GENETIC ANALYSIS OF ANIMAL BREEDING DATA COMBINING MAJOR GENE AND POLYGENIC INHERITANCE

Ina HOESCHELE

Department of Dairy Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061, U.S.A.

SUMMARY

Presented are statistical procedures for detection of major loci and for genetic evaluation with field data combining major gene and polygenic inheritance when the genotype membership of some or all individuals is unknown. For detection, a graphical technique and testing genetic hypotheses based on a mixed model accounting for environmental, additive, and nonadditive polygenic effects are proposed. For genetic evaluation, a Bayes procedure is used to estimate major genotypic effects, frequencies, polygenic effects and heritability within major genotype. Results from simulation and application to muscular hypertrophy in cattle are presented.

Key Words: Major genes, modifying genes, genetic evaluation, mixture data, Bayesian inference
1. Introduction

Animal breeding theory for quantitative traits is based on the polygenic model of inheritance. It postulates the breeding value of an individual to be the sum of small gene effects at many loci. Although this model approximately holds for many traits, genes with large effects on some traits, major genes, exist (e.g., Hanset, 1982). Examples are the muscular hypertrophy gene in cattle and pigs, the dwarf gene in cattle, the Booroola gene in sheep, and the rapid post-weaning growth gene in mice. Typically in field data, phenotypic variation is not entirely determined by the major genotypes, but arises mainly from environmental differences, and additive and nonadditive effects of polygenes, also called modifying genes. Often, the major genotype of an individual is unknown. This makes it difficult to exploit polygenic variation within major genotype by selection. Hence, more sophisticated techniques for detection of major genes, and for estimation of major genotypic effects, polygenic effects, and polygenic variance and heritability within major genotype based on mixed model analysis (Henderson, 1973) are required.

2. Methodology

2.1 Model for the data

Consider the mixed linear model

\[ y_{ik} = g_k + x_i^\beta + z_i^u + e_{ik} \]  \[ 1 \]

where \( y_{ik} \) is an observation on the \( i^{th} \) individual, \( g_k \) is the \( k^{th} \) major genotypic mean with \( k=1, \ldots, m \), \( \beta \) is a \( t \times 1 \) vector of systematic environmental factors, \( u \) is a \( q \times 1 \) vector of polygenic effects,
\(x'_i\) and \(z'_i\) are the \(i\)th rows of the incidence matrices \(X\) and \(Z\), respectively, and \(e_{ik}\) is a residual with \(\text{var}(e_{ik}) = \sigma^2_e\). The vector \(u\) may be partitioned into additive direct, maternal and nonadditive effects with \(u \sim N(0, G)\). If \(u\) is a vector of additive effects, then \(G = A\sigma^2_u\) with \(A\) being a matrix of additive genetic relationships, and \(\sigma^2_u\) represents additive genetic variance.

Denote major genotype membership by \(I_i = k, k \in \{1, 2, \ldots, m\}\). As an example, at a major locus with two alleles \(B\) and \(b\), an individual can have major genotype membership \(BB(I_i = 1), Bb(I_i = 2)\) or \(bb(I_i = 3)\). If the major genotype membership \(I_i\) is unknown, \(y_i\) has an \(m\)-component mixture distribution, conditional on \(\theta' = [g', \beta', u']\), with mean

\[
E_h(y_i \mid p, \theta, \sigma^2_e) = \sum_{k=1}^{m} p(I_i = k)g_k + x'_i\beta + z'_iu
\]

and variance

\[
\text{var}_h(y_i \mid p, \theta, \sigma^2_e) = \sum_{k=1}^{m} p(I_i = k)(g_k - \mu_g)^2 + \sigma^2_e
\]

where \(p\) is an \(m \times 1\) vector with elements \(p(I_i = k)\), the probabilities that the \(i\)th individual has major genotype membership \(k \in \{1, 2, \ldots, m\}\), \(\mu_g = \sum_{k=1}^{m} p(I_i = k)g_k\). Also, \(h\) denotes expectation and variance with respect to the density

\[
h(y_i \mid p, \theta, \sigma^2_e) = \sum_{k=1}^{m} p(I_i = k) f(y_i \mid I_i = k, \theta, \sigma^2_e). \tag{4}
\]

