Theor Appl Genet (1988) 76:81–92

Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance *

I. Hoeschele **

Department of Animal Science, Iowa State University, Ames, IA 50011, USA

Received July 6, 1987; Accepted November 11, 1987
Communicated by D. Van Vleck

Summary. A procedure for genetic evaluation with field data is proposed for situations in which there is mixed major gene and polygenic inheritance and the major genotype membership of some or of all individuals is unknown. Location parameters (fixed environmental, major genotype and polygenic effects), major genotype frequencies and variance components are estimated by the modal values of joint and marginal posterior distributions. The method is described for continuous and discontinuous data as well as for univariate and multivariate evaluations. Results from a simulation study are presented.

Key words: Major genes – Estimation of breeding values – Mixed inheritance – Bayesian inference

Introduction

In animal breeding, mixed linear (Henderson 1973) and nonlinear (e.g., Gianola and Foulley 1983; Harville and Mee 1984; Foulley et al. 1987) model methodology is a powerful and widely used tool for estimating breeding values of candidates for selection. This methodology is based on multifactorial models that include environmental factors such as herd-year-season, age and sex, and polygenic breeding values. Breeding values are usually considered to be the sum of the effects of many genes all having "small" effects. This model of polygenic inheritance satisfactorily fits data on important traits such as milk and fat yield.

However, this does not mean that genes with "large" effects (i.e., single loci that account for an appreciable amount of genetic variation) do not exist. Such loci have been called major loci (Hanset 1982; Roberts and Smith 1982), and evidence for the coexistence of one or few major loci and polygenic effects has been found. Some examples are the muscle hypertrophy in cattle and pigs, the Booroola gene in sheep, the recessive dwarf gene in beef cattle and the rapid postweaning growth gene in mice. Other traits considered as potential candidates for mixed major gene and polygenic inheritance include twinning, calving ease and liability to certain diseases.

In segregation analysis (e.g., Elston and Stewart 1971; Morton and McLean 1974; Bonney 1986) used to investigate the mode of inheritance of a trait, polygenic effects and polygenic heritability are either assumed to not exist or are not separately distinguished.

Data resulting from mixed major gene and polygenic inheritance require statistical methods for detecting major genotypes and for genetic evaluations including major genotypes and polygenic breeding values. This paper presents a method for genetic evaluation with partly or fully unknown major genotype membership of individuals. It requires evidence for the existence of a major locus and for the number of major genotypes. Detection of major genotypes is addressed in a different communication (Hoeschele 1988).

Methodology

Model

Consider the mixed linear model

$$y_{ik} = \mathbf{w}'_{ik}\mathbf{g} + \mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u} + e_{ik} \tag{1}$$

where y_{ik} is an observation on the i^th individual, $\mathbf{g}$ is an $m \times 1$ vector of major genotypic means, $\boldsymbol{\beta}$ is a $(p \times 1)$

[*] Journal Paper No. J-12728 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. 1901

[**] Present address: Department of Dairy Science, Virginia Tech, Blacksburg, VA 24061, USA

vector of systematic environmental factors and $\mathbf{u}$ is a $q \times 1$ vector of polygenic breeding values or sire and dam effects; $\mathbf{w}'_{ik}$ is a row vector having a one in column k if the individual has the k^{th} major genotype and zero elsewhere, $\mathbf{x}'_i$ and $\mathbf{z}'_i$ are the i^{th} rows of the incidence matrices $\mathbf{X}$ and $\mathbf{Z}$ and e_{ik} is a residual with $\mathrm{var}(e_{ik}) = \boldsymbol{\sigma}_e^2$. Denote known major genotype membership (i.e., known $\mathbf{w}'_{ik}$) by $I_i = k$, $k \in \{1, 2, \ldots, m\}$, and let $\mathbf{w}'_{ik}\mathbf{g} \equiv g_k$. Then, assuming normality

$$y_i \mid I_i = k, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2 \sim N(g_k + \mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u}, \sigma_e^2)$$

and

$$\mathrm{cov}(y_i, y_{i'} \mid I_i = k, I_{i'} = k', \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2) = 0. \tag{2}$$

The density function of y_i conditional on the genotype membership and on the parameters $\boldsymbol{\theta}' = [\mathbf{g}', \boldsymbol{\beta}', \mathbf{u}']$ and σ_e^2 will be denoted by $f(y_i \mid I_i = k, \boldsymbol{\theta}, \sigma_e^2)$.

If the major genotype membership I_i is unknown, y_i has an m-component mixture distribution with mean

$$E_h(y_i \mid \mathbf{p}, \boldsymbol{\theta}, \sigma_e^2) = \sum_{k=1}^{m} p(I_i = k)\, g_k + \mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u} \tag{3}$$

and variance

$$\mathrm{var}_h(y_i \mid \mathbf{p}, \boldsymbol{\theta}, \sigma_e^2) = \sum_{k=1}^{m} p(I_i = k)(g_k - \mu_g)^2 + \sigma_e^2 \tag{4}$$

where $\mathbf{p}$ is an $m \times 1$ vector with elements $p(\mathbf{I}_i = k)$, the probabilities that the i^{th} individual has major genotype membership $k \in (k = 1, \ldots, m)$ and $\mu_g = \sum_{k=1}^{m} p(I_i = k)\, g_k$. Also, h denotes expectation and variance with respect to the density of y_i marginal with respect to the uncertain genotype membership

$$h(y_i \mid \mathbf{p}, \boldsymbol{\theta}, \sigma_e^2) = \sum_{k=1}^{m} p(I_i = k)\, f(y_i \mid I_i = k, \boldsymbol{\theta}, \sigma_e^2). \tag{5}$$

In (5), $h(\cdot)$ is called an m-component mixture density and $f(\cdot)$ a component density (Titterington et al. 1985).

Statistical inference

Prior information. Prior information on $\boldsymbol{\theta}$ and $\mathbf{p}$ will be used by specifying Normal priors for the location parameters and a Dirichlet prior for the unknown probabilities of major genotype membership, so

$$\begin{array}{lll}
\mathbf{g} \sim N(\mathbf{1}\mu_g, \boldsymbol{\Sigma}_g), & g_k \in (-\infty, \infty) & k = 1, \ldots, m \\
\boldsymbol{\beta} \sim N(\mathbf{1}\mu_\beta, \boldsymbol{\Sigma}_\beta), & \beta_i \in (-\infty, \infty) & i = 1, \ldots, p \\
\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}), & u_i \in (-\infty, \infty) & i = 1, \ldots, q \\
\mathbf{p} \sim D(\boldsymbol{\alpha}) & p_k \in [0, 1] & k = 1, \ldots, m
\end{array}$$

where N and D denote Normal and Dirichlet distributions. It is convenient to assume that $\mathbf{g}$, $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{p}$ are independent a priori and that prior knowledge about $\mathbf{g}$, $\boldsymbol{\beta}$, and $\mathbf{p}$ is vague, implying $\boldsymbol{\Sigma}_g \to \infty$, $\boldsymbol{\Sigma}_\beta \to \infty$ or equivalently, $\boldsymbol{\Sigma}_g^{-1} \to \mathbf{0}$, $\boldsymbol{\Sigma}_{\sim\beta}^{-1} \to \mathbf{0}$, and $\boldsymbol{\alpha}_k = 1$ for all $k \in \{1, 2, \ldots, m\}$. Then we can write the prior density function as

$$\begin{aligned}
l(\boldsymbol{\theta}, \mathbf{p}) &= l(\mathbf{g})\, l(\boldsymbol{\beta})\, l(\mathbf{u})\, l(\mathbf{p}) && (6)\\
&= C\, l(\mathbf{u}) \\
&= C^* \exp\{-\tfrac{1}{2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\,\sigma_u^{-2}\} && (7)
\end{aligned}$$

where C and C* represent appropriate constants. In (7), $\mathbf{A}$ is a matrix of known additive genetic relationship between the elements in $\mathbf{u}$, σ_u^2 is polygenic variance and $\mathbf{G} = \mathbf{A}\sigma_u^2$. We will assume first that σ_u^2 is known.

Likelihood. Using (5) and assuming σ_e^2 is known, the joint likelihood of all observations is

$$h(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2) = \sum_{\mathbf{K}} p(\mathbf{I} = \mathbf{K})\, f(\mathbf{y} \mid \mathbf{I} = \mathbf{K}, \boldsymbol{\theta}, \sigma_e^2) \tag{8}$$

where $\mathbf{K}$ is a particular $N \times 1$ vector of major genotype memberships, N is the total number of observations, the summation $\sum_{\mathbf{K}}$ represents a nested sum of the form $\sum_{k_1=1}^{m} \left\{ \sum_{k_2=1}^{m} \left\{ \ldots \left\{ \sum_{k_N=1}^{m} \right.\right.\right.$, and $p(\mathbf{I} = \mathbf{K})$ is the joint probability of N particular genotype memberships in $\mathbf{K}$. This form of the likelihood is difficult to handle, and simplifications are required.

Assume that the data consist of records on unrelated parents and their progeny. Given the major genotype memberships of the sire and dam, the genotype memberships of the progeny are conditionally independent. If there is no assortative mating with respect to the major genotype, those of the sire and dam are also independent. Then, simplified forms of the likelihood can be found and are presented in Appendices 1 and 2.

