Génét. Sél. Evol., 1987, **19** (1), 83-102

Sire evaluation with uncertain paternity

J.L. FOULLEY *,**, D. GIANOLA ** and D. PLANCHENAULT ***

* *I.N.R.A., Station de Génétique quantitative et appliquée*
Centre de Recherches Zootechniques, F 78350 Jouy-en-Josas

** *Department of Animal Sciences, University of Illinois, Urbana, Illinois 61801, U.S.A.*

*** *Institut d'Elevage et de Médecine Vétérinaire des Pays Tropicaux,*
10, rue Pierre-Curie, F 94704 Maisons-Alfort Cedex

Summary

A sire evaluation procedure is proposed for situations in which there is uncertainty with respect to the assignment of progeny to sires. The method requires the specification of the prior probabilities p_{ij} that progeny i is out of sire j. Inferences about location parameters (« fixed » environmental and group effects and transmitting abilities of sires) are based on Bayesian statistical procedures. Modal values of the posterior distribution of these parameters are taken as point estimators. Finding this mode entails solving a nonlinear system of equations and several algorithms are suggested. The methodology is described for univariate evaluations obtained from normal or binary traits. Estimation of unknown variances is also addressed. A small numerical example is presented to illustrate the procedure. Potential applications to livestock breeding are discussed.

Key words : Sire evaluation, uncertain paternity, Bayesian methods.

Résumé

Evaluation des pères dans le cas de paternité incertaine

Une méthode d'évaluation des pères est proposée en situation d'incertitude vis-à-vis de l'assignation des descendants à leurs pères. La méthode requiert la spécification des probabilités *a priori* p_{ij} que le descendant i provienne du père j. L'inférence des paramètres de position (effets « groupe » et de milieu, considérés comme fixes et valeurs génétiques transmises des pères) est basée sur des procédures statistiques bayésiennes. Les valeurs modales de la distribution *a posteriori* de ces paramètres ont été prises comme estimateurs ponctuels. La recherche du mode nécessite la résolution d'un système d'équations non linéaire pour lequel plusieurs algorithmes sont proposés. La méthodologie est développée dans le cadre univariate pour des caractères normaux et binaires. Le cas de variances inconnues est également abordé. Un petit exemple numérique est présenté à titre d'illustration. Enfin, les applications possibles aux espèces domestiques sont discutées.

Mots clés : Evaluation des reproducteurs, paternité incertaine, méthodes bayésiennes.

I. Introduction

There are situations such as in multiple-sire matings under pastoral conditions where sire evaluation is complicated because of uncertainty with respect to the assignment of progeny to sires. Using information from red blood cell types, major histocompatibility markers or precise records on breeding period and gestation length, it is possible to specify the probabilities (p_{ij}) that a given offspring (i = 1, ..., n) has been sired by different males (j = 1, ..., m). In the absence of such information, it is reasonable to state that individual males in a given set, e.g., bulls breeding in the same paddock, are sires with equal probability. This problem was studied by POIVEY & ELSEN (1984) within the framework of selection index and its restrictive assumptions. The purpose of this paper is to present a more general and flexible methodology able to cope with several sources of variation including unknown fixed effects and variance components. The procedure is along the lines of linear and nonlinear mixed model methodology (HENDERSON, 1973 ; GIANOLA & FOULLEY, 1983a, b). Continuous and discontinuous variation are examined in this paper to illustrate the power and generality of the approach.

II. Normally distributed data

A. *Methodology*

Consider the usual univariate linear model :

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of records, $\boldsymbol{\beta}$ is an $l \times 1$ vector of « fixed » effects (e.g., genetic groups, « nuisance » environmental factors), $\mathbf{u}$ is an $m \times 1$ vector of random transmitting abilities of sires, $\mathbf{X}$ and $\mathbf{Z}$ are instance matrices, and $\mathbf{e}$ is a vector of residuals. The matrices $\mathbf{X}$ and $\mathbf{Z}$ are known (non-random), if the sires of the progeny with records in $\mathbf{y}$ are identified. In other words, the above model holds *conditionally* on $\mathbf{X}$ and $\mathbf{Z}$.

Let $\mathcal{L}_{ij}$ define the situation in which male j is the true sire of progeny i. The conditional distribution of the record y_i given $\mathcal{L}_{ij}$, the location parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ and the residual variance σ_e^2 can be written as

$$y_i|\mathcal{L}_{ij}, \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2 \sim \text{NIID}\ (\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}, \sigma_e^2) \qquad [1]$$

where NIID stands for normal, independent and identically distributed ; $\mathbf{z}_{ij}$ is an $m \times 1$ vector having a 1 in position j and 0's elsewhere. Put $\mathbf{w}_{ij}' = [\mathbf{x}_i', \mathbf{z}_{ij}']$, $\boldsymbol{\theta}' = [\boldsymbol{\beta}', \mathbf{u}']$ and define $\mu_{ij} = \mathbf{w}'_{ij}\boldsymbol{\theta}$. Inferences about $\boldsymbol{\theta}$ can be obtained conveniently via Bayes theorem, and this has also been done in other genetic evaluation problems (RÖNNINGEN, 1971 ; DEMPFLE, 1977 ; LEFORT, 1980 ; GIANOLA & FERNANDO, 1986). The prior distribution of $\boldsymbol{\theta}$ is « naturally » taken as the conjugate of [1] (COX & HINKLEY, 1974) so

$$\boldsymbol{\theta}|\boldsymbol{\Sigma} \sim \text{N}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}) \qquad [2]$$

where $\boldsymbol{\alpha}' = [\boldsymbol{\delta}', \mathbf{0}]$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_\beta & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_u \end{bmatrix}$$

It will be assumed from now on that prior knowledge about $\boldsymbol{\beta}$ is vague so as to mimic the traditional mixed model analysis. Hence, the prior distribution of $\boldsymbol{\theta}$ is strictly proportional to the marginal prior distribution of $\mathbf{u}$. However, the notation of [2] above is retained to present a more general expression for the posterior distribution of the vector $\boldsymbol{\theta}$. The matrix $\boldsymbol{\Sigma}_u = \mathbf{A}\,\sigma_u^2$, where $\mathbf{A}$ is the matrix of additive relationships between sires, and σ_u^2 is the variance between sires, equal to one quarter of the additive genetic variance.

Because the observations are conditionally independent, the likelihood function can be written as :

$$f(\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2) = \prod_i f(y_i|\boldsymbol{\theta}, \sigma_e^2) \qquad [3A]$$

where

$$f(y_i|\boldsymbol{\theta}, \sigma_e^2) = \sum_j p_{ij} \cdot f(y_i|\mathscr{L}_{ij}, \boldsymbol{\theta}, \sigma_e^2) \qquad [3B]$$

because $\sum_j p_{ij} = 1$. The mean of the distribution in [3B] is

$$E(y_i|\boldsymbol{\theta}, \sigma_e^2) = \underset{\mathscr{L}_{ij}}{E} [E(y_i|\mathscr{L}_{ij}, \boldsymbol{\theta}, \sigma_e^2)]$$

$$= \mathbf{x}_i'\boldsymbol{\beta} + \underset{\mathscr{L}_{ij}}{E}(\mathbf{z}_{ij}')\,\mathbf{u} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{p}_i'\mathbf{u} = \mu_i$$

where $\mathbf{p}_i' = [p_{i1}, \ldots, p_{ij}, \ldots, p_{im}]$ is a $1 \times m$ row vector containing the probabilities p_{ij} of $\mathscr{L}_{ij}$ (progeny i out of sire j). As shown in Appendix A, the variance of the distribution [3B] is

$$\mathrm{Var}(y_i|\boldsymbol{\theta}, \sigma_e^2) = \sigma_e^2 + \sum_j p_{ij}(\mu_{ij} - \mu_i)^2$$

The posterior distribution of $\boldsymbol{\theta}$ (assuming that the dispersion parameters are known, can be written from, [1], [2], [3A] and [3B] as

$$f(\boldsymbol{\theta}|\mathbf{y}, \sigma_e^2, \sigma_u^2) \propto f(\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2) \cdot f(\boldsymbol{\theta}|\sigma_u^2)$$

$$\propto \{\prod_i \sum_j p_{ij} \exp[-(y_i - \mathbf{w}_{ij}'\boldsymbol{\theta})^2/2\sigma_e^2] \cdot \exp[-(\boldsymbol{\theta} - \boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\alpha})/2] \qquad [4]$$

which is not in the form of a normal distribution. Hence, the mean of this distribution cannot be a linear function of the data.