In [4], \(h(.)\) is an \(m\)-component mixture density, and \(f(.)\) is a component density (Titterington et al., 1985) with mean and variance

\[
y_i \mid I_i = k, \theta, \sigma^2_e \sim (g_k + x'_i\beta + z'_iu, \sigma^2_e)
\]

2.2 Graphical technique for detection of major loci

A variety of graphical techniques dealing with mixture data have been developed. For mixtures, the normal quantile-quantile (Q-Q) plot
has been found very competitive with respect to sensitivity to
departures from normality, and a numerical procedure based on the plot
can be devised to estimate component means (major genotypic means) and
mixing weights (major genotypic frequencies). Let $F^{-1}(P)$ and $\Phi^{-1}(P)$
denote the inverse cumulative distribution functions (c.d.f.) of $v$ and
the standard normal deviate $x$, and let $P$ be the cumulative probability
($0 \leq P \leq 1$). Plotting the mixture quantiles $F^{-1}(P) = v$ against the
standard normal quantiles $\Phi^{-1}(P) = x$ gives a characteristic $s$-shaped
configuration (Fowlkes, 1979) due to the nonlinear relationship $\partial v/\partial x$.

Because of the superposition of systematic environmental and
polygenic factors, the observed data cannot be used as mixture
quantiles to construct the Q-Q plot (Hoeschele, 1988a). Hence we
fit the linear mixed model ignoring major genotypes: $y = XB + Zu + e$.

Best Linear Unbiased Predictions (Henderson, 1973) of $u_i$ (\hat{u}_i), are
obtained for continuous data. Estimates of $u_i$ on an underlying scale
can be obtained from categorical data by using the method of Gianola
and Foulley (1983). The quantities $v_i = \hat{u}_i / \sigma_u$ are used as
mixture quantiles to construct the Q-Q plot. These quantities are
dependent, but transformation to uncorrelated variables would affect
the configuration of the plot. Often, there are three genotypes at a
major locus. Therefore, the following function was used to fit the
Q-Q plot configuration of a three-component mixture:

$$
v = b_1 + b_2 x + \frac{1}{b_3 - e} - \frac{b_4 (b_6 - x)}{b_5 (x - b_7)} + \frac{1}{b_8 + e}
$$

In [5], the $v's$ ($x's$) are the mixture (standard normal) quantiles, and
the $b's$ are unknown parameters to be estimated with a nonlinear
leastsquares algorithm.

Setting second derivatives of [5] w.r.t. x equal to zero gives a polynomial in x with roots \( x^*_1 \) and \( x^*_2 \). Using Harding (1949), estimates of the mixing weights (here, genotype frequencies) are

\[
P_1 = \Phi(x^*_1), \quad P_2 = \Phi(x^*_2) - P_1 \quad \text{and} \quad P_3 = 1 - P_2 - P_1.
\]

These estimates are employed to partition the ordered set of v's into three subsets, from which component means (\( u_k^* \), \( k=1,2,3 \)) and variances (\( \sigma^2_k \)) are estimated. Crude estimates of the major genotypic means are

\[
\hat{s}_k = \hat{\mu} + \hat{\mu}_k \sigma_u, \quad \text{where} \quad \mu \text{ is an overall mean.}
\]

2.3 Genetic evaluation

2.3.1 Estimation of major locus parameters and polygenic effects

Bayesian inference uses prior information on \( \Theta \) and \( p \) by specifying normal prior distributions for \( g, \beta, \) and \( u \), and a Dirichlet distribution for \( p \). If prior knowledge on \( g, \beta, \) and \( p \) is vague, \( \text{var}(u) = A \sigma^2_u \), and \( \sigma^2_u \) and \( \sigma^2_e \) are known, the prior density can be written as

\[
\ell(\Theta, p) = \ell(g) \ell(\beta) \ell(u | \sigma^2_u) \ell(p) = C \exp\left(-\frac{1}{2} u' A^{-1} u \sigma^{-2} \right). \tag{6}
\]