Inferences. Using Bayes theorem, the product of the likelihood in (A1.3) and the prior density in (7) is proportional to the joint posterior density of the location parameters $\boldsymbol{\theta}$ and the major genotype frequencies $\mathbf{p}$,

$$h(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma}) = \text{constant } h(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2) \cdot l(\mathbf{u} \mid \sigma_u^2)\, l(\mathbf{p}) \tag{9}$$

with $\boldsymbol{\sigma}' = [\sigma_e^2, \sigma_u^2]$ assumed known. Inferences about $\boldsymbol{\theta}$ and $\mathbf{p}$ may be obtained from (9). An alternative approach is to consider the marginal posterior densities of $\boldsymbol{\theta}$ and $\mathbf{p}$:

$$\begin{aligned}
t(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\sigma}) &= \int_{R_p} h(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma})\, d\mathbf{p} \\
&= \int_{R_p} f(\boldsymbol{\theta} \mid \mathbf{p}, \mathbf{y}, \boldsymbol{\sigma})\, t(\mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma})\, d\mathbf{p} && (10)
\end{aligned}$$

where $t(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\sigma})$ is the marginal posterior density of $\boldsymbol{\theta}$, taking into account uncertainty about $\mathbf{p}$, and $t(\mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma})$ is the marginal posterior density of $\mathbf{p}$ obtained from

$$\begin{aligned}
t(\mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma}) &= \int_{R_\theta} h(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma})\, d\boldsymbol{\theta} \\
&= \int_{R_\theta} f(\mathbf{p} \mid \boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\sigma})\, t(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\sigma})\, d\boldsymbol{\theta}. && (11)
\end{aligned}$$

If $t(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\sigma})$ were reasonably peaked, integrating $\boldsymbol{\theta}$ out could be approximated by evaluating the conditional density of $\mathbf{p}$ at the mode $\hat{\boldsymbol{\theta}}$ of $t(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\sigma})$ (Box and Tiao 1973) so that

$$t(\mathbf{p}|\mathbf{y},\boldsymbol{\sigma}) \doteq f(\mathbf{p}|\hat{\boldsymbol{\theta}},\mathbf{y},\boldsymbol{\sigma}). \tag{12}$$

Also, we have

$$f(\mathbf{p}|\hat{\boldsymbol{\theta}},\mathbf{y},\boldsymbol{\sigma}) = g(\mathbf{y}|\mathbf{p},\hat{\boldsymbol{\theta}},\boldsymbol{\sigma})\, l(\mathbf{p}) \text{ constant.}$$

Combining (12) and (10) then provides an approximation to the marginal posterior density of $\boldsymbol{\theta}$:

$$t(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\sigma}) \doteq \int_{R_p} f(\boldsymbol{\theta}|\mathbf{p},\mathbf{y},\boldsymbol{\sigma})\, f(\mathbf{p}|\hat{\boldsymbol{\theta}},\mathbf{y},\boldsymbol{\sigma})\, d\mathbf{p}. \tag{13}$$

Inferences about the unknown parameters in $\boldsymbol{\theta}$ and $\mathbf{p}$ will be made by point estimation using the joint $(\boldsymbol{\theta},\mathbf{p})$ mode of (9) or approximating the marginal $\boldsymbol{\theta}$ and $\mathbf{p}$ modes of (10) and (11), respectively. The mode represents the most likely vector of values for $\boldsymbol{\theta}$ and $\mathbf{p}$, given the data, and with "large" sample sizes, the posterior mode is close to the posterior mean (Berger 1985). This approach will be referred to as Maximum A-Posteriori Estimation (MAPE). A similar approach has been developed by Foulley et al. (1986) for genetic evaluation with uncertain paternity.

Algorithms of computation

First, consider estimating the unknown location parameters and major genotype frequencies by the mode of the joint posterior distribution of $\boldsymbol{\theta}$ and $\mathbf{p}$, $h(\boldsymbol{\theta},\mathbf{p}|\mathbf{y},\boldsymbol{\sigma})$. The vector of modal values is found by using an iterative algorithm that converges to a maximum of (9). However, because (9) cannot be assumed to have a single maximum (Titterington et al. 1985), this requires using starting values close to its global maximum. The logarithm of (9) using (A 1.3) is:

$$\begin{aligned}\log h(\boldsymbol{\theta},\mathbf{p}|\mathbf{y},\boldsymbol{\sigma}) &= \sum_{S=1}^{N_S} \log\left[\sum_{k=1}^{m} p(I_S=k)\, f(y_S|I_S=k,\boldsymbol{\theta},\sigma_e^2)\right] \\ &+ \sum_{D=1}^{N_D} \log\left[\sum_{l=1}^{m} p(I_D=l)\, f(y_D|I_D=l,\ \boldsymbol{\theta},\sigma_e^2)\right] \\ &+ \log\left[\sum_{K_S}\sum_{K_D}\prod_{S=1}^{N_S} p(I_S=k_S|y_S)\prod_{D=1}^{N_D} p(I_D=k_D|y_D)\right. \\ &\left.\cdot\prod_{i=1}^{N_p}\sum_{r=1}^{m} p(I_i=r|I_{S(i)}=k_{S(i)},I_{D(i)}=k_{D(i)})\, f(y_i|I_i=r,\boldsymbol{\theta},\sigma_e^2)\right] \\ &- 1/2\,\mathbf{u}'\mathbf{G}^{-1}\mathbf{u} + \text{constant}\end{aligned} \tag{14}$$

assuming $l(\mathbf{p}) = $ constant. An iterative algorithm can be obtained by differentiating (14) with respect to $\boldsymbol{\theta}$ and $\mathbf{p}$ and setting the derivatives equal to zero. Because of the nested summation, differentiation of (14) is very difficult. However, derivatives of the logarithms of (A 2.1), (A 2.2), (A 2.3), (A 2.4) and (A 2.5) are given in Appendix 3.

The derivatives of the log posterior density (14), using the derivatives of the log likelihoods in Appendix 3, have the general form

$$\frac{\partial \log h(\boldsymbol{\theta},\mathbf{p}|\mathbf{y},\boldsymbol{\sigma})}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{N}\sum_{k=1}^{m} p(I_i=k|\mathbf{y})(y_i-\boldsymbol{\Delta}'_{ik}\boldsymbol{\theta})\bar{\sigma}_e^2\,\boldsymbol{\Delta}_{ik} - \begin{bmatrix}\mathbf{0}_{\dim(\mathbf{g})+\dim(\boldsymbol{\beta})}\\ \bar{\sigma}_u^2\mathbf{A}^{-1}\mathbf{u}\end{bmatrix} \tag{15}$$

where N is now the total number of individual with records. Equating (15) to zero and rearranging gives a nonlinear system of equations in $\boldsymbol{\theta}$:

$$\begin{bmatrix}\mathbf{D}^{[l]} & \mathbf{Q}'^{[l]}\mathbf{X} & \mathbf{Q}'^{[l]}\mathbf{Z}\\ \mathbf{X}'\mathbf{0}^{[l]} & \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{Q}^{[l]} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda\end{bmatrix}\begin{bmatrix}\mathbf{g}^{[l+1]}\\ \boldsymbol{\beta}^{[l+1]}\\ \mathbf{u}^{[l+1]}\end{bmatrix}\begin{matrix}=\\=\\=\end{matrix}\begin{bmatrix}\mathbf{Q}'^{[l]}\mathbf{y}\\ \mathbf{X}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\end{bmatrix} \tag{16}$$

where

$$\mathbf{D}^{[l]} = \text{Diag}\left\{\sum_{i=1}^{N} p^{[l]}(I_i=k|\mathbf{y}_{m\times m})\right\}$$

$$\mathbf{Q}^{[l]} = \{p^{[l]}(I_i=k|\mathbf{y}_{N\times m})\}$$

and the $p^{[l]}(I_i=k|\mathbf{y})$ values are those given in Appendix 3 evaluated at $\boldsymbol{\theta}^{[l]}$ and $\mathbf{p}^{[l]}$. In relation to Bayesian classification, $p(I_i=k|\mathbf{y})$ represents the posterior probability that the i^th^ individual has major genotype membership k (Geysser 1982) if $\boldsymbol{\theta}$ were known, and with $\sum_{k=1}^{m} p(I_i=k|\mathbf{y}) = 1$.

Also, $\mathbf{G} = \mathbf{A}\sigma_u^2$, and $\lambda = \sigma_e^2|\sigma_u^2$. The unknown unconditional probabilities of major genotype membership $p(I_i=k)$ are estimated by

$$\hat{p}^{[l]}(I_i=k) = \frac{1}{N}\sum_{i=1}^{N} p^{[l]}(I_i=k|\mathbf{y}) \qquad k=1,\ldots,m. \tag{17}$$

The estimator (17) can be derived by differentiating $\log h(\boldsymbol{\theta},\mathbf{p}|\mathbf{y},\sigma)$ with respect to $\mathbf{p}$ and equating the derivative to zero. For estimating $\boldsymbol{\theta}$ and $\mathbf{p}$, (16) and (17) are combined in an iterative scheme. Iteration starts with a set of initial guesses $\boldsymbol{\theta}^{[0]}$ and $\mathbf{p}^{[0]}$ and stops when a convergence criterion such as $\{[\hat{\boldsymbol{\theta}}^{[l+1]}-\hat{\boldsymbol{\theta}}^{[l]}]\ [\hat{\boldsymbol{\theta}}^{[l+1]}-\hat{\boldsymbol{\theta}}^{[l]}]/\dim(\boldsymbol{\theta})\}^{0.5} < \varepsilon$ is satisfied, ε being an arbitrarily small number.

An alternative iteration scheme can be obtained using Newton's method (Kennedy and Gentle 1980) requiring second derivatives of (14). This algorithm converges rapidly even for nonlinear functions not quadratic in the parameters.

The second approach consists of estimating $\boldsymbol{\theta}$ by the corresponding marginal posterior mode. The vector of first derivatives of the logarithm of the marginal posterior density of $\boldsymbol{\theta}$ in (10) is

$$\begin{aligned}\frac{\partial \log}{\partial\boldsymbol{\theta}} t(\boldsymbol{\theta}|\mathbf{y},\sigma) &= \underset{\mathbf{p}|\mathbf{y},\boldsymbol{\sigma}}{\mathrm{E}}\left[\frac{\partial \log}{\partial\boldsymbol{\theta}} h(\boldsymbol{\theta},\mathbf{p}|\mathbf{y},\sigma)\right] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{m}\underset{\mathbf{p}|\mathbf{y},\boldsymbol{\sigma}}{\mathrm{E}}[p(I_i=k|\mathbf{y})](y_i-\boldsymbol{\Delta}'_{ik}\boldsymbol{\theta})\,\sigma_e^{-2}\,\boldsymbol{\Delta}_{ik} \\ &\quad - \begin{bmatrix}\mathbf{0}_{\dim(\mathbf{g})+\dim(\boldsymbol{\beta})}\\ \sigma_u^{-2}\mathbf{A}^{-1}\mathbf{u}\end{bmatrix}\end{aligned} \tag{18}$$

where $\underset{\mathbf{p}|\mathbf{y},\boldsymbol{\sigma}}{\mathrm{E}}$ denotes expectation with respect to the density $\mathrm{t}(\mathbf{p}|\mathbf{y},\boldsymbol{\sigma})$. Based on (18), $\boldsymbol{\theta}$ is estimated by iterating with (16) and replacing the $\mathrm{p}(\mathrm{I}_i = \mathrm{k}|\mathbf{y})$ values in **D** and **Q** by $\underset{\mathbf{p}|\mathbf{y},\boldsymbol{\sigma}}{\mathrm{E}} \mathrm{p}(\mathrm{I}_i = \mathrm{k}|\mathbf{y})$. This approach is more demanding computationally because it requires calculating, for each record and in each round of iteration, the quantity

$$\underset{\mathbf{p}|\mathbf{y},\boldsymbol{\sigma}}{\mathrm{E}} [\mathrm{p}(\mathrm{I}_i = \mathrm{k}|\mathbf{y})] \doteq \int_0^1 \ldots \int_0^1 \mathrm{p}(\mathrm{I}_i = \mathrm{k}|\mathbf{y}) \cdot \mathrm{f}(\mathbf{p}|\hat{\boldsymbol{\theta}}^{[1]}, \mathbf{y}, \boldsymbol{\sigma})\, \mathrm{dp}_1 \ldots \mathrm{dp}_{m-1} \cdot \qquad (19)$$

Also, the $\boldsymbol{\theta}$-mode of $\mathrm{h}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \boldsymbol{\sigma})$ may often closely approximate the mode of $\mathrm{t}(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\sigma})$. It can be shown that the posterior density of **p** in (12) is a mixture of Dirichlet densities, $\mathrm{f}(\mathbf{p}|\hat{\boldsymbol{\theta}}^{[1]}, \mathbf{y}, \boldsymbol{\sigma}) = \sum_{\mathrm{K}} \mathrm{w}_{\mathrm{K}} \mathrm{D}(\mathrm{n}_1 + \alpha_1, \ldots, \mathrm{n}_m + \alpha_m)$, where the w_{K} are appropriate weights, n_k $(\mathrm{k} = 1, \ldots, \mathrm{m})$ is the number of observations with major genotype-membership k, and the summation goes over all sets of positive integers $(\mathrm{n}_1, \ldots, \mathrm{n}_k)$ with $\sum_{k=1}^{m} \mathrm{n}_k = N$.