The selection rule which maximizes the expected transmitting ability of a fixed number of selected sires is the mean of the posterior distribution [4] (GOFFINET & ELSEN, 1984 ; FERNANDO & GIANOLA, 1986). Because the expected value of this distribution is difficult to obtain in closed form, we calculate the modal value of $\boldsymbol{\theta}$ and regard the $\mathbf{u}$ component of this mode as an approximation to the optimum selection

rule in the sense described above ; this is a reasonable approximation as sample size increases (ZELLNER, 1971).

B. *Computations*

Finding the maximum of [4] with respect to $\boldsymbol{\theta}$ requires setting to $\mathbf{0}$ the first derivatives of [4] with respect to this vector. Letting $L(\boldsymbol{\theta})$ be the log-posterior density, we obtain :

$$\frac{\delta L(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}} = (\sigma_e^2)^{-1} \left[- \Sigma^{-1}\sigma_c^2(\boldsymbol{\theta} - \boldsymbol{\alpha}) + \sum_i \sum_j q_{ij}\, \mathbf{w}_{ij}(y_i - \mathbf{w}_{ij}'\boldsymbol{\theta}) \right] \qquad [5]$$

where :

$$q_{ij} = \frac{p_{ij}\, \phi\, [(y_i - \mathbf{w}_{ij}'\boldsymbol{\theta})/\sigma_e]}{\sum_j p_{ij}\, \phi\, [(y_i - \mathbf{w}_{ij}'\boldsymbol{\theta})/\sigma_e]} \qquad [6]$$

and ϕ (.) is the standard normal density function. Observe that q_{ij} is the posterior probability that progeny i is out of sire j, and that this probability is maximum when the residual $y_i - \mathbf{w}_{ij}'\boldsymbol{\theta}$ is null. This is so because in this instance the model under $\mathscr{L}_{ij}$ would fit perfectly to the data. Equating [5] to $\mathbf{0}$ gives a nonlinear system of equations on $\boldsymbol{\theta}$ so an iterative procedure is required to solve it.

Although several algorithms can be used for this purpose, the simple form of [5] suggests to implement a functional iteration. Setting [5] to $\mathbf{0}$ and rearranging yields :

$$\begin{bmatrix} \sum_i \mathbf{x}_i\mathbf{x}_i' & \sum_i \mathbf{x}_i \sum_j q_{ij}\mathbf{z}_{ij}' \\ \text{symm.} & \sum_i \sum_j q_{ij}\mathbf{z}_{ij}\mathbf{z}_{ij}' + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \sum_i \mathbf{x}_i y_i \\ \sum_i \sum_j q_{ij}\mathbf{z}_{ij} y_i \end{bmatrix} \qquad [7]$$

because prior information about $\boldsymbol{\beta}$ is vague and $\sum_j q_{ij} = 1$; $\lambda = \sigma_e^2/\sigma_u^2 = (4/h^2) - 1$, where h^2 is heritability. Note that the coefficient matrix and the right-hand sides depend on $\boldsymbol{\theta}$ as q_{ij} is a function of $\boldsymbol{\beta}$ and $\mathbf{u}$; this is clear from [6]. Defining :

$\mathbf{Q} = \{q_{ij}\}$: an n × m matrix of posterior probabilities,

and :

$\mathbf{D}_c = \text{Diag}\ \{\sum_i q_{ij}\}$: an m × m diagonal matrix, whose elements can be thought of as the posterior expected value of the number of progeny of sire j,

the above system can be written in terms of the iterative scheme :

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Q}^{[k-1]} \\ \mathbf{Q}'^{[k-1]}\mathbf{X} & \mathbf{D}_c^{[k-1]} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{[k]} \\ \mathbf{u}^{[k]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Q}'^{[k-1]}\mathbf{y} \end{bmatrix} \qquad [8]$$

where [k] indicates the iterate number. In [8], the matrices $\mathbf{Q}$ and $\mathbf{D}_c$ are evaluated at the « current » values of $\boldsymbol{\beta}$ and $\mathbf{u}$, through updating q_{ij} in [6].

One possible way of starting iteration is to take $q_{ij}^{[0]} = p_{ij}$ for all values of i and j. Thus $\mathbf{Q}^{[0]} = \mathbf{P} = \{p_{ij}\}$, and $\mathbf{D}_c^{[0]} = \mathbf{\Delta}_c = \text{Diag}\{\sum_i p_{ij}\}$, and these values can be viewed as the « natural » ones to adopt prior to the data.

In practice, uncertainty is only with respect to a small subset of the sires that need to be evaluated. The progeny can be classified into 2 groups : $\mathbf{I}_1$, pertaining to individuals having sires unambiguously identified, and $\mathbf{I}_2$ corresponding to progeny with parentage under « dispute ». Similarly, sires can be allocated to 2 groups : $\mathbf{J}_1$, with all their progeny in set $\mathbf{I}_1$, and $\mathbf{J}_2$, with some progeny in $\mathbf{I}_1$ and some progeny in $\mathbf{I}_2$. The data vector can be partitioned into three mutually exclusive and exhaustive components :

$$\begin{aligned} &\mathbf{y}_{11} \text{ if } \{i \in \mathbf{I}_1 \cap j \in \mathbf{J}_1\} \\ &\mathbf{y}_{12} \text{ if } \{i \in \mathbf{I}_1 \cap j \in \mathbf{J}_2\} \\ &\mathbf{y}_{22} \text{ if } \{i \in \mathbf{I}_2 \cap j \in \mathbf{J}_2\} \end{aligned} \qquad [9]$$

because the set $\{i \in \mathbf{I}_2 \cap j \in \mathbf{J}_1\}$ is empty. The vector of transmitting abilities can be partitioned as $[\mathbf{u}_1, \mathbf{u}_2]$, corresponding to sires in $\mathbf{J}_1$ and $\mathbf{J}_2$, respectively, so.

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}$$

Likewise

$$\mathbf{X}' = [\mathbf{X}_{11}', \mathbf{X}_{12}', \mathbf{X}_{22}']$$

correspond to the three partitions in [9] above. Further

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Z}_{11} & \mathbf{O} \\ \mathbf{O} & \mathbf{Z}_{12} \\ \mathbf{O} & \mathbf{Q}_{22} \end{bmatrix}, \mathbf{P} = \begin{bmatrix} \mathbf{Z}_{11} & \mathbf{O} \\ \mathbf{O} & \mathbf{Z}_{12} \\ \mathbf{O} & \mathbf{P}_{22} \end{bmatrix} \qquad [10]$$

with $\mathbf{Z}_{11} = \{p_{ij} = 0 \text{ or } 1\}$, $\mathbf{Z}_{12} = \{p_{ij} = 0 \text{ or } 1\}$, $\mathbf{Q}_{22} = \{0 < q_{ij} < 1\}$, $\mathbf{P}_{22} = \{0 < p_{ij} < 1\}$, as per the partitions in [9]. Using this notation, equations [8] become :

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_{11}'\mathbf{Z}_{11} & \mathbf{X}_{12}'\mathbf{Z}_{12} + \mathbf{X}_{22}'\mathbf{Q}_{22}^{[k-1]} \\ & \mathbf{Z}_{11}'\mathbf{Z}_{11} + \mathbf{A}^{11}\lambda & \mathbf{A}^{12}\lambda \\ \text{symm.} & & \mathbf{Z}_{12}'\mathbf{Z}_{12} + \mathbf{D}_{C_{22}}^{[k-1]} + \mathbf{A}^{22}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{[k]} \\ \mathbf{u}_1^{[k]} \\ \mathbf{u}_2^{[k]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_{11}'\mathbf{y}_{11} \\ \mathbf{Z}_{12}'\mathbf{y}_{12} + \mathbf{Q}_{22}'^{[k-1]}\mathbf{y}_{22} \end{bmatrix} \qquad [11]$$

where $\mathbf{D}_{C_{22}}$ is a diagonal matrix with elements calculated as before but for the progeny and sires in the third partition of [9]. Again, iteration can be started by replacing the « posterior » $\mathbf{Q}$ and $\mathbf{D}$ matrices in [11] by their « prior » counterparts, $\mathbf{P}$ and $\mathbf{\Delta}$, of appropriate order. The above equations illustrate clearly the modifications needed in the mixed model equations to take into account uncertain paternity. The portions in the coefficient matrix and right-hand sides pertaining to records where paternity is unambi-

guous ($\mathbf{y}_{11}$ and $\mathbf{y}_{12}$) are the usual ones. The incidence matrix $\mathbf{Z}_{22}$ that would arise if paternity of animals with records in $\mathbf{y}_{22}$ were certain, is replaced by a matrix $\mathbf{Q}$ of posterior probabilities. These are updated during the course of iteration to take into account the contribution of the data. Likewise, $\mathbf{Z}'_{22}\mathbf{Z}_{22}$ is replaced by the $\mathbf{D}$ matrix, which is a function of the posterior probabilities q_{ij}, as already indicated. Because $\mathbf{Q}_{22}$ is usually a small matrix, [8] or [11] will converge rapidly. If functional iteration is slow to converge, algorithms such as Newton-Raphson can be employed (Appendix B).