Using [4], the joint likelihood of all observations is

\[
h(y | \Theta, p, \sigma^2_e) = \sum_{I=1}^{N} p(I=K) f(y | I=K, \Theta, \sigma^2_e) \tag{7}
\]

where \( K \) is a particular \( N \times 1 \) vector of major genotype memberships, \( N \) is the total number of observations, the summation \( \Sigma \) represents a nested sum of the form \( \Sigma_{k_1=1}^{m} \Sigma_{k_2=1}^{m} \Sigma_{k_3=1}^{m} \), and \( p(I=K) \) is the joint probability of \( N \) particular genotype memberships in \( K \). Specifications of this likelihood are given in the appendix and in Hoeschele (1988b).
The joint posterior density of the location parameters $\theta$ and the major genotype frequencies $p$ with $\sigma = [\sigma_u^2, \sigma_e^2]$ is

$$h(\theta, p | y, \sigma) = \text{constant } h(y | \theta, p, \sigma^2)e^{\sum_{u} \log \sigma_u^2} e^{\log p}.$$  

[8]

Inferences about $\theta$ and $p$ can be obtained from [8]. An alternative approach is to consider the marginal posterior densities of $\theta$ and $p$,

$$t(\theta | y, \sigma)$$

and

$$t(p | y, \sigma).$$

Because of the large order of $\theta$, the mode of the joint or marginal posterior distributions of $\theta$ and $p$ is chosen as a point estimator, and is found by solving the nonlinear equations

$$
\begin{bmatrix}
D^{[\ell]} & Q^{[\ell]}
\end{bmatrix}
\begin{bmatrix}
X'Q^{[\ell]}
\end{bmatrix}
\begin{bmatrix}
X'Z
\end{bmatrix}
\begin{bmatrix}
6^{[\ell+1]}
\end{bmatrix}
= 
\begin{bmatrix}
Q^{[\ell]}y
\end{bmatrix}
\begin{bmatrix}
X'y
\end{bmatrix}
\begin{bmatrix}
Z'y
\end{bmatrix}
[9]
$$

where $D^{[\ell]} = \text{Diag} \left( \sum_{i=1}^{N} p^{[\ell]}(I_1=k | y) \right)_{m \times m}$, $Q^{[\ell]} = (p^{[\ell]}(I_1=k | y))_{N \times m}$, and the form of the posterior probabilities of genotype membership

$p(I_1=k | y)$ depends on whether [9] is derived from the joint or marginal posterior densities of $\theta$ and $p$. Estimates $\hat{p}^{[\ell]}$ are obtained from

$$\hat{p}^{[\ell]}(I_1=k) = \frac{1}{N} \sum_{i=1}^{N} p^{[\ell]}(I_1=k | y) \quad k = 1, \ldots, m.$$  

[10]


### 2.3.2 Estimation of variance components and heritability

When variances are unknown, estimation of $\theta$ and $p$ requires accommodation of the unknown parameters $\sigma$. Also, interest may be directly in estimating polygenic variance and heritability within major genotype. Estimation of $\sigma$ will be by Marginal Maximum Likelihood.
(MML; Harville and Mee, 1984; Hoeschele et al., 1987). It consists of finding the mode of the marginal posterior density of variance components by employing flat prior densities:

\[
\log t(\sigma | y) = \mathbb{E} \left[ h(\sigma, \theta, p, y) \right] \quad \text{[11]}
\]

\[\sigma_i^2 \]

where \(\sigma_i^2 \equiv \sigma_u^2 \) or \(\sigma_e^2 \). Because \(h(\theta, p, y, \sigma)\) is not in the form of a normal density and \(\dim(\theta)\) is usually large, we employ the approximations:

\[\hat{\theta} | y, \sigma \sim (\hat{\theta}_h, C_h) \quad \text{or} \quad \hat{\theta} | y, \sigma \sim (\hat{\theta}_t, C_t)\]

where \(\hat{\theta}_h \) and \(C_h \) \(\hat{\theta}_t \) and \(C_t \) are solution and inverted coefficient matrix of [9] at convergence derived from the joint posterior \(h(\theta, p, y, \sigma)\) [marginal posterior \(t(\theta | y, \sigma)\)]. Resulting estimators are:

\[
\hat{\sigma}_u^2(\lambda+1) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{m} \hat{e}_{ik}^2}{N - \dim(\theta) + \text{tr}(A^{-1}[C_u]_\lambda)}
\]

\[
\hat{\sigma}_e^2(\lambda+1) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{m} \hat{e}_{ik}^2}{N - \dim(\theta) + \text{tr}(A^{-1}[C_u]_\lambda)}
\]

where \(\hat{e}_{ik} = y_i - \hat{\theta}_k \), \(\Delta \equiv \{x^\prime, z^\prime \}, \hat{\theta} \) \(\text{\hat{\theta}_h} \) or \(\hat{\theta}_t \), respectively, \(\hat{\theta}_u \) and \(C_u \) are either \(\hat{\theta}_h \) and \(C_u(h) \) or \(\hat{\theta}_t \) and \(C_u(t) \), \(C_u \) is the qxq part of \(C \) referring to \(u \), and \(\lambda[\lambda] = \hat{\sigma}_e^2(\lambda) / \hat{\sigma}_e^2 \). Estimation of \(\theta \) and \(p \), and \(\sigma \) proceeds by iterating jointly with [9], [10] and [12].

2.3.3 Testing of genetic hypotheses

We wish to test the null hypothesis \(H_0: \bar{g}_1 = \bar{g}_2 = \ldots = \bar{g}_m \) (no major locus segregating) versus the alternative \(H_A: \bar{g}_k \neq \bar{g}_k \) for all \(k \neq k \in \{1, \ldots, m\} \). Under \(H_0 \), the vector of unknown parameters is \(\gamma_0 = [u, \bar{\beta}', u'] \), where \(u \) is the overall mean. Under \(H_A \), the vector of
unknown parameters is \( y'_A = \{p', g', \beta', u'\} \). In the Bayesian framework, a criterion for discriminating between \( H_0 \) and \( H_A \) is the ratio of "averaged likelihoods", with the prior density serving as a weight function (Zellner, 1971). This is

\[
\Lambda_{1a} = \frac{\sup_{u, \beta} \int_{R_u} h(y|u, \beta, \sigma^2_e) \frac{f(u|\sigma^2)}{\sigma^2_u} du}{\sup_{p, g, \beta} \int_{R_u} h(y|p, g, \beta, \sigma^2_e) \frac{f(u|\sigma^2)}{\sigma^2_u} du}
\]

Assuming equal prior probabilities of the hypotheses, i.e., \( p(H_0) = p(H_A) = .5 \), and equal losses for accepting \( H_0(H_A) \) when \( H_A(H_0) \) is true, \( H_0 \) is rejected if \( \Lambda_{1a} < 1 \). Maximizing the denominator of (13) numerically over the range of \( p, g \) and \( \beta \) becomes difficult when the order of \( y, \beta \) and \( u \) is large, and we might approximate [13] by

\[
\Lambda_{1a}^* = \frac{\sup_{u, \beta, u} h(y|u, \beta, \sigma^2_e) \frac{f(u|\sigma^2)}{\sigma^2_u}}{\sup_{p, g, \beta, u} h(y|p, g, \beta, \sigma^2_e) \frac{f(u|\sigma^2)}{\sigma^2_u}}
\]

The approximation [14] will be close to \( \Lambda_{1a} \) only if most of the probability mass of the likelihood is concentrated over a small region about the mode. Outside the Bayesian approach, hypothesis testing would employ an approximate chi-square distribution of \(-2 \log \Lambda_{1a} \).

This does not hold when some elements of \( p \) are zero, i.e., on the boundary of the parameter space. Then, \( p \) should be integrated out of the denominator of [13] by specifying a Dirichlet prior, or transformed to \( p = t(p) \) such that \( t(0) \neq a \neq b \) for \( p \in [a, b] \).
3. Application to simulated data

Phenotypes (y) were generated by a mixed model including herd-year-season effect (hys), major genotype (g), polygenic effect (u), and residual (e).