Partly known genotype membership

Determination of the major genotype membership by biochemical methods can be expensive and complicated and, therefore, is not performed for all individuals. Studies on mixture distributions (Titterington et al. 1985) have shown that it is clearly worthwhile to include unclassified observations in discriminant analyses. If selection based on functions of the data is ongoing, unclassified records should not be discarded to avoid selection bias. Using Titterington et al. (1985), the appropriate log posterior density for partly known major genotype membership is

$$\log \mathrm{h}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \boldsymbol{\sigma}) = (14) + \sum_{k=1}^{m} \sum_{i=1}^{n_k} \log \mathrm{f}(\mathrm{y}_i|\mathrm{I}_i = \mathrm{k}, \boldsymbol{\theta}, \sigma_e^2) + \sum_{k=1}^{m} \mathrm{n}_k \log \mathrm{p}(\mathrm{I} = \mathrm{k}) - \tfrac{1}{2} \mathbf{u}' \mathbf{G}^{-1} \mathbf{u} + \text{constant} \qquad (20)$$

where $N_1 = N_S + N_D + N_p$ is the number of individuals with unknown genotype membership, n_k is the number of individuals with known genotype membership k, and $\sum_{k=1}^{m} \mathrm{n}_k = N_2$ is the total number of individuals with known genotype membership.

Based on (20), the following equations to estimate $\boldsymbol{\theta}$ and **p** are obtained:

$$\begin{bmatrix} \mathbf{D}_1^{[1]} + \mathbf{W}_2' \mathbf{W}_2 & \mathbf{Q}_1'^{[1]} \mathbf{X}_1 + \mathbf{W}_2' \mathbf{X}_2 & \mathbf{Q}_1'^{[1]} \mathbf{Z}_1 + \mathbf{W}_2' \mathbf{Z}_2 \\ \mathbf{X}_1' \mathbf{Q}_1^{[1]} + \mathbf{X}_2' \mathbf{W}_2 & \mathbf{X}_1' \mathbf{X}_1 + \mathbf{X}_2' \mathbf{X}_2 & \mathbf{X}_1' \mathbf{Z}_1 + \mathbf{X}_2' \mathbf{Z}_2 \\ \mathbf{Z}_1' \mathbf{Q}_1^{[1]} + \mathbf{Z}_2' \mathbf{W}_2 & \mathbf{Z}_1' \mathbf{X}_1 + \mathbf{Z}_2' \mathbf{X}_2 & \mathbf{Z}_1' \mathbf{Z}_1 + \mathbf{Z}_2' \mathbf{Z}_2 + \mathbf{A}^{-1} \lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{g}} \\ \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix}^{[1+1]} = \begin{bmatrix} \mathbf{Q}_1'^{[1]} \mathbf{y}_1 + \mathbf{W}_2' \mathbf{y}_2 \\ \mathbf{X}_1' \mathbf{y}_1 + \mathbf{X}_2' \mathbf{y}_2 \\ \mathbf{Z}_1' \mathbf{y}_1 + \mathbf{Z}_2' \mathbf{y}_2 \end{bmatrix} \qquad (21)$$

$$\hat{\mathrm{p}}^{[1]}(\mathrm{I} = \mathrm{k}) = \frac{\sum_{i=1}^{N_1} \mathrm{p}^{[1]}(\mathrm{I}_i = \mathrm{k}|\mathbf{y}) + \mathrm{n}_k}{N_1 + N_2} \qquad \mathrm{k} = 1, \ldots, \mathrm{m} \qquad (22)$$

where

$$\mathbf{D}_1 = \mathrm{Diag}\left\{\sum_{i=1}^{N_1} \mathrm{p}(\mathrm{I}_i = \mathrm{k}|\mathbf{y})\right\}_{N_1 \times m},$$

and

$$\mathbf{Q}_1 = \{\mathrm{p}(\mathrm{I}_i = \mathrm{k}|\mathbf{y})\}_{N_1 \times N_1};$$

indices 1 and 2 refer to individuals with unknown and known genotype memberships, and $\mathbf{W}_2$ is the known part of the incidence matrix of **g**.

Estimation of variance components and heritability

The estimation equations in (16) and (21) require knowing the polygenic and residual variances, i.e., σ_u^2 and σ_e^2. Usually these are unknown and replaced by estimates. Gianola et al. (1986) justify the use of REML (Restricted Maximum Likelihood) estimates of σ_u^2 and σ_e^2 in place of the true values, and Hoeschele et al. (1987) suggest estimating σ_u^2 and σ_e^2 by "Marginal Maximum Likelihood", which reduces to REML under normality. "Marginal Maximum Likelihood" estimation consists of finding the mode of the marginal posterior density of the variances and employing flat prior densities, i.e., $\mathrm{f}(\sigma_e^2) = \text{constant}$, and $\mathrm{f}(\sigma_u^2) = \text{constant}$ (Gianola and Fernando 1986). This is done by equating the derivatives of the log posterior density to zero:

$$\frac{\partial \log}{\partial \sigma_i^2} \mathrm{t}(\boldsymbol{\sigma}|\mathbf{y}) = \int_{R_{\boldsymbol{\theta}}} \int_{R_{\mathbf{p}}} \left[\frac{\partial \log}{\partial \sigma_i^2} \mathrm{h}(\boldsymbol{\sigma}, \boldsymbol{\theta}, \mathbf{p}|\mathbf{y})\right] \mathrm{h}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \boldsymbol{\sigma})\, \mathrm{d}\mathbf{p}\, \mathrm{d}\boldsymbol{\theta} = \underset{(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \boldsymbol{\sigma})}{\mathrm{E}} \left[\frac{\partial \log}{\partial \sigma_i^2} \mathrm{h}(\boldsymbol{\sigma}, \boldsymbol{\theta}, \mathbf{p}|\mathbf{y})\right] \qquad (23)$$

where $\sigma_i^2 \equiv \sigma_u^2$, or $\sigma_i^2 \equiv \sigma_e^2$. Because $\mathrm{h}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \boldsymbol{\sigma})$ is not in the form of a normal density and $\dim(\boldsymbol{\theta})$ is usually large, the integration (23) is difficult computationally. Berger (1985) suggests normal approximations to non-normal posterior densities and gives a heuristic proof based on "large" sample size, implying a sharply concentrated posterior density that is approximately normal. Here, we might consider the approximations

$$(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma}) \dot{\sim} N(\hat{\boldsymbol{\theta}}_h, \mathrm{C}_h) \quad \text{or} \quad (\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma}) \dot{\sim} \mathrm{N}(\hat{\boldsymbol{\theta}}_t, \mathrm{C}_t) \qquad (24)$$

where $\hat{\boldsymbol{\theta}}_h$ and C_h [$\hat{\boldsymbol{\theta}}_t$ and C_t] are solution and inverted coefficient matrix of [16] at convergence derived from the joint posterior $\mathrm{h}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \boldsymbol{\sigma})$ [marginal posterior $\mathrm{t}(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\sigma})$].

Because asymptotic normality does not hold on the boundary of the parameter space, $h(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma})$ will not be approximately normal if the true parameter p_k is zero for some $k \in \{1, 2, \ldots, m\}$. For $\sigma_i^2 \equiv \sigma_u^2$, Hoeschele et al. (1987) showed that

$$\frac{\partial \log}{\partial \sigma_u^2} t(\boldsymbol{\sigma} \mid \mathbf{y}) = \mathop{\mathrm{E}}_{(\mathbf{u} \mid \mathbf{y}, \boldsymbol{\sigma})} \left[\frac{\partial \log}{\partial \sigma_u^2} l(\mathbf{u} \mid \sigma_u^2) \right]. \qquad (25)$$

Using (7), (24) and equating (25) to zero gives the estimator of σ_u^2:

$$\hat{\sigma}_u^{2^{[l+1]}} = \frac{\hat{\mathbf{u}}'^{[l]} \mathbf{A}^{-1} \hat{\mathbf{u}}^{[l]}}{q - \mathrm{tr}(\mathbf{A}^{-1} \mathbf{C}_{uu}^{[l]}) \lambda^{[l]}} \qquad (26)$$

where $\hat{\mathbf{u}}$ and $\mathbf{C}_{uu}$ are either $\hat{\mathbf{u}}_h$ and $\mathbf{C}_{uu(h)}$ or $\hat{\mathbf{u}}_t$ and $\mathbf{C}_{uu(t)}$, $\mathbf{C}_{uu}$ is the $q \times q$ part of $\mathbf{C}$ referring to $\mathbf{u}$, and $\lambda^{[l]} = \hat{\sigma}_e^{2^{[l]}} / \hat{\sigma}_u^{2^{[l]}}$.

For $\sigma_i^2 = \sigma_e^2$ and $l(\boldsymbol{\sigma}) = \text{constant}$, (23) becomes

$$\frac{\partial \log}{\partial \sigma_e^2} t(\boldsymbol{\sigma} \mid \mathbf{y}) = \mathop{\mathrm{E}}_{(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma})} \left[\frac{\partial \log}{\partial \sigma_e^2} h(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2) \right]. \qquad (27)$$

Using the likelihood functions (A2.1), (A2.2), (A2.3), (A2.4) and (A2.5),

$$f(y_i \mid I_i = k, \boldsymbol{\theta}, \sigma_e^2) = \frac{1}{\sqrt{2\pi}\,\sigma_e} \exp\left\{ -\frac{1}{2\sigma_e^2} (y_i - \Delta'_{ik} \boldsymbol{\theta})^2 \right\},$$

and following Gianola et al. (1986) and Foulley et al. (1986) lead to the estimator

$$\hat{\sigma}_e^{2^{[l+1]}} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{m} \hat{p}^{[l]}(I_i = k \mid \mathbf{y})\, \hat{e}_{ik}^{2^{[l]}}}{N - \dim(\boldsymbol{\theta}) + \mathrm{tr}(\mathbf{A}^{-1} \mathbf{C}_{uu}^{[l]}) \lambda^{[l]}} \qquad (28)$$

where $\hat{e}_{ik} = y_i - \Delta'_{ik} \hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\theta}}$ is $\hat{\boldsymbol{\theta}}_h$ or $\hat{\boldsymbol{\theta}}_t$, respectively.

