expressions of $v_{m\ell}$ and $r_{m\ell}$, one obtains the usual coefficients encountered in genetic evaluation for binary (Foulley *et al.*, 1983 ; Foulley *et al.*, 1987 *d*) and Poisson variates (Foulley *et al.*, 1987 *c*). Finally, the form of the system in (25) indicates that the analysis is carried out conditionally to the different possible genotypes of progeny with appropriate weighting factors on the left and right handsides depending on the posterior probabilities of genotypes. This generates two sources of nonlinearity as shown clearly in the formula for $r_{m\ell}$ (26c), one due to the form of the distribution and the second to uncertainty about genotype of progeny.

B. *Estimation of dispersion parameters*

The value of $\boldsymbol{\theta}$ calculated from (25) is the mode of the condition distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\Gamma}$ and the data. It remains to replace $\boldsymbol{\Gamma}$ by its marginal ML estimator. For the sake of simplicity we will consider the case of just one random factor **u** such as sire transmitting abilities or individual additive genetic values. Then, the unknown is the vector $\boldsymbol{\gamma}_u$ (see 11) formed by the r(r + 1)/2 different elements of the $\mathbf{G} = \{g_{k\ell}\}$ matrix of « u » components of variance and covariance. The general procedure for discrete variates has been presented by Harville & Mee (1984), Foulley *et al.* (1987 *b* & *c*). These authors have shown that maximization of the logposterior density of $\boldsymbol{\gamma}_u$ with respect to $\boldsymbol{\gamma}_u$ using a diffuse prior, results in the equation

$$E_C\left\{\frac{\partial}{\partial \boldsymbol{\gamma}_u}\log[p(\mathbf{u}|\boldsymbol{\gamma}_u)]\right\} = \mathbf{0} \qquad (30)$$

here E_C indicates expectation with respect to the density $p(\mathbf{u}|y, \boldsymbol{\gamma}_u)$.

Within the framework of a normal distribution for $\mathbf{u}|\boldsymbol{\gamma}_u$, the formal solution to (30) is very general whatever form of the likelihood. Provided some approximations are made about the first two moments of $\mathbf{u}|y$, $\boldsymbol{\gamma}_u$, computations amount to iterate with

$$g_{k\ell}^{[s+1]} = [\mathbf{u}'^{[s]}_k \mathbf{u}^{[s]}_\ell + \text{tr}\,(\mathbf{C}^{[s]}_{k\ell})]/q \qquad (31)$$

where

$\mathbf{u}^{[s]}_k$, $\mathbf{u}^{[s]}_\ell$ are solutions to (25) in $\mathbf{u}_k$ and $\mathbf{u}_\ell$ given $\boldsymbol{\gamma}_u = \boldsymbol{\gamma}^{[s]}_u$

$\mathbf{C}^{[s]}_{k\ell}$ is the (q × q) submatrix pertaining to genotypes k and ℓ in the **u** part of the inverse of the coefficient matrix in (25).

C. *Computation of the posterior probabilities*

Under the same assumptions as in II (no genetic relationships among parents), formula (17) may be simply written as

$$(17) = \prod_i \Pr[G_i = g_i|y_i, \boldsymbol{\theta}^*(\boldsymbol{\gamma}^*)].$$

Hence, after having estimated $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, the posterior probability that sire i has genotype k will be calculated as

$$\Pr[G_i = k|\mathbf{y}_i, \boldsymbol{\theta}^*(\boldsymbol{\gamma}^*)] = N_{ik}/D_i \tag{32a}$$

with

$$N_{ik} = \Pr(G_i = k) \prod_{j=1}^{n_i} \sum_{\ell=1}^{r} \Pr(G^*_{ij} = \ell|G_i = k)\, p[y_{ij}|(G^*_{ij} = \ell), \boldsymbol{\theta}^*(\boldsymbol{\gamma}^*)] \tag{32b}$$

$$D_i = \sum_{k=1}^{r} N_{ik}. \tag{32c}$$

The last terms in (32 b) are computed from (13 a & b), (14 a & b) and (15 a, b & c) replacing $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^*(\boldsymbol{\gamma}^*)$.