The following parameter sets were used:

<table>
<thead>
<tr>
<th>Parameter set</th>
<th>p(B)</th>
<th>$\sigma_y^2$</th>
<th>$\sigma_{hys}^2$</th>
<th>$\sigma_g^2$</th>
<th>$\sigma_u^2$</th>
<th>$\sigma_e^2$</th>
<th>$h^2%$</th>
<th>t/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS 1</td>
<td>.3</td>
<td>50^2</td>
<td>25^2</td>
<td>22.4^2</td>
<td>11.2^2</td>
<td>35.4^2</td>
<td>25</td>
<td>35.4</td>
</tr>
<tr>
<td>PS 2</td>
<td>.3</td>
<td>50^2</td>
<td>25^2</td>
<td>18.0^2</td>
<td>17.3^2</td>
<td>35.4^2</td>
<td>25</td>
<td>27.8</td>
</tr>
<tr>
<td>PS 3</td>
<td>.3</td>
<td>46.7^2</td>
<td>25^2</td>
<td>0.0</td>
<td>17.3^2</td>
<td>35.4^2</td>
<td>14</td>
<td>0.0</td>
</tr>
</tbody>
</table>

where p(B) is the frequency of allele B at a major locus with two alleles B and b, heritability $h^2 = (\sigma_g^2 + \sigma_u^2)/(\sigma_{hys}^2 + \sigma_g^2 + \sigma_u^2 + \sigma_e^2)$, and $t = g_{BB} - g_{bb}$, assuming additive gene action. Three data sets (I,II,III) were generated by a sire model and PS 1, 2, and 3, including 5000 records on 100 sires in 400 hys. Data set IV was generated with an animal model, PS 1, and consisted of 150 records. Each design was unbalanced and replicated ten times.

The Q-Q plots showed evidence for the presence of major genotypes when compared to the control parameter set 3. Crude estimates of major genotypic means and frequencies (used as starting values in the Bayes estimation) were obtained from the plot and are shown in Table 1.
Bayes estimates were computed based on [8], and are shown in Tables 2 and 3. Estimates of genotype frequencies were close to the true values but were found highly dependent on the starting values when using values different from those obtained from the Q-Q plot. The displacement effect $t = g_1 - g_3$ was underestimated, however, less severely when 50 percent of the genotype memberships were known.

**Table 2**: Estimates of the major genotype frequencies ($\hat{p}$) and their empirical standard errors ($s^p_\hat{p}$)

<table>
<thead>
<tr>
<th>Major Genotype</th>
<th>True Values</th>
<th>Data Set</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p$</td>
<td>$s^p_\hat{p}$</td>
<td>$p$</td>
<td>$s^p_\hat{p}$</td>
<td>$p$</td>
</tr>
<tr>
<td>AA = 1</td>
<td>.09</td>
<td>.074</td>
<td>.010</td>
<td>.064</td>
<td>.005</td>
</tr>
<tr>
<td>Aa = 2</td>
<td>.42</td>
<td>.412</td>
<td>.017</td>
<td>.397</td>
<td>.007</td>
</tr>
</tbody>
</table>
Table 3. Estimates of the major genotypic values (g) and their empirical standard errors (s_g)

<table>
<thead>
<tr>
<th>Major genotype</th>
<th>Data Set</th>
<th>Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I a)</td>
<td>II b)</td>
</tr>
<tr>
<td>g</td>
<td>ĝ</td>
<td>ĝ</td>
</tr>
<tr>
<td>s_g</td>
<td>s_ĝ</td>
<td>s_ĝ</td>
</tr>
<tr>
<td>AA = 1</td>
<td>477.4</td>
<td>18.9</td>
</tr>
<tr>
<td>Aa = 2</td>
<td>446.4</td>
<td>17.0</td>
</tr>
<tr>
<td>aa = 3</td>
<td>417.9</td>
<td>20.5</td>
</tr>
</tbody>
</table>

a) True values are g_1 = 486, g_2 = 450 and g_3 = 414
b) True values are g_1 = 478, g_2 = 450 and g_3 = 422
c) 50% of major genotype memberships known.

Estimation of polygenic effects was evaluated by comparing mean true polygenic effect of selected sires u̅ with unknown and known major genotype memberships. No significant differences were found (Table 4).