III. Binary data

A. *Methodology*

The data are now binary responses so $y_i = 0$ or 1. The model used here is based on the concept of « liability » originally developed by WRIGHT (1934), where it is assumed that there is an underlying normal variable rendered binary via an abrupt threshold. Genetic evaluation procedures based on threshold models have been discussed by several authors (GIANOLA & FOULLEY, 1983a,b ; FOULLEY *et al.*, 1983 ; FOULLEY & GIANOLA, 1984 ; HARVILLE & MEE, 1984 ; GILMOUR *et al.*, 1985 ; HÖSCHELE *et al.*, 1986).

The notation of the preceding section is retained, with the understanding that the parameters are now those of the underlying distribution. The conditional distribution of a binary response is taken as :

$$f(y_i|\mathcal{L}_{ij}, \boldsymbol{\theta}) \sim \text{Binomial } \{\Phi [(-1)^{y_i}\mu_{ij}]\} \qquad [12]$$

where $\Phi(.)$ is the standardized normal cumulative distribution function. The parameter μ_{ij} is the difference between the threshold and the mean of the statistical « subpopulation » defined by indexes i, j (GIANOLA & FOULLEY, 1983a) expressed in units of standard deviation. Assuming the prior distribution is as in [2] and replacing the normal density in [3B] by [12], the posterior density can be written as :

$$f(\boldsymbol{\theta}|\mathbf{y}, \sigma_u^2) \propto \{\prod_i \sum_j p_{ij} \Phi [(-1)^{y_i}\mathbf{w}'_{ij}\boldsymbol{\theta}]\} \cdot \exp[-(\boldsymbol{\theta} - \boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\alpha})/2] \qquad [13]$$

because the residual standard deviation is equal to 1.

Finding the $\boldsymbol{\theta}$ – mode of [13] involves solving a system with a higher order of nonlinearity than the one stemming from [5] so Newton-Raphson is used here instead of functional iteration as done in the previous section. The derivatives needed are :

$$\frac{\delta L(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}} = -\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\alpha}) + \sum_i \sum_j \pi_{ij}\mathbf{w}_{ij} \qquad [14]$$

$$\frac{-\delta^2 L(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}\delta \boldsymbol{\theta}'} = \boldsymbol{\Sigma}^{-1} + \sum_i \sum_j r_{ij}\mathbf{w}_{ij}\mathbf{w}'_{ij} \qquad [15]$$

where

$$\pi_{ij} = \frac{(-1)^{y_i}p_{ij}\,\phi\,(\mu_{ij})}{\sum_j p_{ij}\,\Phi\,[(-1)^{y_i}\mu_{ij}]} \qquad [16]$$

and

$$r_{ij} = \pi_{ij} (\pi_{ij} + \mu_{ij}) \qquad [17]$$

Letting

$$\mathbf{R} = \{r_{ij}\},\ n \times m$$

$$\mathbf{E}_r = \text{Diag}\ \{\sum_j r_{ij}\},\ n \times n$$

$$\mathbf{E}_c = \text{Diag}\ \{\sum_i r_{ij}\},\ m \times m$$

$$\mathbf{\Pi} = \{\pi_{ij}\},\ n \times m$$

the Newton-Raphson equations can be written after algebra as :

$$\begin{bmatrix} \mathbf{X}'\mathbf{E}_r^{[k-1]}\mathbf{X} & \mathbf{X}'\mathbf{R}^{[k-1]} \\ \mathbf{R}'^{[k-1]}\mathbf{X} & \mathbf{E}_c^{[k-1]} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\beta}^{[k]} \\ \Delta\mathbf{u}^{[k]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{\Pi}^{[k-1]}\mathbf{1}_m \\ \mathbf{\Pi}'^{[k-1]}\mathbf{1}_n - \mathbf{A}^{-1}\lambda\mathbf{u}^{[k-1]} \end{bmatrix} \qquad [18]$$

where the variance ratio $\lambda = 1/\sigma_u^2$ because the residual variance is unity, $\Delta\boldsymbol{\beta}^{[k]} = \boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}^{[k-1]}$, $\Delta\mathbf{u}^{[k]} = \mathbf{u}^{[k]} - \mathbf{u}^{[k-1]}$, and $\mathbf{1}_m$, $\mathbf{1}_n$ are vectors of ones of appropriate order. One possible way to start iteration would be to use equations [8] with $\mathbf{Q}$ replaced by $\mathbf{P}$, $\mathbf{D}_c$ replaced by $\mathbf{\Delta}_c$, and $\mathbf{y}$ replaced by a vector of 0 and 1's indicating the absence or presence of the attribute in the progeny in question. The values of $\boldsymbol{\beta}$ and $\mathbf{u}$ so obtained would be used to calculate π_{ij} and r_{ij} in [16] and [17] to then proceed iterating with [18] above.

B. *Analogy with the normal case*

Write π_{ij} in [16] as

$$\pi_{ij} = \left\{\frac{p_{ij}\Phi\left[(-1)^{y_i}\mu_{ij}\right]}{\sum_j p_{ij}\Phi[(-1)^{y_i}\mu_{ij}]}\right\} \cdot \left\{\frac{(-1)^{y_i}\phi(\mu_{ij})}{\Phi[(-1)^{y_i}\mu_{ij}]}\right\}$$

$$= \{q^*_{ij}\} \cdot \{v_{ij}\}$$

The expression q^*_{ij} is directly comparable to q_{ij} of [6] for the normal case. Both can be interpreted as the posterior probabilities that progeny i is out of sire j, and are similar to formulae arising in multivariate classification problems (LINDEMAN *et al.*, 1980, p. 196). In the discrete case and given $\mathcal{L}_{ij}$, if μ_{ij} is large progeny i would be expected to respond with high probability in the first category and q^*_{ij} will be larger when the response is actually in the first rather than in the second category. The expression for v_{ij} (with a minus sign) is the « normal score » discussed by GIANOLA & FOULLEY (1983a, p. 216 ; 1983b, p. 143).

IV. Estimation of unknown variances

The point estimators of location described above are the modes of posterior distributions of $\boldsymbol{\theta}$ conditionally on the variances σ_u^2 and σ_e^2 in the normal case, or to σ_u^2

in the situation of binary responses. When these variances are unknown, Box & TIAO (1973) and O'HAGAN (1976) have given arguments indicating that inferences could be made from the distribution $f(\boldsymbol{\theta}|\sigma_e^2 = \hat{\sigma}_e^2, \sigma_u^2 = \hat{\sigma}_u^2)$, where the variances are replaced by the modal values of the marginal posterior distribution of the variances. In the absence of prior information about the variances, these modal values are those obtained from the method of restricted maximum likelihood (HARVILLE, 1974, 1977). This approach was employed by GIANOLA *et al.* (1986) in the context of optimum prediction of breeding values and these authors view the resulting predictors as belonging to the class of empirical Bayes estimators. The general principles involved in finding the modal values of the posterior distribution of the variances are given below.

FOULLEY *et al.* (1986) and GIANOLA *et al.* (1986) showed that maximization of $f(\sigma_u^2, \sigma_u^2|\mathbf{y})$ with respect to the variances in the absence of prior information about these parameters leads to the equations :

$$E_c\left[\frac{\delta}{\delta\sigma_u^2} ln\ f(\mathbf{u}|\sigma_u^2)\right] = 0 \qquad [19]$$

where E_c indicates expectation with respect to the distribution $f(\mathbf{u}|\sigma_e^2, \sigma_u^2, \mathbf{y})$. Further, and now taking expectation with respect to $f(\boldsymbol{\theta}|\sigma_e^2, \sigma_u^2, \mathbf{y})$, we need to satisfy :

$$E_c\left[\frac{\delta}{\delta\sigma_e^2} ln\ f(\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2)\right] = 0 \qquad [20]$$

The derivation is based on the decomposition of the posterior distribution of all unknowns, $f(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_e^2|\mathbf{y})$, into

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_e^2) \cdot f(\mathbf{u}|\sigma_u^2) \cdot f(\boldsymbol{\beta}) \cdot f(\sigma_u^2) \cdot f(\sigma_e^2)$$

It should be noted that the likelihood function does not depend on σ_u^2, which is true both in the normal and binary cases. Also, when flat priors are taken for the variances, $f(\sigma_u^2)$ and $f(\sigma_e^2)$ do not appear in the above decomposition.