For unknown σ_e^2 and σ_u^2, four nonlinear systems of equations have to be solved simultaneously, namely, (16), (17), or (21), (22), (26) and (28).

"Polygenic heritability" (h_p^2), i.e., heritability within major genotype, and the relative contribution of the variance explained by the major genotypes to phenotypic variance (h_{ML}^2) are defined as

$$h_p^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}, \quad h_{ML}^2 = \frac{\sigma_{ML}^2}{\sigma_{ML}^2 + \sigma_u^2 + \sigma_e^2} \qquad (29, 30)$$

where $\sigma_{ML}^2 = \sum_{k=1}^{m} p_k (g_k - \mu_g)^2$ and $\mu_g = \sum_{k=1}^{m} p_k g_k$.

An estimate of σ_{ML}^2 may be obtained from

$$\hat{\sigma}_{ML}^2 = \sum_{k=1}^{m} \hat{p}_k (\hat{g}_k - \hat{\mu}_g)^2 + \text{constant} \qquad (31)$$

where the constant is a correction factor given in Searle (1971, p. 476) for the two-way nested classification with, e.g., $n_{j.}$ being progeny group size of sire j and n_{jk} replaced by $\hat{P}_{jk}$ with $\hat{P}_{jk} = \sum_{i=1}^{n_{j.}} \hat{p}(I_i = k \mid \mathbf{y})$.

Genotype-environment interaction

A given environmental difference may have more effect on some genotypes than on others, i.e., the major genotypes differ in their environmental sensitivity. Consequently, the model needs to account for environment and major genotype interaction and for heterogeneous residual variance. Similarly, interactions with the polygenic background of different lines or breeds are possible (Roberts and Smith 1982). This can be achieved by implementing the following modifications. First, replace (2) by

$$(y_i \mid I_i = k, \mathbf{g}\boldsymbol{\beta}, \mathbf{u}, \sigma_k^2) \sim N(\mathbf{x}'_{ik} \mathbf{g}\boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u}, \sigma_k^2) \qquad (32)$$

where $\mathbf{g}\boldsymbol{\beta} = \{(g\beta)_{ik}\}$ and $(g\beta)_{ik}$ is the joint effect of the k^th^ major genotype ($k = 1, \ldots, m$) and the i^th^ "fixed" effect in $\boldsymbol{\beta}$ ($i = 1, \ldots, p$). Secondly, the vector of first partial derivatives of the log posterior density using normal component densities becomes

$$\frac{\partial \log}{\partial \boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} \sum_{k=1}^{m} p(I_i = k \mid \mathbf{y})(y_i - \Delta'_{ik} \boldsymbol{\theta}) \sigma_k^{-2} \Delta_{ik} - \begin{bmatrix} \mathbf{0}_{\dim(\mathbf{g}\boldsymbol{\beta})} \\ \bar{\sigma}_u^2 \mathbf{A}^{-1} \mathbf{u} \end{bmatrix} \qquad (33)$$

where $\Delta'_{ik} = [\mathbf{x}'_{ik} : \mathbf{z}'_i]$, and $\boldsymbol{\theta}' = [\mathbf{g}\boldsymbol{\beta}', \mathbf{u}']$. Equating (33) to zero and solving for $\boldsymbol{\theta}$ gives the following nonlinear system of equations to estimate $\mathbf{g}\boldsymbol{\beta}$ and $\mathbf{u}$:

$$\begin{bmatrix} \mathbf{D}^{[l]} & \mathbf{Q}'^{[l]} \mathbf{Z} \\ \mathbf{Z}' \mathbf{Q}^{[l]} & \mathbf{Z}' \mathbf{R}^{-1^{[l]}} \mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{g}\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix}^{[l+1]} = \begin{bmatrix} \mathbf{Q}'^{[l]} \mathbf{y} \\ \mathbf{Z}' \mathbf{y} \end{bmatrix} \qquad (34)$$

where

$$\mathbf{D}^{[l]} = \mathrm{Diag} \left\{ \sigma_k^{-2} \sum_{j=1}^{n_{i.}} \hat{p}^{[l]}(I_{ij} = k \mid \mathbf{y}) \right\}_{mp \times mp},$$

$$\mathbf{Q}^{[l]} = \{ \sigma_k^{-2} \hat{p}^{[l]}(I_{ij} = k \mid \mathbf{y}) \}_{N \times mp},$$

$$\mathbf{R}^{-1^{[l]}} = \mathrm{Diag} \left\{ \sum_{k=1}^{m} \sigma_k^{-2} \hat{p}(I_{ij} = k \mid \mathbf{y} \right\}_{N \times N},$$

$n_{i.}$ is the number of observations in the i^ih^ level of $\boldsymbol{\beta}$, and $ij \left(i = 1, \ldots, p; j = 1, \ldots, n_{i.}; \sum_{i=1}^{p} n_{i.} = N \right)$ is used to identify an element in $\mathbf{y}$ instead of i ($i = 1, \ldots, N$) employed elsewhere. Also, σ_k^2 may be estimated by approximately modifying (28).

Large number of alleles at the major locus

The number of alleles at a major locus may be large with one allele having an outlier effect. In this situation, we have to distinguish between three major genotypes denoted by $g_1 \equiv g_{aa}$, $g_2 \equiv g_{aw}$ and $g_3 \equiv g_{ww'}$. Let a be the outlier allele effect and $w, w' \sim N(0, \sigma_w^2)$. Then, we can write $\mu_{aa} = E(g_{aa}) = g_{aa}$, $\mu_{aw} = E(g_{aw}) = \mu + a$ and $\mu_{ww'} = E(g_{ww'}) = \mu$ with μ being the mean genotypic value. One approach is to estimate $[\mu_{aa}, \mu_{aw}$ and $\mu]$ in place of g and proceed as before if σ_w^2 is small. Otherwise, heterogeneity of the polygenic variance requires an appropriate modification of the G matrix with

$$\operatorname{var}(u \mid \mu_{aa}) = \operatorname{var}(u) = \sigma_u^2,\ \operatorname{var}(u \mid \mu_{aw}) = \operatorname{var}(u + w),$$

and

$$\operatorname{var}(u \mid \mu_{ww'}) = \operatorname{var}(u + w + w').$$

Extension to multivariate analyses

The power of the proposed method may be increased by a multivariate analysis of correlated traits. Consider, for instance, two highly correlated traits. Then, the posterior probabilities of major genoype membership $p(I_i = k \mid y)$ will probably be less flat than those estimated in single-trait analyses because separating between m overlapping univariate normal distributions is more difficult than between m bivariate distributions.

Suppose that $\mathbf{y}_i$ is the vector of observations on r continuous traits for the i^th^ individual, and assume that

$$\mathbf{y}_i \mid I_i = k, \boldsymbol{\theta}, \boldsymbol{\Sigma}_e \sim N_r(\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_e) \tag{35}$$

where $\boldsymbol{\theta}' = [\mathbf{g}', \boldsymbol{\beta}', \mathbf{u}']$ is of dimension $r \times (m + p + q)$ (assuming the same model for each trait), $\boldsymbol{\Sigma}_e$ is the $r \times r$ residual covariance matrix, N_r denotes the r-variate normal distribution, $\boldsymbol{\mu}_{ik} = (\boldsymbol{\Delta}'_{ik} \otimes \mathbf{I}_r)\boldsymbol{\theta}$ (assuming equal incidence matrices for all traits), and $\otimes$ denotes the Kronecker product. The log posterior density of $\boldsymbol{\theta}$ and $\mathbf{p}$ is

$$\begin{aligned}&\log h(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\Sigma}_e, \boldsymbol{\Sigma}_u)\\ &\quad = \log h(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \boldsymbol{\Sigma}_e) - 1/2\, \mathbf{u}'(\mathbf{A}^{-1} \otimes \boldsymbol{\Sigma}_u^{-1})\, \mathbf{u}.\end{aligned} \tag{36}$$

The conditional densities of the data are replaced by the r-variate density $f_r(\mathbf{y}_i \mid I_i = k, \boldsymbol{\theta}, \boldsymbol{\Sigma}_e)$, $\boldsymbol{\Sigma}_u$ is the $r \times r$ polygenic covariance matrix and $\boldsymbol{\theta}$ is ordered by level. The gradient vector of (36) is

$$\begin{aligned}&\frac{\partial \log}{\partial \boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}, \boldsymbol{\Sigma}_e, \boldsymbol{\Sigma}_u)\\ &\quad = \sum_{i=1}^{N} \sum_{k=1}^{m} p(I_i = k \mid \mathbf{y})(\boldsymbol{\Delta}_{ik} \otimes \mathbf{I}_r) \cdot \boldsymbol{\Sigma}_e^{-1}[\mathbf{y}_i - (\boldsymbol{\Delta}_{ik} \otimes \mathbf{I}_r)' \boldsymbol{\theta}]\\ &\qquad - (\mathbf{A}^{-1} \otimes \boldsymbol{\Sigma}_u^{-1})\, \mathbf{u}\end{aligned} \tag{37}$$

where $p(I_i = k \mid \mathbf{y})$ is obtained by using the formulae given in Appendix 3 and replacing the densities f(.) by $f_r(.)$.

With the data ordered by individual and the parameters in $\boldsymbol{\theta}$ by level, the following system of equations is obtained:

$$\left[\begin{pmatrix} \mathbf{D}^{[l]} & \mathbf{Q}'^{[l]}\mathbf{X} & \mathbf{Q}'^{[l]}\mathbf{Z} \\ \mathbf{Q}'^{[l]}\mathbf{X} & \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Q}'^{[l]}\mathbf{Z} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{pmatrix} \otimes \mathbf{I}_r + \begin{pmatrix} \mathbf{0}_{r(m+p) \times r(m+p)} & \mathbf{0}_{r(m+p) \times rq} \\ \mathbf{0}_{rq \times r(m+p)} & \mathbf{A}^{-1} \otimes \boldsymbol{\Sigma}_e \boldsymbol{\Sigma}_u^{-1} \end{pmatrix}\right] \begin{bmatrix} \hat{\mathbf{g}} \\ \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix}^{[l+1]} = \begin{bmatrix} (\mathbf{Q}'^{[l]} \otimes \mathbf{I}_r)\mathbf{y} \\ (\mathbf{X}' \otimes \mathbf{I}_r)\mathbf{y} \\ (\mathbf{Z}' \otimes \mathbf{I}_r)\mathbf{y} \end{bmatrix}. \tag{38}$$

This approach can be extended to other situations of multitrait analysis of continuous and discontinuous data (e.g., Foulley et al. 1982; Hoeschele et al. 1986; Foulley et al. 1987b).