Table 4. Realized genetic response (u̅) from selection on estimates of polygenic effects computed under unknown (U) and known (K) genotype membership

<table>
<thead>
<tr>
<th>Genotype membership</th>
<th>% of candidates selected</th>
<th>Data Set</th>
<th>Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>ĝ</td>
<td>s_ĝ</td>
<td>ĝ</td>
<td>s_ĝ</td>
</tr>
<tr>
<td>U</td>
<td>10%</td>
<td>6.42 ns</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>4.21 ns</td>
<td>1.35</td>
</tr>
<tr>
<td>K</td>
<td>10%</td>
<td>7.09</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>5.79</td>
<td>1.37</td>
</tr>
</tbody>
</table>

ns: difference not significant at α = .05
With data set I and II, variance components were estimated. True values and estimates are presented in Table 5. The results show that polygenic variance tends to be overestimated when major genotype membership is unknown.

Table 5. Estimation of variance components within major genotype

<table>
<thead>
<tr>
<th>Data Set I</th>
<th>Data Set II</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_g^2$</td>
<td>$\sigma_e^2$</td>
</tr>
<tr>
<td>True Values</td>
<td>31.4</td>
</tr>
<tr>
<td>Estimates</td>
<td>65.3</td>
</tr>
</tbody>
</table>

The hypothesis of a segregating major locus with 3 genotypes was examined using [14]. The test criterion was $-\log \Delta_{1a}^*$ which should be close to zero in absence of a major locus. Except for the control data set III, the values in Table 6 indicate evidence for a segregating major locus.

Table 6. Approximate criterion for discriminating between polygenic and mixed inheritance.

<table>
<thead>
<tr>
<th>Data set</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-\log \Delta_{1a}^*$</td>
<td>59.3</td>
<td>10.1</td>
<td>1.1</td>
<td>76.3</td>
</tr>
</tbody>
</table>
4. Application to field data

In the Belgian White and Blue cattle, muscular hypertrophy is caused by a recessive major gene. Two genotypes can be distinguished: the double-muscled (mh/mh) and the conventional or dual-purpose type (mh/+; +/+). Hanset and Michaux (1985) stated: "...within the double-muscled class, there is significant genetic variation in the degree of fleshiness which could be exploited by selection ... A reliable method of identification of the 3 genotypes is needed." Selection within the double-muscled type could allow correction for undesirable side effects (calving difficulty, viability...). The usefulness of the presented method for this purpose is demonstrated using a data set on 45 AI sires of the types mh/mh and mh/+ mated to dual-purpose cows and having 10 female progeny each, scored for muscling on a 50–150 point scale at the age of one year.

A mixed model including sire, district, year and feeding level was fitted. The Q-Q plot revealed that about 11 of the 45 sires were mh/+. An adequate specification of the likelihood for this design lead to estimation equations [11]. Variance components were estimated using [14]. Major genotypic means were 84.2, 87.1, and 126.7 points. Estimates of residual and genetic variances were $\sigma^2_e = (5.84 \text{ points})^2$ and $\sigma^2_u = (5.87 \text{ points})^2$ implying a within major genotype heritability of $h^2 = .5$. For each sire, its polygenic effect was estimated, and the posterior probability of major genotypic membership was computed using formula A6 in the Appendix. For 11 sires, the posterior probability of having genotype mh/+ exceeded 65%, and was less or equal to 50% for the others. The maximum difference between
5. Conclusions

With unknown major genotype membership, the BLUE of g, the BLUP of \( \tilde{u} \) and the REML estimate of \( \tilde{g} \) cannot be computed, because parts of the \( X \) matrix are unknown. Foulley et al. (1986) dealt with unknown \( Z \) in the context of uncertain paternity. Because interest is primarily in \( \tilde{u} \) and var(\( \tilde{u} \)), and the number of animals, environmental, and non-additive or maternal effects that need to be accounted for is large, complex segregation analysis (Morton and McLean, 1974; Bonney 1986) cannot be used. An approximate Bayes approach was developed because due to the size of the problem, exact posterior means

\[
E(\theta | y, \sigma) = \sum_{K} E(\theta | I=K, y, \sigma) f(I=K|y, \sigma)
\]

often cannot be computed.