Solving [19] and [20] simultaneously for the unknown variances leads to an iterative scheme involving the expressions :

$$\sigma_u^{2[k+1]} = \{[\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}]^{[k]} + \sigma_e^{2[k]} tr(\mathbf{A}^{-1}\mathbf{C}_{uu}^{[k]})\}/m \qquad [21]$$

and

$$\sigma_e^{2[k+1]} = \frac{\mathbf{y}'\mathbf{y} - [\boldsymbol{\theta}'\mathbf{W}'\mathbf{y}]^{[k]} - [\lambda\ \mathbf{u}'\mathbf{A}^{-1}\ \mathbf{u}]^{[k]}}{n - tr\ (\mathbf{C}^{[k]}\mathbf{M}^{[k]})} \qquad [22]$$

where

- k is iterate number,
- **C** is the inverse of the coefficient matrix in Newton-Raphson (Appendix B), or of [18] when observations are binary,
- $\mathbf{C}_{uu}$ is the submatrix of **C** corresponding to the u-effects,
- **M** is the coefficient matrix in [8] or [18] without $\mathbf{A}^{-1}\lambda$,
- $\mathbf{W} = [\mathbf{X}, \mathbf{Q}]$

It should be noted that in the binary case the residual variance is not estimated because it is taken as equal to one. The derivation of [22] is given in Appendix C. Equation

[21], however, holds in both cases. The conditional expectations are taken as if the « true » values of the variance components were those found in the previous iteration. As pointed out by GIANOLA *et al.* (1986), [20] and [21] arise in the EM algorithm (DEMPSTER *et al.*, 1977) when applied to estimation by restricted maximum likelihood, and the resulting estimates are never negative.

V. Numerical application

A small data set from a progeny test of *Blonde d'Aquitaine* sires carried out in France was used to illustrate the methods presented in this paper. The data set is the same as the one utilized by FOULLEY *et al.* (1983), with some modifications, as illustrated in table 1. There were 47 calving records including information on region of origin of the heifer, calving season, sex and sire of calf, and birth weight (BW) and calving ease (CE) as response variables. CE was recorded as an all-or-none trait with « easy » and « difficult » calvings coded as 0 or 1, respectively. As shown in table 1, paternity was uncertain in the case of records 1, 2, 3 and 39. For the first three records, information on breeding periods and gestation lengths led to an assignment to natural service sires 7 and 8 of probabilities equal to $\frac{1}{4}$ and $\frac{3}{4}$, respectively. In the case of record 39, artificial insemination sires 1 and 2 were assigned probabilities of $\frac{1}{2}$ and $\frac{1}{2}$, respectively.

A. *Model*

Birth weight was regarded as following a normal distribution, and CE was treated as a binomial trait. Both traits were analyzed using the model

$$y_{ijk/m} = H_i + A_j + S_k + f_l + e_{ijk/m} \quad [23]$$

where H_i is the effect of region i of origin of heifer (i = 1, 2), A_j is the effect of the *jth* season of calving (j = 1, 2), S_k is the effect of sex of calf k (k = 1 for males or 2 for females), f_l is the transmitting ability of the *lth* sire of heifer (l = 1, ..., 8), and $e_{ijk/m}$ is a residual with variance σ_e^2. The vectors $\boldsymbol{\beta}$ and $\mathbf{u}$ were

$$\boldsymbol{\beta}' = (H_1 + A_2 + S_2,\ H_2 + A_2 + S_2,\ A_1 - A_2,\ S_1 - S_2) \quad [24]$$

and

$$\mathbf{u}' = (f_1,\ f_2,\ \ldots,\ f_8) \quad [25]$$

Prior knowledge about $\boldsymbol{\beta}$ was assumed to be vague. Heritability was .25 for both traits, and σ_e^2 was 5 kg for BW and 1 for CE, the discrete trait. In forming the relationship matrix **A**, it was assumed that the artificial insemination sires (1 through 6) were unrelated, and that the natural service sires 7 and 8 were non-inbred sons of 5 and 4, respectively.

TABLE 1

Description of records from Blonde d'Aquitaine *heifers used in the calculations.*

Record no.	Heifer origin	Calving season	Sex [a] of calf	Sire	BW [b]	CD [c]	Record no.	Heifer origin	Calving season	Sex [a] of calf	Sire	BW [b]	CD [c]
1 [d]	1	1	M	7(1/4) ; 8(3/4)	41.0	E	24	1	2	M	4	47.0	D
2 [d]	1	1	M	7(1/4) ; 8(3/4)	37.5	E	25	1	2	F	4	51.0	D
3 [d]	1	1	F	7(1/4) ; 8(3/4)	41.5	E	26	1	2	F	4	39.0	E
4	1	2	F	1	40.0	E	27	2	1	M	4	44.5	E
5	1	2	F	1	43.0	E	28 [e]	1	1	M	5	40.5	D
6	1	2	F	1	42.0	E	29	1	1	F	5	43.5	E
7	1	2	F	1	35.0	E	30 [e]	1	2	M	5	42.5	D
8	2	1	F	1	46.0	E	31	1	2	M	5	48.8	D
9	2	1	F	1	40.5	E	32	1	2	M	5	38.5	E
10	2	2	F	1	39.5	E	33	1	2	M	5	52.0	E
11	1	1	M	2	41.4	E	34	1	2	F	5	48.0	E
12	1	1	M	2	43.0	D	35	2	1	F	5	41.0	E
13	1	2	F	2	34.0	E	36	2	1	M	5	50.5	D
14	1	2	M	2	47.0	D	37	2	2	M	5	43.7	D
15	1	2	M	2	42.0	E	38	2	2	M	5	51.0	D
16	2	2	M	2	44.5	E	39 [f]	1	1	F	1(1/2) ; 6(1/2)	51.6	D
17	2	2	M	2	49.0	E	40	1	1	M	6	45.3	D
18	1	1	M	3	41.6	E	41	1	1	F	6	36.5	E
19	2	1	M	3	36.0	E	42	1	2	M	6	50.5	E
20	2	1	F	3	42.7	E	43	1	2	M	6	46.0	D
21	2	2	F	3	32.5	E	44	1	2	M	6	45.0	E
22	2	2	F	3	44.4	E	45	1	2	F	6	36.0	E
23	2	2	M	3	46.0	E	46	2	1	F	6	43.5	E
							47	2	1	F	6	36.5	E

(a) M : male ; F : female.
(b) BW : birth weight (kg).
(c) CD : calving difficulty ; E : easy, D : difficult.
(d) assigned to sire 1 in Foulley *et al.* (1983).
(e) classified as E in Foulley *et al.* (1983).
(f) assigned to sire 6 in Foulley *et al.* (1983).

In this example, the sets needed to define [9] were : $\mathbf{I}_2 = \{1, 2, 3, 39\}$ (with $\mathbf{I}_1$ being the complement), $\mathbf{J}_1 = \{2, 3, 4, 5\}$ and $\mathbf{J}_2 = \{1, 6, 7, 8\}$. Thus, the matrix $\mathbf{P}_{22}$ in [10] was

$$\mathbf{P}_{22} = \begin{bmatrix} 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

For BW, the nonlinear system [11] was solved using 3 algorithms : functional iteration, and Newton-Raphson and scoring as described in Appendix B. For CE, computations were carried out with [18] ; starting values were calculated as discussed earlier. Iteration stopped when the square root of the average squared correction was less than 10^{-5}. Variance components were estimated for both traits using the procedures outlined in section **III**.

Sire evaluations ignoring uncertainty on paternity were also calculated so as to further illustrate the procedures. This was done by assigning progenies 1, 2, 3 to sire 1 and record 39 to sire 6.