V. Discussion-Conclusion

It has been implicitly assumed till now that one record per progeny is available. Taking into account several performances per animal can be easily achieved using a « repeatability » model as follows :

$$p(\mathbf{y}_i|\mathbf{g}_i, \boldsymbol{\alpha}) = \prod_{jt} p[y_{ijt}|(G^*_{ij} = g^*_{ij}), \boldsymbol{\alpha}]$$

where t is the subscript for the tth performance within progeny ij. The location parameter used in the likelihood is then, for progeny m given genotype ℓ :

$$\mu_{m\ell t} = \mathbf{w}'_{m\ell t}\, \boldsymbol{\theta}$$

where, as previously $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \mathbf{u}')$, and $\mathbf{u}$ can be parameterized as $\mathbf{u}' = (\mathbf{s}', \mathbf{p}')$, $\mathbf{s}$ being a vector of sire effects and $\mathbf{p}$ a vector of permanent environmental effects for a given progeny within sire. The corresponding dispersion parameters become

$$\boldsymbol{\Gamma}_u = \begin{pmatrix} \mathbf{I}\sigma_s^2 & \mathbf{O} \\ \mathbf{O} & \mathbf{I}\sigma_p^2 \end{pmatrix}$$

and the coefficient $\rho_k = (\sigma^2_{sk} + \sigma^2_{pk})/(\sigma^2_{sk} + \sigma^2_{pk} + \sigma^2_{\ell k})$

designates the usual repeatability coefficient for genotype k which may be assumed constant as in Elsen *et al.* (1988). The procedure described in the previous chapter IV is still valid especially the method for estimating dispersion parameters which can be easily extended to several sets of random factors (Foulley *et al.*, 1987 *c*).

It is also worth mentioning that the approach followed in chapter IV provides as a by-product a genetic evaluation method for traits with a mixed model of inheritance (Elston & Stewart, 1971 ; Morton & McLean, 1974 ; Lalouel *et al.*, 1983) and with completely or partially unknown genotypic information. In that case, the coefficient matrix in (25) may be very large especially with field data sets. Some problems encountered in solving such large non linear systems (*e.g.* convergence properties, precision, computing costs) have been recently discussed by Mistzal & Gianola (1987) for sire evaluation programs dealing with threshold traits. Anyhow, this is another example of how the Bayesian paradigm can be used to solve problems in that area which cannot be readily addressed *via* the BLUP machinery.

However with the method proposed, nuisance parameters have been averaged out and not exactly integrated out as it should be. Therefore, it is important to bear in mind that this procedure as others (HASSTEDT, 1982) is an approximation, the domain of validity of which should be more carefully addressed using for instance realistic examples with Monte-Carlo simulation techniques.

Some practical provisions can be suggested to apply this procedure shrewdly. First, genetic and phenotypic parameters must be taken as known to reduce the degree of nonlinearity of the problem especially when a limited number of sires is tested. Secondly, one has better to choose at start a simple parameterization with no specific effects and variances according to genotypes of progeny. Finally, when the distribution of performance is clearly multimodal, one may expect some of the coefficients q to have extreme values. In order to improve the assignment of genotypes to sires, one might use for instance some prior information about genotypic means in calculating the q's at the first round of iteration. Anyhow, we are truly conscious that the approximations proposed in (17) and (18 a & b) may severely limit the potential interest of this methodology as long as formula (6) cannot be calculated efficiently by numerical procedures of integration.

Received June 1, 1987.
Accepted September 9, 1987.

Acknowledgements

The authors would like to thank the two anonymous referees for valuable comments which helped to improve the manuscript. They are also grateful to Ina HÖSCHELE to let them know after this paper was submitted that she had independently dealt with a related problem in an article entitled « genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance » and submitted to *Theoretical & Applied Genetics.*

References

DAVIS G.H., MONTGOMERY G.W., ALLISON A.J., KELLY R.W., BRAY A.R., 1982. Segregation of a major gene influencing fecundity in progeny of Booroola sheep. *N.Z. J. Agr. Res.*, **25,** 525-529.

DEMPSTER A.P., LAIRD N.M., RUBIN R.B., 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.*, **B, 39,** 1-20.

ELSEN J.M., VU TIEN KHANG J., LE ROY P., 1988. A statistical model for genotype determination at a major locus in a progeny test design. *Génét. Sél. Evol.*, **20,** 211-226

ELSTON N.E., STEWART J., 1971. A general model for the genetic analysis of pedigree data. *Hum. Hered.*, **21,** 523-542.