Multimodality of the surface of the posterior density requires "good" starting values for finding the mode \( M(\theta | y, \sigma) \) used as an approximation to \( E(\theta | y, \sigma) \), and the degree of approximation will limit the value of the method. Simulation results suggest that the method will be potentially useful if the major locus accounts for at least ten percent of phenotypic variance.

As shown by Foulley and Elsen (1988), this method can also be used to compute posterior probabilities of genotype memberships of individuals.

Under strong linkage disequilibrium and close linkage, the approach could incorporate known genetic markers linked to a quantitative trait locus by writing the unconditional probabilities of genotype membership as functions of recombination frequency.
Extensions to multitrait analyses, nonnormal data and interaction of genotypes with environmental or polygenes are not discussed for reasons of space.

REFERENCES


APPENDIX I

(1) Sire model with $N_p(S)$ records available only on progeny of sire $S$ assuming each offspring has a different dam.

$$h_1(y|\theta, \beta, \sigma_e^2) = \prod_{S=1}^{N_S} \sum_{k_S=1}^{m} p(I_S=k_S) \sum_{i=1}^{N_p(S)} \sum_{k_i=1}^{m} p(I_i^s=k_i, I_S^i=k_S^i) f(y_i^s|I_i^s=k_i, \theta, \sigma_e^2) \} \quad [A1]$$

An approximation to $[A1]$ would be to pretend that the probabilities of major genotype memberships of individuals having the same sire are independent:

$$h_1(y|\theta, \beta, \sigma_e^2) \approx \prod_{S=1}^{N_S} \sum_{k_S=1}^{m} p(I_S=k_S) \sum_{i=1}^{N_p(S)} \sum_{k_i=1}^{m} p(I_i^s=k_i) f(y_i^s|I_i^s=k_i, \theta, \sigma_e^2). \quad [A2]$$

(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.

$$h_2(y|\theta, \beta, \sigma_e^2)$$

$$= \prod_{S=1}^{N_S} \sum_{k_S=1}^{m} p(I_S=k_S) f(y_S|I_S=k_S, \theta, \sigma_e^2) \prod_{D=1}^{N_D} \sum_{\ell_D=1}^{m} p(I_D=\ell_D) f(y_D|I_D=\ell_D, \theta, \sigma_e^2)$$

$$= \prod_{S=1}^{N_S} \sum_{k_S=1}^{m} \sum_{r=1}^{N} \sum_{k=1}^{m} \sum_{\ell=1}^{m} [ \sum_{i=1}^{N_p(S)} \sum_{r=1}^{N} \sum_{k=1}^{m} \sum_{\ell=1}^{m} p(I_S^i=k_i, I_D^i=\ell_i) p(I_D^i=k_D^i) f(y_i^s|I_i^s=r_i, \theta, \sigma_e^2) \] \quad [A3]$$
(3) As (2), but each of $N$ matings can produce several ($N_i^1$) offspring.

$$h_3 (\gamma|\theta, \Omega, \sigma^2_e)$$

$$= \prod_{S=1}^N \sum_{k=1}^m p(I_S=k) f(y_S|k, \theta, \sigma^2_e) \prod_{D=1}^N \sum_{\ell=1}^m p(I_D=\ell) f(y_D|I_D=\ell, \theta, \sigma^2_e)$$

$$= \prod_{i=1}^{N_S} \sum_{k=1}^m \sum_{\ell=1}^m p(I_S(i)=k|y_S(i)) p(I_D(i)=\ell|y_D(i)) \prod_{j=1}^{N_i^1} \sum_{r=1}^m p(I_{ij}=r|I_S(i)=k, I_D(i)=\ell)$$

$$f(y_{ij}|I_{ij} = r, \theta, \sigma^2_e)$$ \[ A4 \]

(4) Records are available on parents and progeny, dams have one offspring and

are nested within sires.