B. *Results*

Results of the analysis conducted for BW are presented in table 2. Irrespective of the algorithm used, the stopping rule of 10^{-5} was satisfied in 4 iterations. The fact that the algorithms were equally fast to converge is undoubtedly related to the limited extent of nonlinearity, as only 4 out of 39 records had ambiguous parentage. Further, a « sharp » assignment of probabilities $\left(\frac{1}{4} \text{vs. } \frac{3}{4}\right)$ was made in 3 out of the 4 records. In this data set, from a practical point of view iteration could have stopped at the second round. Differences between the analyses conducted ignoring uncertainty and taking it into account were minimal. Sire 1 was the most affected because 3 records assigned to him in the case of certain paternity were assigned to sires 7 or 8 when paternity was uncertain.

The analysis of calving ease is shown in table 3. When paternity was certain, 5 iterates were required to converge. On the other hand, 13 iterations were required when uncertainty was taken into account. This is so because with a binary trait there are 2 sources of nonlinearity when paternity is uncertain : one due to the fact that the model is nonlinear, and the second due to the uncertainty itself. The second source of nonlinearity was responsible for the 8 additional iterations.

Estimates of variance components in this example were $\sigma_u^2 \simeq 0$ and $\sigma_e^2 = 22.81$ for BW, and $\sigma_u^2 = .096$ for CE. The latter value gives an estimate of heritability of .35 in the underlying scale. For BW, more than 400 iterates were needed for estimates of variance components to converge, and 193 iterations were required for CE. It is well known that the EM algorithm is extremely slow to converge (THOMPSON, 1979), especially in small samples. However, alternative parameterizations of the model or numerical shortcuts (e.g., SCHAEFFER, 1979 ; MISZTAL & SCHAEFFER, 1986) can be used to reduce the computational burden.

TABLE 2

Estimates of parameters and of their posterior precision : birthweight.

Parameters	Certain paternity [a]	Uncertain paternity [a,b]	
		First iterate	Final solution
$H_1 + A_2 + S_2$	41.598 ± 1.493	41.455	41.456 ± 1.528
$H_2 + A_2 + S_2$	42.341 ± 1.719	42.203	42.205 ± 1.758
$A_1 - A_2$	− 1.269 ± 1.506	− 1.273	− 1.274 ± 1.618
$S_1 - S_2$	3.144 ± 1.528	3.294	3.293 ± 1.602
u_1	− 0.486 ± 1.086	0.080	0.076 ± 1.151
u_2	− 0.368 ± 1.117	− 0.364	− 0.364 ± 1.118
u_3	− 0.749 ± 1.141	− 0.730	− 0.730 ± 1.140
u_4	0.492 ± 1.165	0.365	0.367 ± 1.160
u_5	0.745 ± 1.061	0.725	0.723 ± 1.062
u_6	0.367 ± 1.085	0.162	0.166 ± 1.104
u_7	0.372 ± 1.238	0.273	0.265 ± 1.227
u_8	0.246 ± 1.261	− 0.085	− 0.080 ± 1.208

(a) $h^2 = .25$ ($\lambda = 15$) and $\sigma_e^2 = 25$ were used. Convergence reached after 4 rounds of iteration for $\epsilon = 10^{-5}$. Estimates of the variances in this data set are : $\hat{\sigma}_e^2 = 22.8133$ and $\hat{\sigma}_u^2 = 5 \times 10^{-6}$.
(b) Posterior standard errors calculated from the Newton-Raphson algorithm.

TABLE 3

Estimates of parameters and of their posterior precision : calving ease.

Parameters	Certain paternity [a]	σ_u^2 known [b]	σ_u^2 estimated [c]
$H_1 + A_2 + S_2$	1.181 ± .463	1.196 ± .488	1.210 ± .495
$H_2 + A_2 + S_2$	1.692 ± .592	1.702 ± .598	1.715 ± .608
$A_1 - A_2$	.008 ± .441	.024 ± .479	0.012 ± .481
$S_1 - S_2$	− 1.152 ± .478	− 1.172 ± .522	− 1.167 ± .526
u_1	.164 ± .241	.020 ± .250	0.027 ± .297
u_2	.059 ± .237	.059 ± .237	0.075 ± .276
u_3	.120 ± .246	.119 ± .246	.162 ± .290
u_4	− .103 ± .243	− .066 ± .243	− .094 ± .285
u_5	− .182 ± .230	− .172 ± .230	− .225 ± .266
u_6	− .057 ± .235	− .018 ± .239	− .027 ± .279
u_7	− .091 ± .251	− .064 ± .251	− 0.081 ± .299
u_8	− .051 ± .255	.032 ± .249	.046 ± .295

(a) $\lambda = 15$ ($h^2 = .25$). Convergence reached after 5 rounds of iteration for $\epsilon = 10^{-5}$.
(b) $\lambda = 15$. Convergence reached after 13 rounds of iteration for $\epsilon = 10^{-5}$.
(c) $\hat{\sigma}_u^2 = .096$ ($\hat{h}^2 = .35$) ; starting with $\sigma_u^2 = 1/15$. Convergence reached after 193 iterations for $\epsilon = 10^{-15}$.

VI. Discussion

The impact of the extent of misidentification on sire evaluation and on estimates of genetic parameters was studied by VAN VLECK (1970a,b) and BONAITI (1975). These authors found that misidentification of sires biased downwards estimates of heritability and of expected genetic progress. Biases in evaluation of sires increased as the fraction of misidentified animals increased.

The approach followed in the present study, as in POIVEY & ELSEN (1984), is to directly take into account in the analysis uncertainty on the assignment of progeny to sires so as to improve prediction of breeding values. However, POIVEY & ELSEN (1984) studied the problem in a selection index framework which requires knowledge of means and variances. The issue was adressed here in a more general manner so as to accommodate different types of distribution (normal or binomial), and less restrictive states of knowledge vis-a-vis fixed effects and variance components.

Using the Bayesian paradigm as in GIANOLA *et al.* (1986), leads to inferences based on a posterior distribution with the uncertainty « integrated » or averaged out. With the $\boldsymbol{\beta}$ and $\mathbf{u}$ components of the mode of the posterior distribution taken as point estimators and predictors of fixed and random effects, respectively, a nonlinear system of equations is obtained. Algorithms for solving these equations are discussed in the paper. Variance components were estimated from the joint posterior distribution of the variances after taking into account uncertainty in the assignment of progeny to sires. The point estimators chosen were the modal values of this distribution ; expressions for computing the estimates iteratively were presented.

Is it possible to use directly the mixed model equations to obtain standard best linear unbiased predictors when paternity is uncertain ? The best linear unbiased predictor of $\mathbf{u}$ is given by

$$\hat{\mathbf{u}} = \text{Cov}(\mathbf{u}, \mathbf{y}')\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

where $\mathbf{V}$ is the variance covariance matrix of the records (HENDERSON, 1973). From results in Appendix A, the diagonal elements of $\mathbf{V}$ in the continuous case are

$$v_{ii} = \sigma_u^2 + \sigma_e^2$$

and the off-diagonals are

$$v_{ii'} = (\sum_j p_{ij}\, p_{i'j} + \sum_{j \neq j'} p_{ij} p_{i'j'} a_{jj'})\sigma_u^2$$

where $a_{jj'}$ is the additive relationship between sires j and j'. It follows that $\mathbf{V}$ is not in the form $\mathbf{ZGZ}' + \mathbf{R}$ needed to put $\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$ so as to establish the equivalence between the best linear unbiased predictor above and the results given by the mixed model equations (HENDERSON, 1984). It is not obvious how to treat the problem of uncertain paternity using standard techniques. The Bayesian solution presented here, on the other hand, offers a clear answer. POIVEY & ELSEN (1984) discussed situations in which the methods presented here could be applied. These include : i) females exposed simultaneously or successively in time to groups of males ; ii) joint use of artificial and natural breeding in sheep flocks and cattle herds in conjunction with estrous synchronization techniques ; and iii) heterospermic progeny testing with ambiguous parentage. A requirement of the procedure is the specification of prior probabili-

ties p_{ij} which can be based on external information such as biochemical polymorphysms or, more likely under extensive conditions, records on breeding dates and gestation lengths. In particular, the methods described here may be potentially useful in situations where natural service sires are used extensively, e.g., pastoral production systems. The computations are feasible, at least for univariate sire evaluations carried out under the assumption of normality and with genetic parameters assumed known. Extensions to the multivariate situation can be done without great conceptual difficulty.

Received February 27, 1986.
Accepted June 24, 1986.