Application to simulated data

Generation of the data

Phenotypic values were generated by using a mixed model including herd-year-season effect (hys_i), major genotype (g_j), polygenic effect (u_k) and residual (e_{ijkl}), so that

$$y_{ijkl} = hys_i + g_j + u_k + e_{ijkl}. \tag{39}$$

Dispersion assumptions were:

$$\operatorname{var}(y_{ijkl}) = \sigma^2 = \sigma_{hys}^2 + \sum_{j=1}^{m} p_j (g_j - \mu_g)^2 + \sigma_u^2 + \sigma_e^2,$$

$\{u_k\} \sim N\left(\mathbf{0}, \mathbf{I}\frac{\sigma_u^2}{4}\right)$ if the u_k's were sire effects (sire model),

$\{u_k\} \sim N(\mathbf{0}, \mathbf{A}\,\sigma_u^2)$ if the u_k's were breeding values (animal model),

$\{hys_i\} \sim N(\mathbf{0}, \mathbf{I}\,\sigma_h^2)$ and $\{e_{ijkl}\} \sim N(\mathbf{0}, \mathbf{I}\,\sigma_e^2)$.

Discontinuous phenotypes were obtained by using (39) and

$$Y_{ijkl} = \begin{cases} 1, & \text{if } \ y_{ijkl} < \Phi^{-1}(0.7) \\ 0, & \text{otherwise.} \end{cases} \tag{40}$$

where 0.7 is the frequency in the category coded by 1.

The following parameter values were used:

Data set	p(A)	σ^2	σ_{hys}^2	σ_{ML}^2	σ_u^2	σ_e^2	h²%	t/2
(1)	0.3	50^2	25^2	$22.4^2 = 0.2\,\sigma^2$	11.2^2	35.4^2	25	35.4
(2)	0.3	50^2	25^2	$18.0^2 = 0.13\,\sigma^2$	17.3^2	35.4^2	25	27.8

Table 1. Estimates of the major genotype frequencies ($\hat{p}$) and their empirical standard errors ($s_{\hat{p}}$)

Major genotype	True values	Data set[a]							
		I		II		III		IV	
		$\hat{p}$	$s_{\hat{p}}$	$\hat{p}$	$s_{\hat{p}}$	$\hat{p}$	$s_{\hat{p}}$	$\hat{p}$	$s_{\hat{p}}$
AA≡1	0.09	0.074	0.010	0.064	0.005	0.082	0.011	0.125	0.031
Aa ≡2	0.42	0.412	0.017	0.397	0.007	0.412	0.015	0.401	0.026
aa ≡3	0.49	0.514	0.027	0.539	0.011	0.506	0.025	0.474	0.090

[a] Estimates are averages of 10 replicates per data set

Table 2. MAPE and BLUE estimates of the major genotypic values ($\hat{g}$) and their empirical standard errors ($s_{\hat{g}}$)

Method	Major genotype	Data set[a]							
		I[b]		II[c]		III[b]		IV[b]	
		$\hat{g}$	$s_{\hat{g}}$	$\hat{g}$	$s_{\hat{g}}$	$\hat{g}$	$s_{\hat{g}}$	$\hat{g}$	$s_{\hat{g}}$
MAPE	AA≡1	477.4	18.9	462.9	13.5	481.1	18.2	470.0	18.7
	Aa ≡2	446.4	17.0	446.4	13.6	445.7	20.2	442.5	20.6
	aa ≡3	417.9	20.5	427.3	13.9	413.3	19.4	426.7	21.6
BLUE	AA≡1	486.6	19.7	479.0	15.1	485.2	19.6	472.4	16.8
	Aa ≡2	451.9	19.5	451.7	13.2	448.2	19.7	438.5	18.3
	aa ≡3	414.8	19.9	422.8	14.3	412.9	19.4	405.9	19.5

[a] Estimates are averages of 10 replicates per data set

[b] True values are $g_1=486$, $g_2=450$ and $g_3=414$

[c] True values are $g_1=478$, $g_2=450$ and $g_3=422$

where p(A) is the frequency of the allele A at the major locus with alleles A and a, h^2 is heritability with $h^2 = (\sigma^2_{ML} + \sigma^2_u)/(\sigma^2_{ML} + \sigma^2_u + \sigma^2_h + \sigma^2_e)$, σ^2_{ML} is the variance accounted for by the major locus and $t = g_{AA} - g_{aa}$ is the displacement effect computed by assuming additive gene action.

Using (39), 2500 progeny records in 200 herd-year-seasons and representing 50 sires were simulated. Also, a small sample of 150 parent and progeny records was generated. The design was unbalanced in both cases. Five data sets were generated with ten replicates each. The data sets are described below:

Data set	Phenotypes	Sample size	Model	Unknown genotype membership in %	$\sigma^2_{ML}/\sigma^2 \times 100\%$
I	continuous	2500	sire model	100%	20%
II	continuous	2500	sire model	100%	13%
III	continuous	2500	sire model	50%	20%
IV	continuous	150	animal model	100%	20%
V	binary	2500	sire model	100%	20%

Results

Major genotypic values, frequencies and polygenic sire effects or breeding values, respectively, were estimated by maximizing the joint posterior density via (16) and (17) using formulae (A 3.1) and (A 3.3) in Appendix 3, for sire models (data sets, I, II, III, V) and (A 3.10) for data set IV. This approach is referred to as Maximum-A-Posteriori Estimation (MAPE). In Tables 1–4, results averaged over ten replicates are reported for data sets I, II and III only for the approximate formula (A 3.3). With (A 3.1), taking into account dependencies between genotype memberships of offspring of the same sire, convergence was very slow or not attained, probably because of increased multimodality of the posterior distribution in (A2.1). For iteration with (16) and (17), a set of starting values obtained by using techniques described by Hoeschele (1988) was employed. Estimates of the major genotype frequencies are presented in Table 1. With that particular set of starting values, the estimates were quite close to the true values, even for the small sample IV. The best estimates were obtained from data set III with 50% known major genotype membership.

Estimates of the major genotypic values are reported in Table 2. For control, BLUE estimates for known genotype membership were also computed. The true displacement effect: $t = g_1 - g_3 = 72.0$ was estimated precisely by BLUE, whereas with 100% unknown genotype membership, t was underestimated by MAPE for about 17% – 40%. With 50% unknown genotype membership, t was only slightly underestimated.

Estimation of polygenic breeding values was evaluated by comparing realized genetic response achieved with fully unknown (MAPE) and fully known (BLUE) major genotype membership. Realized genetic response was computed as the mean true breeding value of the candidates selected according to MAPE and BLUP estimates. Table 3 shows that genetic response was reduced by unknown genotype membershisp.

Table 3. Realized genetic responses (ū) from selection using MAPE and BLUP estimates of polygenic effects and their empirical standard errors $s_{\bar{u}}$

Method	% of candidates selected	Data set[a]			
		I		II	
		ū	$s_{\bar{u}}$	ū	$s_{\bar{u}}$
MAPE	10	6.42[ns]	1.85	12.34[ns]	1.91
	20	4.21[ns]	1.35	9.63[ns]	1.52
BLUP	10	7.09	1.51	10.57	1.81
	20	5.79	1.37	7.62	1.50

[a] Estimates are averages of 10 replicates per data set.
[ns]: Difference between MAPE and BLUP not significant at $\alpha = 0.05$.

For data set I, estimates of **p** and **g** and mean true breeding value of the 10% highest ranking sires obtained from the solutions to (16) and (17) at convergence by using different sets of starting values are presented in Table 4. Estimates of the major genotype frequencies were mostly affected by the choice of starting values. If the starting values for **p** were far away from the true values, so were the estimates, and this, of course, also influenced the estimates of **g**. Realized genetic response was practically unaffected by the starting values. Hoeschele (1988) described techniques to obtain a set of starting values for fully unknown genotype membership.

With data set I, variance components were estimated. For the sire model, true values were $\sigma_s^2 = \frac{1}{4}\sigma_u^2 = 31.4$, $\sigma_e^{2*} = \frac{3}{4}\sigma_s^2 + \sigma_e^2 = 1347.2$, and $\sigma_{ML}^2 = 500.0$ Average estimates from (26), (28) and (31) were $\sigma_s^2 = 65.3$, $\hat{\sigma}_e^{2*} = 1366.0$ and $\hat{\sigma}_{ML}^2 = 336.6$, indicating that polygenic variance tends to be overestimated when major genotype membership is fully unknown.

Two iterative schemes, (16) and Newton's method, were suggested to obtain estimates of location parameters maximizing the posterior density. With data set I, the set of starting values described by Hoeschele (1988) and the stopping rule $\sqrt{\Delta'\Delta \,|\, \dim(\boldsymbol{\theta})} < 10^{-4}$, (16) required 12–20 (5–8) iterates. In 50% of the replicates, convergence was not attained with Newton's method because iteration was stopped after round 30 or if the matrix $[-\partial^2 F/\partial\gamma\,\partial\gamma']^{-1}_{\gamma=\hat{\gamma}^{[1]}}$ was not positive definite. Poor performance of Newton's method when mixture components are not well separated has been reported by Titterington et al. (1985).