FOULLEY J.L., FREBLING J., 1985. Critères de détection indirecte des taureaux porteurs de la translocation 1-29 à partir des caryotypes de leurs descendants. *Génét. Sél. Evol.*, **17,** 341-350.

FOULLEY J.L., 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. Evol.*, **15,** 407-424.

FOULLEY J.L., GIANOLA D., PLANCHENAULT D., 1987 *a*. Sire evaluation with uncertain paternity. *Génét. Sél. Evol.*, **19,** 83-102.

FOULLEY J.L., IM S., GIANOLA D., HÖSCHELE I., 1987 *b*. Empirical Bayes estimation of parameters for n polygenic binary traits. *Génét. Sél. Evol.*, **19**, 197-224.

FOULLEY J.L., GIANOLA D., IM S., 1987 *c*. Genetic evaluation for traits distributed as Poisson-binomial with reference to reproductive traits. *Theor. Appl. Genet.*, **73**, 870-877.

FOULLEY J.L., GIANOLA D., IM S., 1987 *d*. Genetic evaluation for discrete polygenic traits in animal beeding. *In : Advances in statistical methods for genetic improvement of animals,* Springer-Verlag, Berlin, forthcoming.

GEPPERT S., KOLLER S., 1938. *Erbmathematik : Theorie der Veterbung in Bevölkerung und Sippe.* 228 p., Quelle und Meyer, Leipzig.

GIANOLA D., FOULLEY J.L., 1983. Sire evaluation for ordered categorical data with a threshold model. *Génét. Sél. Evol.*, **15**, 201-224.

GIANOLA D., FOULLEY J.L., FERNANDO R.L., 1986. Prediction of breeding values when variances are not known. *Génét. Sél. Evol.*, **18**, 485-498.

HANSET R., MICHAUX C., 1985 *a*. On the genetic determinism of muscular hypertrophy in the Belgian White and Blue cattle breed. 1. Experimental data. *Génét. Sél. Evol.*, **17**, 359-368.

HANSET R., MICHAUX C., 1985 *b*. On the genetic determinism of muscular hypertrophy in the Belgian White and Blue cattle breed. 2. Population data. *Génét. Sél. Evol.*, **17**, 369-386.

HARVILLE D.A., MEE R.W., 1984. A mixed model procedure for analyzing ordered categorical data. *Biometrics,* **40**, 393-408.

HASSTEDT S.J., 1982. A mixed-model likelihood approximation on large pedigrees. *Comput. Biomed. Res.*, **15**, 295-307.

LALOUEL J.M., RAO D.C., MORTON N.E., ELSTON R.C., 1983. A unified model for complex segregation analysis. *Am. J. Hum. Genet.*, **35**, 816-826.

LI C.C., SACKS L., 1954. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics,* **10**, 347-360.

MISZTAL I., GIANOLA D., 1987. Computing aspects of a nonlinear method of sire evaluation for categorical data. *J. Dairy Sci.*, **70**, suppl. 1, 124 (abstr.).

MORRIS C.A., DAY A.M., 1986. Ovulation results from cattle herds with high twinning frequency. *In :* DICKERSON G.E., JOHNSON R.K. (ed.), *3rd World Congress on Genetics applied to Livestock Production, Lincoln, Nebraska, July 16-22, 1986,* **11**, 96-100, Editorial Garsi, Madrid.

MORTON N.E., MCLEAN C.J., 1974. Analysis of family resemblance. 3. Complex segregation analysis of quantitative traits. *Am. J. Hum. Genet.*, **26**, 489-503.

PIPER L.R., BINDON B.M., 1982. Genetic segregation for fecundity in Booroola Merino sheep. *In* : BARTON R.A., SMITH W.C. (ed.), *Proceedings of the World Congress on Sheep and Beef Cattle Breeding,* vol. 1 (Technical), 395-400, The Dunmore Press, Palmerston North.

ROLLINS W.C., TANAKA M., NOTT C.F.G., THIESSEN R.B., 1972. On the mode of inheritance of double muscled conformation in bovines. *Hilgardia,* **41**, 433-456.

VISSAC B., 1972. L'hypertrophie musculaire d'origine génétique ou caractère culard. *Ann. Génét. Sél. Anim.*, **4**, 87-97.

WRIGHT S., 1934. An analysis of variability in number of digits in an inbred strain of Guinea pigs. *Genetics,* **19**, 506-536.