Acknowledgements

This research was conducted while J.L. FOULLEY was a George A. Miller Visiting Scholar at the University of Illinois, on sabbatical leave from INRA. J.L. FOULLEY wishes to thank the Direction des Productions animales and Direction des relations internationales, INRA, for their support. D. GIANOLA acknowledges support from the Illinois Agriculture Experiment Station and of Grant No. US-805-84 from BARD-The United States-Israel Binational Agricultural Research and Development Fund.

References

BONAITI B., 1975. Influence du taux d'erreur dans les filiations paternelles sur la valeur de la sélection sur descendance chez les ovins. *1ères Journées de la Recherche Ovine et Caprine, Paris, 2-4 décembre 1975, SPEOC-ITOVIC*, 166-172.

BOX G.E.P., TIAO G.C., 1983. *Bayesian inference in statistical analysis*. 588 pp., Addison-Wesley, Reading, Massachussets.

COX D.R., HINKLEY D.V., 1974. *Theoretical statistics*. 511 pp., Chapman and Hall, London.

DEMPFLE L., 1977. Relation entre BLUP et estimateurs bayesiens. *Ann. Génét. Sél. Anim.*, **9**, 27-32.

DEMPSTER A.P., LAIRD N.M., RUBIN D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statis. Soc.*, B, **39**, 1-38.

FERNANDO R.L., GIANOLA D., 1986. Optimal properties of the conditional mean as a selection criterion. *Theor. Appl. Genet.* (submitted).

FOULLEY J.L., GIANOLA D., THOMPSON R., 1983. Prediction of genetic merit from data on binary and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. *Génét. Sél. Evol.*, **15**, 401-424.

FOULLEY J.L., GIANOLA D., 1984. Estimation of genetic merit from bivariate « all or none » responses. *Génét. Sél. Evol.*, **16**, 285-306.

FOULLEY J.L., IM S., GIANOLA D., HÖSCHELE I., 1987. Empirical Bayes estimation of parameters for n polygenic binary traits. *Génét. Sél. Evol.*, **19** (in press).

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

GIANOLA D., FOULLEY J.L., 1983b. New techniques of prediction of breeding value for discontinuous traits. *Proc. 32nd Annual National Breeders Round-table, St. Louis, Missouri, May 6, 1983*, 28 pp., Mimeo.

GIANOLA D., FERNANDO R.L., 1986. Bayesian methods in animal breeding theory. *J. Anim. Sci.*, **63**, 217-244.

GIANOLA D., FOULLEY J.L., FERNANDO R.L., 1986. Prediction of breeding values when variances are not known. *Génét. Sél. Evol.* (submitted).

GILMOUR A.R., ANDERSON R.D., RAE A.L., 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72**, 593-599.

GOFFINET B., ELSEN J.M., 1984. Critère optimal de sélection : quelques résultats généraux. *Génét. Sél. Evol.*, **16**, 307-318.

HARVILLE D.A., 1974. Bayesian inference for variance components using only error contrasts. *Biometrika*, **72**, 593-599.

HARVILLE D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.*, **72**, 320-338.

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

HENDERSON C.R., 1973. Sire evaluation and genetic trends. *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush.* American Society of Animal Science and American Dairy Science Association, 10-41, Champaign, Illinois.

HENDERSON C.R., 1984. *Applications of linear models in animal breeding.* 462 pp., University of Guelph, Guelph, Ontario.

HÖSCHELE Ina, FOULLEY J.L., COLLEAU J.J., GIANOLA D., 1986. Genetic evaluation for multiple binary responses. *Génét. Sél. Evol.*, **18**, 299-320.

LEFORT G., 1980. Le modèle de base de la sélection, justification et limites. *In* Legay J.M. *et al.* (éd.), *Biométrie et Génétique*, **4**, 1-14, Société Française de Biométrie. INRA, Département de Biométrie.

LINDEMAN R.H., MERENDA P.F., GOLD R.Z., 1980. *Introduction to bivariate and multivariate analysis.* 444 pp., Scott, Foresman & Co., Glenview, Illinois.

MISZTAL I., SCHAEFFER L.R., 1986. Nonlinear model for describing convergence of iterative methods of variance component estimation. *J. Dairy Sci.*, **69**, 2209-2213.

O'HAGAN A., 1976. On posterior joint and marginal modes. *Biometrika*, **63**, 329-333.

POIVEY J.P., ELSEN J.M., 1984. Estimation de la valeur génétique des reproducteurs dans le cas d'incertitude sur les apparentements. I. Formulation des indices de sélection. *Génét. Sél. Evol.*, **16**, 445-454.

RÖNNINGEN K., 1971. Some properties of the selection index derived by « Henderson's mixed model method ». *Z. Tierz. Züchtungsbiol.*, **88**, 186-193.

SCHAEFFER L.R., 1979. Estimation of variance and covariance components for average daily gain and backfat thickness in swine. *In* : Searle S.R., Van Vleck L.D. (ed.), *« Variance components and animal breeding ». Proc. of a Conference in Honor of C.R. Henderson*, 123-137. Cornell University, Ithaca, New York.

THOMPSON R., 1979. Sire evaluation. *Biometrics*, **35**, 339-353.

VAN VLECK L.D., 1970a. Misidentification in estimating the paternal sib correlation. *J. Dairy Sci.*, **53**, 1469-1475.

VAN VLECK L.D., 1970b. Misidentification and sire evaluation. *J. Dairy Sci.*, **53**, 1697-1702.

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

ZELLNER A., 1971. *An introduction to Bayesian inference in econometrics*, 431 pp., John Wiley & Sons, New York.

Appendix A

Variance-covariance structure of the data (frequentist viewpoint)

The starting point is [3B]. As shown in the text, omitting the conditioning on σ_e^2 for the sake of simplicity :

$$E(y_i|\boldsymbol{\theta}) = \mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{p}_i'\mathbf{u} \qquad [A1]$$

The variance of the distribution can be obtained by writing :

$$\begin{aligned} Var(y_i|\boldsymbol{\theta}) &= \underset{\mathscr{L}_{ij}}{E}\,[Var(y_i|\boldsymbol{\theta}, \mathscr{L}_{ij})] + \underset{\mathscr{L}_{ij}}{Var}\,[E(y_i|\boldsymbol{\theta}, \mathscr{L}_{ij}] \\ &= \sigma_e^2 + \underset{\mathscr{L}_{ij}}{E}\,(\mu_{ij} - \mu_i)^2 = \sigma_e^2 + \sum_j p_{ij}\,(\mu_{ij} - \mu_i)^2 \end{aligned} \qquad [A2]$$

which follows from [1] and [A1]. Likewise

$$Cov(y_{i'}\, y_{i'}|\boldsymbol{\theta}) = \underset{\mathscr{L}_{ij}}{E}\ \underset{\mathscr{L}_{i'k}}{E}\,[Cov(y_{i'}\, y_{i'}|\boldsymbol{\theta}, \mathscr{L}_{ij}, \mathscr{L}_{i'k})] + Cov(\mu_{ij}, \mu_{i'k})$$

where the covariance is taken with respect to the joint distribution of $\mathscr{L}_{ij}$ and $\mathscr{L}_{i'k}$. The first term in the above equations is clearly null because the observations are conditionally independent. Assuming $P(\mathscr{L}_{ij} \cap \mathscr{L}_{i'k}) = p_{ij} \cdot p_{i'k}$, we get

$$\begin{aligned} Cov(y_{i'}\, y_{i'}|\boldsymbol{\theta}) &= \sum_j \sum_k p_{ij} p_{i'k}\, \mu_{ij}\mu_{i'k} - \mu_i\mu_{i'} \\ &= 0 \end{aligned} \qquad [A3]$$

We consider now the variance-covariance structure unconditionally on $\boldsymbol{\theta}$. Applying the same strategy, one can write :

$$Var(y_i) = \underset{\boldsymbol{\theta}}{E}\,[Var(y_i|\boldsymbol{\theta})] + \underset{\boldsymbol{\theta}}{Var}\,[E(y_i|\boldsymbol{\theta})] \qquad [A4]$$

The second term is the variance of μ_i taken with respect to the distribution of $\boldsymbol{\theta}$. Arguing from a classical viewpoint ($\boldsymbol{\beta}$ fixed, $\mathbf{u}$ random) we have :

$$\underset{\boldsymbol{\theta}}{Var}\,(\mu_i) = \mathbf{p}_i'\,\mathbf{A}\,\mathbf{p}_i\sigma_u^2 \qquad [A5]$$