Conclusions

The method proposed in this paper attempts to combine aspects of mixed model methodology and complex segre-

Table 4. Effect of different sets of starting values on the MAPE estimates of the major genotype frequencies (p̂), effects (ĝ) and realized genetic response (ū)

Starting values[a]	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_3$	ū
[1A, 2A, 3A]	0.100	0.448	0.452	177.0	450.9	425.7	6.53
[1A, 2B, 3A]	0.035	0.461	0.504	486.8	455.7	426.6	6.53
[1A, 2C, 3A]	0.066	0.419	0.515	481.0	454.3	427.3	6.53
[1A, 2D, 3A]	0.342	0.347	0.311	462.5	439.6	421.5	6.53
[1B, 2A, 3A]	0.100	0.418	0.482	477.1	451.4	426.9	6.53
[1C, 2A, 3A]	0.100	0.609	0.291	478.1	446.4	420.5	6.53
[1A, 2A, 3B]	0.082	0.427	0.490	478.6	452.8	426.7	6.53

[a] Triplet with starting values $[p^0, g^0, (\beta, u)^0]$
1A: $g^{0\prime} = [486.0, 450.0, 414.0]$, 1B: $g^{0\prime} = [468.0, 450.0, 432.0]$, 1C: $g^{0\prime} = [550.0, 450.0, 350.0]$,
2A: $p^{0\prime} = [0.09, 0.42, 0.49]$, 2B: $p^{0\prime} = [0.03, 0.43, 0.54]$, 2C: $p^{0\prime} = [0.058, 0.388, 0.554]$, 2D: $p^{0\prime} = [0.33, 0.33, 0.34]$,
3A: BLUP of u and BLUE of β ignoring the major genotype, 3B: $\beta^0 = 0$, $u^0 = 0$

gation analysis. The BLUE of $\mathbf{g}$ and BLUP of $\mathbf{u}$ (Henderson 1973) cannot be computed because parts of the $\mathbf{X}$ matrix (relating elements in $\mathbf{g}$ and $\mathbf{y}$) are unknown, and the dispersion assumption $\operatorname{var}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$ known with $\mathbf{R} = \mathbf{I}\,\sigma_e^2$ does not hold because of [4]. Complex segregation analysis (Morton and McLean 1974; Bonney 1984, 1986) does not account for or distinguish separately polygenic parameters and is computable only for small pedigrees. Distinction between major locus and polygenic parameters and multitrait analyses are particularly useful when major genotypic effects are beneficial on some and harmful on other traits of interest.

Results of the simulation study indicate that the proposed method can provide useful estimates of major locus and polygenic parameters, although with fully unknown genotype membership the precision of estimation is clearly reduced and will deteriorate with a degree of dominance not equal to 1/2 (implying the genotypic mean of the heterozygote being closer to that of the dominant homozygote) and a displacement effect smaller than considered here. Precision of estimation also depends on the accuracy of approximating the marginal posterior means of $\mathbf{g}$, $\boldsymbol{\beta}$ and $\mathbf{u}$ considered as the best estimators (Gianola et al. 1986) by their joint posterior mode. Multimodality of the surface of the posterior distribution in (A 1.3), however, poses a difficult problem. The results obtained with sire models (data sets I, II and III) suggest that it is probably better to ignore dependencies between major genotype memberships. As an alternative, one could consider computing the marginal posterior mean $E(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\sigma})$. If the distribution of the data conditional on major genotype membership is normal, $E(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\sigma}) = \Sigma\, E(\boldsymbol{\theta} \mid \mathbf{I} = \mathbf{K}, \mathbf{y}, \boldsymbol{\sigma})\; f(\mathbf{I} = \mathbf{K} \mid \mathbf{y}, \boldsymbol{\sigma})$. Also, $E(\boldsymbol{\theta} \mid \mathbf{I} = \mathbf{K}, \mathbf{y}, \boldsymbol{\sigma}) = \hat{\boldsymbol{\theta}}$ is the vector of solutions to the mixed model equations (Henderson 1973) given $\mathbf{I} = \mathbf{K}$, and $f(\mathbf{I} = \mathbf{K} \mid \mathbf{y}, \boldsymbol{\sigma})$ would have to be approximated by $f(\mathbf{I} = \mathbf{K} \mid \hat{\boldsymbol{\theta}}, \mathbf{y}, \boldsymbol{\sigma})$, so that iteration would be required. However, for application to large field data sets as considered here, computation of $E(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\sigma})$ is probably not feasible because of the large dimension of $\mathbf{K}$ and the necessity to solve the mixed model equations many times.

If the data are categorial with threshold character (Falconer 1965), the likelihood function in (A 1.3) with normal components is no longer adequate. Gianola and Foulley (1983) and Foulley et al. (1987a) suggested a method of estimating breeding values with categorial data based on the polygenic model of inheritance. In their approach, estimates of environmental and genetic effects in a hypothetical normal scale are obtained by maximizing the posterior density proportional to a product bi-, multinomial, or poisson likelihood and the normal prior density in (7). This method was extended to mixed major gene and polygenic inheritance by specifying the likelihood as a mixture density of m bi- or multinomial components. The posterior density was maximized with respect to the unknown effects by using Fisher's scoring or the Newton-Raphson procedure (Gianola and Foulley 1983). However, the bad performance of these algorithms experienced with continuous data was observed with categorial data (data set V) and fully unknown major genotype membership to an even larger extent. In the majority of trials, convergence was extremely slow or was not achieved because of a nonpositive definite inverted coefficient matrix. These difficulties probably could be overcome by using alternative algorithms such as the method of steepest descent or robust nonlinear regression algorithms (Kennedy and Gentle 1980).

For continuous data, the method assumes that the data are normally distributed conditionally on genotype membership and the parameters. Departures from normality would require transformations such as the Box-Cox approach or estimation of the power transformation (Gianola et al. 1988). Analysis of field-collected data requires adjusting for a large number of environmental differences, nonadditive genetic effects and selection. This can be achieved by an appropriate choice of the mixed model in (1) and by making inferences from a posterior distribution under certain conditions unaffected by selection (Gianola and Fernando 1986). In this situation, the proposed method and approaches suggested by Hoeschele (1988) may be potentially useful for discriminating between purely polygenic and mixed major gene and polygenic inheritance and for estimating genetic parameters and breeding values under fully or partly unknown genotype membership.

Acknowledgements. This research was conducted while I. Hoeschele was a visiting scientist at Iowa State University on leave from the Institute of Animal Breeding and Animal Husbandry, Hohenheim University, Stuttgart, West Germany. Financial support from the Deutsche Forschungsgemeinschaft and from the Department of Animal Science, Iowa State University, is gratefully acknowledged. Dr. R. L. Fernando, Dr. M. F. Rothschild, Dr. J. L. Foulley and Dr. D. Gianola are thanked for comments on the manuscript.

Appendix 1

Consider a vector $\mathbf{y}$ that consists only of records on one unrelated sire-dam pair (y_S, y_D) and their n_{SD} progeny $(y_1, \ldots, y_{n_{SD}})$:

$$h(y_S, y_D, y_1, \ldots, y_{n_{SD}} \mid \boldsymbol{\theta}, \mathbf{p}\,\sigma_e^2) = \sum_{k_S=1}^{m} \sum_{k_D=1}^{m} \sum_{\mathbf{K}_{SD}} f(I_S = k_S, I_D = k_D, \mathbf{I}_{SD} = \mathbf{K}_{SD}, y_S, y_D, y_1, \ldots, y_{n_{SD}} \mid \boldsymbol{\theta}, \sigma_e^2) \qquad (\text{A}\,1.1)$$

with $\mathbf{I}_{SD} = \mathbf{K}_{SD}$ standing for $I_1 = k_1 \ldots, I_{n_{SD}} = k_{n_{SD}}$. Based on the assumptions stated, the density in the right-hand side of (A 1.1) becomes

$$\begin{aligned}
&\mathrm{f}(\mathrm{I_S}=\mathrm{k_S}, \mathrm{y_S}, \mathrm{I_D}=\mathrm{k_D}, \mathrm{y_D}, \mathbf{I}_{\mathrm{SD}}=\mathbf{K}_{\mathrm{SD}}, \mathrm{y}_1, \ldots, \mathrm{y}_{\mathrm{n_{SD}}} \mid \boldsymbol{\theta}, \sigma_e^2)\\
&\quad = \mathrm{f}(\mathrm{y_S} \mid \boldsymbol{\theta}, \sigma_e^2)\, \mathrm{f}(\mathrm{y_D} \mid \boldsymbol{\theta}, \sigma_e^2)\, \mathrm{f}(\mathrm{I_S}=\mathrm{k_S} \mid \mathrm{y_S})\, \mathrm{f}(\mathrm{I_D}=\mathrm{k_D} \mid \mathrm{y_D})\\
&\qquad \cdot \mathrm{f}(\mathbf{I}_{\mathrm{SD}}=\mathbf{K}_{\mathrm{SD}} \mid \mathrm{I_S}=\mathrm{k_S}, \mathrm{I_D}=\mathrm{k_D})\\
&\qquad \cdot \mathrm{f}(\mathrm{y}_1, \ldots, \mathrm{y}_{\mathrm{n_{SD}}} \mid \mathbf{I}_{\mathrm{SD}}=\mathbf{K}_{\mathrm{SD}}, \boldsymbol{\theta}, \sigma_e^2).
\end{aligned}$$

Using this and (5) in (A 1.1) gives, after rearrangement

$$\begin{aligned}
&\mathrm{h}(\mathrm{y_S}, \mathrm{y_D}, \mathrm{y}_1, \ldots, \mathrm{y}_{\mathrm{n_{SD}}} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2)\\
&\quad = \sum_{\mathrm{k_S}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_S}=\mathrm{k_S})\, \mathrm{f}(\mathrm{y_S} \mid \mathrm{I_S}=\mathrm{k_S}, \boldsymbol{\theta}, \sigma_e^2)\\
&\qquad \cdot \sum_{\mathrm{k_D}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_D}=\mathrm{k_D})\, \mathrm{f}(\mathrm{y_D} \mid \mathrm{I_D}=\mathrm{k_D}, \boldsymbol{\theta}, \sigma_e^2)\\
&\qquad \cdot \sum_{\mathrm{k_S}=1}^{\mathrm{m}} \sum_{\mathrm{k_D}=1}^{\mathrm{m}} \{\mathrm{p}(\mathrm{I_S}=\mathrm{k_S} \mid \mathrm{y_S})\, \mathrm{p}(\mathrm{I_D}=\mathrm{k_D} \mid \mathrm{y_D})\\
&\qquad \cdot \prod_{\mathrm{i}=1}^{\mathrm{n_{SD}}} \sum_{\mathrm{k_i}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_i}=\mathrm{k_i} \mid \mathrm{I_S}=\mathrm{k_S}, \mathrm{I_D}=\mathrm{k_D})\, \mathrm{f}(\mathrm{y_i} \mid \mathrm{I_i}=\mathrm{k_i}, \boldsymbol{\theta}, \sigma_e^2)\} \qquad \text{(A 1.2)}
\end{aligned}$$

where {.} indicates nested summation.