$$h_4 (\gamma|\theta, \Omega, \sigma^2_e)$$

$$= \prod_{S=1}^{N_S} \sum_{k=1}^m p(I_S=k) f(y_S|I_S=k, \theta, \sigma^2_e) \prod_{D=1}^{N_D} \sum_{\ell=1}^m p(I_D=\ell) f(y_D|I_D=\ell, \theta, \sigma^2_e)$$

$$= \prod_{i=1}^{N_S} p(I_S=k) \prod_{i=1}^{N_D} p(I_D=\ell) \prod_{r=1}^m p(I_{ij}=r|I_S=k, I_D=\ell)$$

$$f(y_{ij}|I_{ij} = r, \theta, \sigma^2_e)$$ \[ A5 \]

Posterior probability that sire $S$ has major genotype membership $k$, based on model 1:

$$P(I_S = k|y) = \frac{\sum_{i=1}^{m} p(I_S = k) \sum_{r=1}^{m} p(I_{Si} = r|I_S = k) f(y_{Si}|I_{Si} = r, \theta, \sigma^2_e)}{\sum_{k=1}^{N_S} \sum_{i=1}^{m} p(I_S = k) \sum_{r=1}^{m} p(I_{Si} = r|I_S = k) f(y_{Si}|I_{Si} = r, \theta, \sigma^2_e)}$$ \[ A6 \]
APPENDIX II: Discussion following presentation

It was pointed out that the presented methodology will only be useful if the desired major allele has not already been fixed in the population through continued selection. The examples (e.g., muscular hypertrophy gene, Booroola gene) show that fixation is not readily achieved or desirable if a considerable amount of phenotypic variation is explained by other sources (environment, modifying genes), if there is no reliable biochemical or other method for determining major genotype memberships, and if the major allele has desirable on some and undesirable effects on other traits of economic importance. In this context, it was asked if the method could be extended to multitrait estimation and accommodate unknown major genotype memberships.

Extension to multitrait estimation has been described in Hoeschele (1988b), and the method accounts for partly (some animals) or fully (all animals) unknown major genotype membership.

Another point mentioned in the discussion was that the method as presented assumes gametic phase equilibrium. Genetic variance was partitioned into variance attributed to the major locus and polygenic variance. Selection, however, will introduce gametic phase disequilibrium. It could not be answered with certainty to what degree gametic phase disequilibrium introducing (negative) genetic covariance terms between loci, would affect the partitioning of genetic variance, i.e., whether it can truly be neglected.

Relative to a question concerning convergence rate when estimating polygenic and residual variance components within major genotype, no severe convergence problems were observed. Because extremely slow convergence and the possibility of a premature termination of iteration arise in the EM-algorithm, an alternative first-derivative algorithm with faster convergence rate was employed.
The remaining questions dealt with further refinements and extensions of the method including heterogeneity of polygenic variance with respect to major genotype membership, and the presence of two linked major loci. Heterogeneity of polygenic variance can be accommodated by estimating polygenic effects (Foulley et al., 1988) and their variances within major genotype membership

\[
(u' = [u_1', u_2', \ldots, u_m']) , \sigma_u = [\sigma_{u_{11}}, \sigma_{u_{12}}, \ldots, \sigma_{u_{22}}, \sigma_{u_{23}}, \ldots, \sigma_{u_{mm}}]
\]

with appropriate modification of [9] and [12]. This, however, will considerably increase the size of the system of equations in [9], and should only be implemented if a considerable amount of interaction between major locus and modifying genes is expected. Extension to two linked major loci (or inclusion of a known marker locus) is theoretically possible, expressing probabilities of joint genotype memberships as functions of the (unknown) recombination frequency. However, the number of possible genotype memberships increases, and, under fully unknown genotype membership with respect to both loci, estimated probabilities of major genotype membership will reflect an increased degree of uncertainty about the true genotype memberships.

Although the method has been primarily developed for estimating major genotypic effects, polygenic variance, heritability, and polygenic effects for selection on several traits of economic importance, when a major locus (with fully or partly unknown genotype membership) is segregating and polygenic variation is present, application to rather different problems was suggested. For example, in the dairy industry, it is likely that growth hormone treatment will soon be used to further increase milk production. Sometimes, the treatment level (e.g., treated versus not treated) will be unknown. Adaptation of the method presented would make it possible to include records with unknown treatment level rather than discarding them.