From [A2]

$$\begin{aligned} \underset{\boldsymbol{\theta}}{E}\,[Var(y_i|\boldsymbol{\theta})] &= \sigma_e^2 + \sum_j p_{ij}\, \underset{\boldsymbol{\theta}}{E}(\mu_{ij} - \mu_i)^2 \\ &= \sigma_e^2 + \sum_j p_{ij}\,(\mathbf{z}_{ij} - \mathbf{p}_i)'\mathbf{A}\,(\mathbf{z}_{ij} - \mathbf{p}_i)\,\sigma_u^2 \\ &= \sigma_e^2 + (1 - \mathbf{p}_i'\,\mathbf{A}\,\mathbf{p}_i)\,\sigma_u^2 \end{aligned} \qquad [A6]$$

because $\mathbf{Z}_{ij}'\,\mathbf{A}\,\mathbf{Z}_{ij} = 1$ (if sires are not inbred), and $\Sigma p_{ij}\mathbf{z}_{ij} = \mathbf{p}_i$. Collecting [A5] and [A6] into [A4] gives

$$Var(y_i) = \sigma_e^2 + \sigma_u^2 \qquad [A7]$$

Finally, we consider the unconditional covariances between records y_i and $y_{i'}$. Writing :

$$\text{Cov}(y_i, y_{i'}) = \underset{\theta}{E}\,[\text{Cov}(y_i, y_{i'}|\theta] + \underset{\theta}{\text{Cov}}\,[E(y_i|\theta), E(y_{i'}|\theta)]$$

we observe as before that the first term is null. Also, from [A1] :

$$\underset{\theta}{\text{Cov}}\,[E(y_i|\theta), E(y_{i'}|\theta)] = \underset{\theta}{\text{Cov}}\,[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{p}_i'\mathbf{u}, \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{p}_{i'}', \mathbf{u}]$$

$$= \mathbf{p}_i' \mathbf{A}\, \mathbf{p}_{i'}\, \sigma_u^2 \qquad \text{[A8]}$$

It should be observed that Var(**y**) cannot be written as $\mathbf{R}\sigma_e^2 + \mathbf{PAP}'\sigma_u^2$ because the diagonal elements of this last matrix expression are not equal to [A7] except in the trivial case $\mathbf{P} = \mathbf{Z}$, i.e., when paternity is certain.

Appendix B

Newton-Raphson and scoring algorithms for normal data

The Newton-Raphson algorithm consists in iterating with :

$$-\left\{\frac{\delta^2 L(\boldsymbol{\theta})}{\delta\boldsymbol{\theta}\,\delta\boldsymbol{\theta}'}\right\}_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[k-1]}} \boldsymbol{\Delta}^{[k-1]} = \left\{\frac{\delta L(\boldsymbol{\theta})}{\delta\boldsymbol{\theta}}\right\}_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[k-1]}} \qquad \text{[B1]}$$

where $\boldsymbol{\Delta}^{[k]} = \boldsymbol{\theta}^{[k]} - \boldsymbol{\theta}^{[k-1]}$ and $\boldsymbol{\theta}^{[k]}$ is the solution at iteration k. The first derivatives are given in [5] and the second derivatives are :

$$\frac{-\delta^2 L(\boldsymbol{\theta})}{\delta\boldsymbol{\theta}\,\delta\boldsymbol{\theta}'} = (\sigma_e^2)^{-1}\left\{\boldsymbol{\Sigma}^{-1}\sigma_e^2 + \sum_i\sum_j q_{ij}\mathbf{w}_{ij}\mathbf{w}_{ij}' - \sum_i\sum_j \mathbf{w}_{ij}\left[\frac{\delta q_{ij}}{\delta\mu_{ij}}(y_i - \mathbf{w}_{ij}')\right]\mathbf{w}_{ij}'\right\} \qquad \text{[B2]}$$

From the definition of q_{ij} in [6], we have :

$$\frac{\delta q_{ij}}{\delta\mu_{ij}} = q_{ij}(1 - q_{ij})\,(y_i - \mu_{ij})/\sigma_e^2$$

Using this in [B2] above and rearranging gives :

$$\frac{-\delta^2 L(\boldsymbol{\theta})}{\delta\boldsymbol{\theta}\,\delta\boldsymbol{\theta}'} = (\sigma_e^2)^{-1}\left\{\boldsymbol{\Sigma}^{-1}\sigma_e^2 + \sum_i\sum_j r_{ij}\mathbf{w}_{ij}\mathbf{w}'_{ij}\right\} \qquad \text{[B3]}$$

where :

$$r_{ij} = q_{ij} - q_{ij}\,(1 - q_{ij})\,(y_i - \mu_{ij})^2/\sigma_e^2 \qquad \text{[B4]}$$

Using [B3] and [5] from the text in [B1] yields, after rearrangement :

$$\left\{\sum_i\sum_j r_{ij}\mathbf{w}_{ij}\mathbf{w}_{ij}' + \boldsymbol{\Sigma}^{-1}\sigma_e^2\right\}\boldsymbol{\theta}^{[k]} = \sum_i\sum_j \mathbf{w}_{ij}q_{ij}y_i - \sum_i\sum_j (q_{ij} - r_{ij})\,\mathbf{w}_{ij}\mathbf{w}_{ij}'\boldsymbol{\theta}^{[k-1]} \qquad \text{[B5]}$$

Some simplification in the calculations can be achieved by replacing r_{ij} by its expectation taken conditionally on $\boldsymbol{\theta}$, σ_e^2 and $\mathcal{L}_{ij}$. From [B4] we obtain directly :

$$E(r_{ij}|\boldsymbol{\theta}, \sigma_e^2, \mathcal{L}_{ij}) = q_{ij}^2$$

This used in [B5] above yields a « scoring » algorithm for solving the nonlinear system of equations.

The system [B5] can be written in matrix notation using the matrices $\mathbf{R}$, $\mathbf{E}_r$, and $\mathbf{E}_c$ of page 89 with r_{ij} as in [B4] instead of [17]. Also, define matrices

$$\mathbf{S} = \{q_{ij} - r_{ij}\} \qquad i = 1, \ldots, n \; ; \; j = 1, \ldots, m$$

$$\mathbf{F}_r = \text{Diag}\,\{\sum_j (q_{ij} - r_{ij})\} \quad i = 1, \ldots, n$$

$$\mathbf{F}_c = \text{Diag}\,\{\sum_i (q_{ij} - r_{ij})\} \quad j = 1, \ldots, m$$

The system [B5] becomes then

$$\begin{bmatrix} \mathbf{X}'\mathbf{E}_r^{[k-1]}\mathbf{X} & \mathbf{X}'\mathbf{R}^{[k-1]} \\ \mathbf{R}'^{[k-1]}\mathbf{X} & \mathbf{E}_c^{[k-1]} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{[k]} \\ \mathbf{u}^{[k]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Q}'^{[k-1]}\mathbf{y} \end{bmatrix} -$$

$$\begin{bmatrix} \mathbf{X}'\,(\mathbf{F}_r^{[k-1]}\mathbf{X}\boldsymbol{\beta}^{[k-1]} + \mathbf{S}^{[k-1]}\mathbf{u}^{[k-1]}) \\ \mathbf{S}'^{[k-1]}\mathbf{X}\boldsymbol{\beta}^{[k-1]} + \mathbf{F}_c^{[k-1]}\mathbf{u}^{[k-1]} \end{bmatrix} \qquad [B6]$$

The algorithm is also described for the case where uncertain paternity is only with respect to a small proportion of the sires evaluated. Here, we partition $\mathbf{R}$ in the same way as $\mathbf{Q}$ in [10], except that $\mathbf{Q}_{22}$ is replaced by $\mathbf{R}_{22}$. Also, put

$$\mathbf{S} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix} ;$$

$$\mathbf{E}_c = \begin{bmatrix} \mathbf{Z}_{11}'\mathbf{Z}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{12}'\mathbf{Z}_{12} + \mathbf{E}_{c_{22}} \end{bmatrix}, \; \mathbf{E}_{c_{22}} = \text{Diag}\,\{\sum_i r_{ij}\} \; (i \in \mathbf{I}_2, \, j \in \mathbf{J}_2) \; ;$$

$$\mathbf{E}_r = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_{r_{22}} \end{bmatrix} \qquad , \; \mathbf{E}_{r_{22}} = \text{Diag}\,\{\sum_j r_{ij}\} \; (i \in \mathbf{I}_2, \, j \in \mathbf{J}_2) \; ;$$

$$\mathbf{F}_c = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{c_{22}} \end{bmatrix} \quad , \ \mathbf{F}_{c_{22}} = \text{Diag } \{\sum_i s_{ij}\} \ (i \in I_2, \ j \in J_2) \ ;$$

$$\mathbf{F}_r = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{r_{22}} \end{bmatrix} \quad , \ \mathbf{F}_{r_{22}} = \text{Diag } \{\sum_j s_{ij}\} \ (i \in I_2, \ j \in J_2) \ ;$$