If all individuals appearing as parents but not as progeny in the data are unrelated, (A 1.2) generalizes to (A 1.3):

$$\begin{aligned}
&\mathrm{h}(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2)\\
&\quad = \prod_{\mathrm{S}=1}^{N_\mathrm{S}} \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_S}=\mathrm{k})\, \mathrm{f}(\mathrm{y_S} \mid \mathrm{I_S}=\mathrm{k}, \boldsymbol{\theta}, \sigma_e^2)\\
&\qquad \cdot \prod_{\mathrm{D}=1}^{N_\mathrm{D}} \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_D}=\mathrm{k})\, \mathrm{f}(\mathrm{y_D} \mid \mathrm{I_D}=\mathrm{k}, \boldsymbol{\theta}, \sigma_e^2)\\
&\qquad \cdot \sum_{\mathbf{K}_\mathrm{S}} \sum_{\mathbf{K}_\mathrm{D}} \left\{ \prod_{\mathrm{S}=1}^{N_\mathrm{S}} \mathrm{p}(\mathrm{I_S}=\mathrm{k_S} \mid \mathrm{y_S}) \prod_{\mathrm{D}=1}^{N_\mathrm{D}} \mathrm{p}(\mathrm{I_D}=\mathrm{k_D} \mid \mathrm{y_D}) \right.\\
&\qquad\qquad \cdot \prod_{\mathrm{i}=1}^{N_\mathrm{p}} \sum_{\mathrm{k_i}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_i}=\mathrm{k_i} \mid \mathrm{I_{S(i)}}=\mathrm{k_{S(i)}}, \mathrm{I_{D(i)}}=\mathrm{k_{D(i)}})\\
&\qquad\qquad \left. \cdot \mathrm{f}(\mathrm{y_i} \mid \mathrm{I_i}=\mathrm{k_i}, \boldsymbol{\theta}, \sigma_e^2) \right\} \qquad \text{(A 1.3)}
\end{aligned}$$

where N_S, N_D and N_p are the numbers of sires, dams and progeny, respectively, $\mathrm{S_{(i)}}$ [$\mathrm{D_{(i)}}$] denotes the sire [dam] of the $\mathrm{k^{th}}$ progeny, the $\mathrm{p}(\mathrm{I_i}=\mathrm{k_i} \mid \mathrm{I_{S(i)}}=\mathrm{k_{S(i)}}, \mathrm{I_{D(i)}}=\mathrm{k_{D(i)}})$ values are known constants found in the genetic transition matrix (Elston and Stewart 1971), and {.} again indicates nested summation. In particular cases, some of which are shown in Appendix 2, the general likelihood (A 1.3) can be simplified.

Appendix 2

1) Sire model with $N_{\mathrm{p(S)}}$ records available only on progeny of sire S assuming each offspring has a different dam.

$$\begin{aligned}
\mathrm{h}_1(\mathbf{y} \mid \boldsymbol{\theta}, \sigma_e^2) &= \prod_{\mathrm{S}=1}^{\mathrm{N_S}} \sum_{\mathrm{k_S}=1}^{\mathrm{m}} \left\{ \mathrm{p}(\mathrm{I_S}=\mathrm{k_S}) \right. \qquad \text{(A 2.1)}\\
&\quad \left. \cdot \prod_{\mathrm{i}=1}^{N_{\mathrm{p(S)}}} \sum_{\mathrm{k_i}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_i}=\mathrm{k_i} \mid \mathrm{I_{S(i)}}=\mathrm{k_{S(i)}})\, \mathrm{f}(\mathrm{y_i} \mid \mathrm{I_i}=\mathrm{k_i}, \boldsymbol{\theta}, \sigma_e^2) \right\}.
\end{aligned}$$

An approximation to (A 2.1) would be to pretend that the probabilities of major genotype memberships of individuals having the same sire are independent:

$$\mathrm{h}_1(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2) \doteq \prod_{\mathrm{i}=1}^{N_\mathrm{p}} \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_i}=\mathrm{k})\, \mathrm{f}(\mathrm{y_i} \mid \mathrm{I_i}=\mathrm{k}, \boldsymbol{\theta}, \sigma_e^2). \qquad \text{(A 2.2)}$$

2) Records are available on parents and N progeny, each sire is mated only to one dam, and N matings produce only one offspring each.

$$\begin{aligned}
&\mathrm{h}_2(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2)\\
&\quad = \prod_{\mathrm{S}=1}^{N} \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_S}=\mathrm{k})\, \mathrm{f}(\mathrm{y_S} \mid \mathrm{I_S}=\mathrm{k}, \boldsymbol{\theta}, \sigma_e^2)\\
&\qquad \cdot \prod_{\mathrm{D}=1}^{N} \sum_{\mathrm{l}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_D}=\mathrm{l})\, \mathrm{f}(\mathrm{y_D} \mid \mathrm{I_D}=\mathrm{l}, \boldsymbol{\theta}, \sigma_e^2) \qquad \text{(A 2.3)}\\
&\qquad \cdot \prod_{\mathrm{i}=1}^{N} \sum_{\mathrm{r}=1}^{\mathrm{m}} \left[\sum_{\mathrm{k}=1}^{\mathrm{m}} \sum_{\mathrm{l}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_{S(i)}}=\mathrm{k} \mid \mathrm{y_{S(i)}})\, \mathrm{p}(\mathrm{I_{D(i)}}=\mathrm{l} \mid \mathrm{y_{D(i)}}) \right.\\
&\qquad\qquad \left. \cdot \mathrm{p}(\mathrm{I_i}=\mathrm{r} \mid \mathrm{I_{S(i)}}=\mathrm{k}, \mathrm{I_{D(i)}}=\mathrm{l}) \right] \mathrm{f}(\mathrm{y_i} \mid \mathrm{I_i}=\mathrm{r}, \boldsymbol{\theta}, \sigma_e^2).
\end{aligned}$$

3) As (2), but each of N matings can produce several (N_i) offspring.

$$\begin{aligned}
&\mathrm{h}_3(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2)\\
&= \prod_{\mathrm{S}=1}^{N} \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_S}=\mathrm{k})\, \mathrm{f}(\mathrm{y_S} \mid \mathrm{k}, \boldsymbol{\theta}, \sigma_e^2)\\
&\quad \cdot \prod_{\mathrm{D}=1}^{N} \sum_{\mathrm{l}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_D}=\mathrm{l})\, \mathrm{f}(\mathrm{y_D} \mid \mathrm{I_D}=\mathrm{l}, \ \boldsymbol{\theta}, \sigma_e^2) \qquad \text{(A 2.4)}\\
&\quad \cdot \prod_{\mathrm{i}=1}^{N} \left[\sum_{\mathrm{k}}^{\mathrm{m}} \sum_{\mathrm{l}}^{\mathrm{m}} \mathrm{p}(\mathrm{I_{S(i)}}=\mathrm{k} \mid \mathrm{y_{S(i)}})\, \mathrm{p}(\mathrm{I_{D(i)}}=\mathrm{l} \mid \mathrm{y_{D(i)}}) \right.\\
&\qquad \left. \cdot \prod_{\mathrm{j}=1}^{N_\mathrm{i}} \sum_{\mathrm{r}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_{ij}}=\mathrm{r} \mid \mathrm{I_{S(i)}}=\mathrm{k}, \mathrm{I_{D(i)}}=\mathrm{l})\, \mathrm{f}(\mathrm{y_{ij}} \mid \mathrm{I_{ij}}=\mathrm{r}, \boldsymbol{\theta}, \sigma_e^2) \right].
\end{aligned}$$

4) Records are available on parents and progeny, dams have one offspring and are nested within sires.

$$\begin{aligned}
&\mathrm{h}_4(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2)\\
&\quad = \prod_{\mathrm{S}=1}^{N_\mathrm{S}} \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_S}=\mathrm{k})\, \mathrm{f}(\mathrm{y_S} \mid \mathrm{I_S}=\mathrm{k}, \boldsymbol{\theta}, \sigma_e^2)\\
&\qquad \cdot \prod_{\mathrm{D}=1}^{N_\mathrm{D}} \sum_{\mathrm{l}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_D}=\mathrm{l})\, \mathrm{f}(\mathrm{y_D} \mid \mathrm{I_D}=\mathrm{l}, \boldsymbol{\theta}, \sigma_e^2) \qquad \text{(A 2.5)}\\
&\qquad \cdot \prod_{\mathrm{S}=1}^{N_\mathrm{S}} \left\{ \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_S}=\mathrm{k} \mid \mathrm{y_s}) \prod_{\mathrm{i}=\mathrm{D}=1}^{N_{\mathrm{D(S)}}} \left[\sum_{\mathrm{r}=1}^{\mathrm{m}} \sum_{\mathrm{l}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_D}=\mathrm{l} \mid \mathrm{y_D}) \right. \right.\\
&\qquad\qquad \left. \left. \cdot \mathrm{p}(\mathrm{I_i}=\mathrm{r} \mid \mathrm{I_S}=\mathrm{k}, \mathrm{I_D}=\mathrm{l}) \right] \mathrm{f}(\mathrm{y_i} \mid \mathrm{I_i}=\mathrm{r}, \boldsymbol{\theta}, \sigma_e^2) \right\}.
\end{aligned}$$

Appendix 3

First derivatives of log (A 2.1)

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}} \log \mathrm{h}_1(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{p}, \sigma_e^2) \qquad \text{(A 3.1)}\\
&\quad = \sum_{\mathrm{S}=1}^{N_\mathrm{S}} \sum_{\mathrm{i}=1}^{N_{\mathrm{p(S)}}} \sum_{\mathrm{r}=1}^{\mathrm{m}} \mathrm{p}(\mathrm{I_{Si}}=\mathrm{r} \mid \mathbf{y})(\mathrm{y_{Si}} - \Delta'_{\mathrm{Si,r}} \boldsymbol{\theta})\, \sigma_e^{-2} \Delta_{\mathrm{Si,r}}
\end{aligned}$$

where

$$p(I_{Si}=r|\mathbf{y})=\sum_{k=1}^{m}\left\{\frac{p(I_S=k)\prod_{i=1}^{N_{p(S)}}\sum_{r=1}^{m}p(I_{Si}=r|I_S=k)\,f(y_{Si}|I_{Si}=r,\boldsymbol{\theta},\sigma_e^2)\,p(I_{Si}=r|I_S=k)\,f(y_{Si}|I_{Si}=r,\boldsymbol{\theta},\sigma_e^2)}{\sum_{k=1}^{m}p(I_S=k)\prod_{i=1}^{N_{p(S)}}\sum_{r=1}^{m}p(I_{Si}=r|I_S=k)\,f(y_{Si}|I_{Si}=r,\boldsymbol{\theta},\sigma_e^2)\sum_{r=1}^{m}p(I_{Si}=r|I_s=k)\,f(y_{Si}|I_{Si}=r,\boldsymbol{\theta},\sigma_e^2)}\right\}$$

and $\Delta'_{Si,r}=[\mathbf{w}'_r:\mathbf{x}'_{Si}:\mathbf{z}'_{Si}]$. (A 3.2)

First derivatives of log (A 2.2):

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log h_1(\mathbf{y}|\boldsymbol{\theta},\mathbf{p},\sigma_e^2)=\sum_{i=1}^{N}\sum_{k=1}^{m}p(I_i=k|\mathbf{y})(y_i-\Delta'_{ik}\boldsymbol{\theta})\,\sigma_e^{-2}\Delta_{ik} \tag{A 3.3}$$

where

$$N=\sum_{S=1}^{N_S}N_{p(S)}\quad\text{and}\quad p(I_i=k|\mathbf{y})=\frac{p(I_i=k)\,f(y_i|I_i=k,\boldsymbol{\theta},\sigma_e^2)}{\sum_{k=1}^{m}p(I_i=k)\,f(y_i|I_i=k,\boldsymbol{\theta},\sigma_e^2)} \tag{A 3.4}$$