With the above notation, the system in B6 can be written as :

$$\begin{bmatrix} \mathbf{X}_{11}'\mathbf{X}_{11} + \mathbf{X}_{12}'\mathbf{X}_{12} + \mathbf{X}_{22}'\mathbf{E}_{r_{22}}^{[k-1]}\mathbf{X}_{22} & \mathbf{X}_{11}'\mathbf{Z}_{11} & \mathbf{X}_{12}'\mathbf{Z}_{12} + \mathbf{X}_{22}'\mathbf{R}_{22}^{[k-1]} \\ & \mathbf{Z}_{11}'\mathbf{Z}_{11} + \mathbf{A}^{11}\lambda & \mathbf{A}^{12}\lambda \\ \text{Symmetric} & & \mathbf{Z}_{12}'\mathbf{Z}_{12} + \mathbf{E}_{c_{22}}^{[k-1]} + \mathbf{A}^{22}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{[k]} \\ \mathbf{u}_1^{[k]} \\ \mathbf{u}_2^{[k]} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_{11}'\mathbf{y}_{11} \\ \mathbf{Z}_{12}'\mathbf{y}_{12} + \mathbf{Q}'^{[k-1]}\mathbf{y}_{22} \end{bmatrix} - \begin{bmatrix} \mathbf{X}'\mathbf{F}_{r_{22}}^{[k-1]}\mathbf{X}_{22} & \mathbf{0} & \mathbf{X}_{22}'\mathbf{S}_{22}^{[k-1]} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{S}_{22}'^{[k-1]}\mathbf{X}_{22} & \mathbf{0} & \mathbf{F}_{c_{22}}^{[k-1]} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{[k-1]} \\ \mathbf{u}_1^{[k-1]} \\ \mathbf{u}_2^{[k-1]} \end{bmatrix} \qquad [B7]$$

The above equations indicate the parts of the system that need to be amended to take uncertain paternity into account. As before, the nonlinearity stems from the contribution of the vector $\mathbf{y}_{22}$ to information about the unknown parameters. In order to start iteration, one may take $\mathbf{R}^{[0]} = \mathbf{P}$, $\mathbf{E}_r^{[0]} = \boldsymbol{\Delta}_r$, $\mathbf{E}_c^{[0]} = \boldsymbol{\Delta}_c$, and values of the vectors $\boldsymbol{\beta}$, $\mathbf{u}_1$ and $\mathbf{u}_2$ obtained by applying linear mixed model methodology upon the vectors $\mathbf{y}_{11}$ and $\mathbf{y}_{12}$.

Appendix C

Derivation of the algorithm used for estimating σ_e^2 with normal data

The estimator of σ_e^2 needs to satisfy [20]. From [3A] and [3B]

$$\frac{\delta}{\delta\sigma_e^2} ln\ f(\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2) = \frac{\delta}{\delta\sigma_e^2} \{\sum_i ln\ [\sum_j p_{ij} f(y_i|\boldsymbol{\theta}, \sigma_e^2, \mathcal{L}_{ij})]\}$$

$$= \sum_i \frac{\sum_j p_{ij}\ \delta f(y_i|\boldsymbol{\theta}, \sigma_e^2, \mathcal{L}_{ij})/\delta\sigma_e^2}{\sum_j p_{ij}\ f(y_i|\boldsymbol{\theta}, \sigma_e^2, \mathcal{L}_{ij})} \qquad [C1]$$

Because $y_i|\boldsymbol{\theta}, \sigma_e^2, \mathcal{L}_{ij} \sim N(\mu_{ij}, \sigma_e^2)$, then

$$\frac{\delta}{\delta\sigma_e^2} f(y_i|\boldsymbol{\theta}, \sigma_e^2, \mathcal{L}_{ij}) = \phi\,[(y_i - \mu_{ij})/\sigma_e]\cdot[(y_i - \mu_{ij})^2/\sigma_e^4 - \sigma_e^{-2}]/2\,\sigma_e \qquad [C2]$$

Using this result in [C1] and then in [20] gives :

$$E_c\,\{\sum_i \sum_j q_{ij}\,[(y_i - \mu_{ij})^2/\sigma_e^4 - \sigma_e^{-2}]\} = 0$$

where q_{ij} is as in [6], and where E_c indicates expectation taken with respect to the conditional distribution $f(\boldsymbol{\theta}|\sigma_e^2, \sigma_u^2, \mathbf{y})$. The expectation in [C2] is difficult to obtain because q_{ij} is a function of $\boldsymbol{\theta}$. If q_{ij} is regarded as a constant a rearrangement of [C2] gives :

$$E_c\,\{\sum_i \sum_j q_{ij}\,(y_i - \mu_{ij})^2/\sigma_e^4\} = n\,\sigma_e^2$$

As done by HARVILLE & MEE (1984) in the context of threshold models, we replace $E(\mu_{ij} = \mathbf{w}_{ij}'\,\boldsymbol{\theta}|\sigma_e^2, \sigma_u^2, \mathbf{y})$ by the mode $\hat{\boldsymbol{\theta}}$ calculated using equations [8] (or the expressions described in Appendix B). Thus, the formula above becomes :

$$\sum_i \sum_j q_{ij} y_i^2 - 2(\sum_i y_i)\,(\sum_j q_{ij}\mathbf{w}_{ij}'\hat{\boldsymbol{\theta}}) + \sum_i \sum_j q_{ij}\hat{\boldsymbol{\theta}}'\mathbf{w}_{ij}\mathbf{w}_{ij}'\hat{\boldsymbol{\theta}} + \sum_i \sum_j q_{ij}\,\mathrm{tr}\{\mathbf{w}_{ij}'\,\mathrm{Var}(\boldsymbol{\theta}|\mathbf{y})\mathbf{w}_{ij}\}$$

$$= n\,\sigma_e^2 \qquad [C3]$$

Now, $\mathrm{Var}(\boldsymbol{\theta}|\mathbf{y}) \simeq \mathbf{C}\,\sigma_e^2$, where $\mathbf{C}$ is the inverse of the coefficient matrix in [B6] or [B7] evaluated at $\hat{\boldsymbol{\theta}}$. Further, at the maximum, [5] must be null which is satisfied when [7] is satisfied. Multiplying both sides of [7] by $\hat{\boldsymbol{\theta}}$ (and remembering that a flat prior is used for $\boldsymbol{\beta}$) yields

$$\hat{\boldsymbol{\theta}}'\sum_i \sum_j q_{ij}\mathbf{w}_{ij}\mathbf{w}_{ij}'\,\hat{\boldsymbol{\theta}} + \lambda\hat{\mathbf{u}}'\mathbf{A}^{-1}\,\hat{\mathbf{u}} = \hat{\boldsymbol{\theta}}'\sum_i \sum_j q_{ij}\mathbf{w}_{ij}y_i$$

With this in mind, we obtain the following for the components of [C3]

i) $\sum_i \sum_j q_{ij}\,y_i^2 = \mathbf{y}'\mathbf{y}$

ii) $-2\,(\sum_i y_i)\,(\sum_j q_{ij}\mathbf{w}_{ij}'\hat{\boldsymbol{\theta}}) = -2\,\hat{\boldsymbol{\theta}}'\sum_i \sum_j q_{ij}\mathbf{w}_{ij}y_i$

iii) $\sum_i \sum_j q_{ij}\,\mathrm{tr}(\mathbf{w}_{ij}\,\mathbf{C}\,\mathbf{w}_{ij}')\,\sigma_e^2 = \sigma_e^2\,\mathrm{tr}(\sum_i \sum_j q_{ij}\mathbf{w}_{ij}\mathbf{w}_{ij}'\mathbf{C})$

$$= \sigma_e^2\,\mathrm{tr}(\mathbf{CM})$$

where $\mathbf{M}$ is the coefficient matrix in [8] without $\mathbf{A}^{-1}$.

Using these results in [C3] leads directly to [22].