First derivatives of log (A 2.3):

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}\log h_2(\mathbf{y}|\boldsymbol{\theta},\mathbf{p},\sigma_e^2)&=\sum_{S=1}^{N}\sum_{k=1}^{m}p(I_S=k|\mathbf{y})(y_S-\Delta'_{Sk}\boldsymbol{\theta})\,\sigma_e^{-2}\Delta_{Sk}+\sum_{D=1}^{N}\sum_{l=1}^{m}p(I_D=l|\mathbf{y})(y_D-\Delta'_{Dl}\boldsymbol{\theta})\,\sigma_e^{-2}\Delta_{Dl}\\&\quad+\sum_{i=1}^{N}\sum_{r=1}^{m}p(I_i=r|\mathbf{y})(y_i-\Delta'_{ir}\boldsymbol{\theta})\,\sigma_e^{-2}\Delta_{ir}\end{aligned} \tag{A 3.5}$$

where

$$p(I_S=k|\mathbf{y})=\frac{p(I_S=k)\,f(y_S|I_S=k,\boldsymbol{\theta},\sigma_e^2)}{\sum_{k=1}^{m}p(I_S=k)\,f(y_S|I_S=k,\boldsymbol{\theta},\sigma_e^2)}, \tag{A 3.6}$$

$p(I_D=l|\mathbf{y})$ is defined analogously, and

$$p(I_i=r|\mathbf{y})=\sum_{k=1}^{m}\sum_{l=1}^{m}\frac{p(I_S=k|y_S)\,p(I_D=l|y_D)\,p(I_i=r|I_S=k,I_D=l)\,f(y_i|I_i=r,\boldsymbol{\theta},\sigma_e^2)}{\sum_{r=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m}p(I_S=k|y_S)\,p(I_D=l|y_D)\,p(I_i=r|I_S=k,I_D=l)\,f(y_i|I_i=r,\boldsymbol{\theta},\sigma_e^2)} \tag{A 3.7}$$

First derivatives of log (A 2.4):

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log h_3(\mathbf{y}|\boldsymbol{\theta},\mathbf{p},\sigma_e^2)=\ldots+\ldots+\sum_{i=1}^{N}\sum_{j=1}^{N_i}\sum_{r=1}^{m}p(I_{ij}=r|\mathbf{y})(y_{ij}-\Delta'_{ijr}\boldsymbol{\theta})\sigma_e^{-2}\Delta_{ijr} \tag{A 3.8}$$

where the first two parts are as in (A 3.5), and

$$\begin{aligned}&p(I_{ij}=r|\mathbf{y})\\&=\sum_{k=1}^{m}\sum_{l=1}^{m}\left\{\frac{p(I_{S(i)}=k|y_{S(i)})\,p(I_{D(i)}=l|y_{D(i)})\left[\prod_{j=1}^{N_i}\sum_{r=1}^{m}p_{r.kl}f(y_{ij}|I_{ij}=r,\boldsymbol{\theta},\sigma_e^2)\right]p_{r.kl}f(y_{ij}|I_{ij}=r,\boldsymbol{\theta},\sigma_e^2)}{\sum_{k=1}^{m}\sum_{l=1}^{m}\left[p(I_{S(i)}=k|y_{S(i)})\,p(I_{D(i)}=l|y_{D(i)})\left[\prod_{j=1}^{N_i}\sum_{r=1}^{m}p_{r.kl}f(y_{ij}|I_{ij}=r,\boldsymbol{\theta},\sigma_e^2)\right]\sum_{r=1}^{m}p_{r.kl}f(y_{ij}|I_{ij}=r,\boldsymbol{\theta},\sigma_e^2)\right]}\right\}\end{aligned} \tag{A 3.9}$$

with $p_{r.kl}=p(I_{ij}=r|I_{S(i)}=k,I_{D(i)}=l)$.

First derivatives of (A 2.5):

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log h_4(\mathbf{y}|\boldsymbol{\theta},\mathbf{p},\sigma_e^2)=\ldots+\ldots+\sum_{S=1}^{N_S}\sum_{i=D=1}^{N_{D(S)}}\sum_{r=1}^{m}p(I_i=r|\mathbf{y})(y_i-\Delta'_{ir}\boldsymbol{\theta})\,\sigma_e^{-2}\Delta_{ir} \tag{A 3.10}$$

where

$$\mathrm{p}(\mathrm{I}_i=\mathrm{r}\,|\,\mathrm{y})=\sum_{k=1}^{m}\left\{\frac{\mathrm{p}(\mathrm{I}_\mathrm{S}=\mathrm{k}\,|\,\mathrm{y}_\mathrm{S})\prod_{i=D=1}^{N_{D(S)}}\sum_{l=1}^{m}\sum_{r=1}^{m}\mathrm{p}(\mathrm{I}_\mathrm{D}=\mathrm{l}\,|\,\mathrm{y}_\mathrm{D})\,\mathrm{p}_{\mathrm{r.kl}}\,\mathrm{f}(\mathrm{y}_i\,|\,\mathrm{I}_i=\mathrm{r},\boldsymbol{\theta},\sigma_e^2)}{\sum_{k=1}^{m}\mathrm{p}(\mathrm{I}_\mathrm{S}=\mathrm{k}\,|\,\mathrm{y}_\mathrm{S})\sum_{i=D=1}^{N_{D(S)}}\sum_{l=1}^{m}\sum_{r=1}^{m}\mathrm{p}(\mathrm{I}_\mathrm{D}=\mathrm{l}\,|\,\mathrm{y}_\mathrm{D})\,\mathrm{p}_{\mathrm{r.kl}}\,\mathrm{f}(\mathrm{y}_i\,|\,\mathrm{I}_i=\mathrm{r},\boldsymbol{\theta},\sigma_e^2)}\right.$$
$$\left.\cdot\sum_{l=1}^{m}\frac{\mathrm{p}(\mathrm{I}_\mathrm{D}=\mathrm{l}\,|\,\mathrm{y}_\mathrm{D})\,\mathrm{p}_{\mathrm{r.kl}}\,\mathrm{f}(\mathrm{y}_i\,|\,\mathrm{I}_i=\mathrm{r}_i\,\boldsymbol{\theta},\sigma_e^2)}{\sum_{l=1}^{m}\sum_{r=1}^{m}\mathrm{p}(\mathrm{I}_\mathrm{D}=\mathrm{l}\,|\,\mathrm{y}_\mathrm{D})\,\mathrm{p}_{\mathrm{r.kl}}\,\mathrm{f}(\mathrm{y}_i\,|\,\mathrm{I}_i=\mathrm{r},\boldsymbol{\theta},\sigma_e^2)}\right\}\qquad\text{(A 3.11)}$$

A suitable expression for $\mathrm{p}(\mathrm{I}_\mathrm{S}=\mathrm{k}\,|\,\mathrm{y}_\mathrm{S})$ and $\mathrm{p}(\mathrm{I}_\mathrm{D}=\mathrm{l}\,|\,\mathrm{y}_\mathrm{D})$ required in (A 3.7), (A 3.9) and (A 3.11) is that obtained in (A 3.6).

References

Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer, Berlin Heidelberg New York Tokyo, 217 pp

Bonney GE (1984) On the statistical determination of major gene mechanisms in continuous human traits: Regressive models. Am J Med Genet 18: 731

Bonney GE (1986) Regressive logistic models for familial disease and other binary traits. Biometrics 42:611

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading, Mass, 588 pp

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523

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

Foulley JL, Calomiti S, Gianola D (1982) Ecriture des èquations du BLUP multicaractères. Genet Sel Evol 14:309

Foulley JL, Gianola D, Planchenault D (1986) Sire evaluation with uncertain paternity. Genet Sel Evol 19:83

Foulley JL, Gianola D, Im S (1987 a) Genetic evaluation of traits distributed as poisson binomial with reference to reproductive characters. Theor Appl Genet 17:870

Foulley JL, Gianola D, Im S, Hoeschele I (1987 b) Empirical Bayes estimation of parameters for n polygenic binary traits. Genet Sel Evol 19:197

Geysser S (1982) Bayesian Discrimination. In: Krishnaiah R, Kananl L (eds) Handbook of statistics, vol 2. Amsterdam, North-Holland, pp 101–120

Gianola D, Foulley JL (1983) Sire evaluation of ordered categorial data with a threshold model. Genet Sel Evol 15:201

Gianola D, Fernando RL (1986) Bayesian methods in animal breeding theory. J Anim Sci 63:217

Gianola D, Foulley JL, Fernando RL (1986) Prediction of breeding values when variances are not known. In: Dickerson GE, Johnson RK (eds) Proc 3rd World Congr Genet Appl Liv Prod Agric Commun. University of Nebraska, Lincoln, Neb, XII:356

Gianola D, Im S, Fernando RL, Foulley JL (1988) Mixed model methodology and the Box-Cox theory of transformations: A Bayesian approach. In: Gianola D, Hammond K (eds) Advances in statistical methods for genetic improvement of livestock. Springer, Berlin Heidelberg New York (in press)

Hanset R (1982) Major genes in animal production, examples and perspectives: cattle and pigs. In: 2nd World Congr Genet Appl Liv Prod, vol VI. Publicaciones Agrarias, Madrid (Spain), p 439

Harville DA, Mee RW (1984) A mixed model procedure for analyzing ordered categorial data. Biometrics 40:393

Henderson CR (1973) Sire evaluation and genetic trends. Proc Anim Breed Genet Symp in honor of Dr JL Lush, Blacksburg, Virginia, July 29, 1972, 10–41, ASAS-ADSA. Champaign, Ill

Hoeschele I (1988) Statistical techniques for detection of major genes in animal breeding data. Submitted to Theor Appl Genet

Hoeschele I, Foulley JL, Colleau JJ, Gianola D (1986) Genetic evaluation for multiple binary responses. Genet Sel Evol 18:299

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

Kennedy NJ, Gentle JE (1980) Statistical computing. Marcel Dekker, New York Basel, 591 pp

Morton NE, McLean CJ (1974) Analysis of family resemblance. III. Complex segregation of quantitative traits. Am J Hum Genet 26:489

Roberts RC, Smith C (1982) Genes with large effects – theoretical aspects in livestock breeding. In: 2nd World Congr Genet Appl Liv Prod, vol VI. Publicaciones Agrarias, Madrid (Spain), p 420

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

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley and Sons, Chichester New York Brisbane Toronto Singapore, 243 pp