DOI: 10.1534/genetics.105.054197

On the Quantitative Genetics of Mixture Characters

Daniel Gianola,*,†,[1] Bjorg Heringstad† and Jorgen Odegaard†

**Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53706 and †Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway*

Manuscript received December 1, 2005
Accepted for publication April 14, 2006

ABSTRACT

Finite mixture models are helpful for uncovering heterogeneity due to hidden structure. Quantitative genetics issues of continuous characters having a finite mixture of Gaussian components as statistical distribution are explored in this article. The partition of variance in a mixture, the covariance between relatives under the supposition of an additive genetic model, and the offspring–parent regression are derived. Formulas for assessing the effect of mass selection operating on a mixture are given. Expressions for the genetic and phenotypic correlations between mixture and Gaussian traits and between two mixture traits are presented. It is found that, if there is heterogeneity in a population at the genetic or environmental level, then genetic parameters based on theory treating distributions as homogeneous can lead to misleading interpretations. Some peculiarities of mixture characters are: heritability depends on the mean values of the component distributions, the offspring–parent regression is nonlinear, and genetic or phenotypic correlations cannot be interpreted devoid of the mixture proportions and of the parameters of the distributions mixed.

FINITE mixture models, used in biology and in genetics since Pearson (1894), are helpful for uncovering heterogeneity due to hidden structure or incorrect assumptions. For instance, unknown loci with major effects can create "bumps" (sometimes quite subtle) in a phenotypic distribution, and this type of heterogeneity may be resolved by fitting a mixture, *i.e.*, by calculating conditional probabilities that a datum is drawn from one of the several potential, yet unknown, genotypes. A brief review of the potential usefulness of mixtures for uncovering major genes is in Lynch and Walsh (1998). Also, many quantitative trait loci detection procedures are based on ideas from mixture models (Haley and Knott 1992).

The quantitative genetics of characters distributed as mixtures has not been studied extensively, although the idea underlies work of, *e.g.*, Latter (1965) and Kimura and Crow (1978). Perhaps this is due to that, until recently, fitting complex hierarchical mixture models to phenotypic data was computationally difficult. However, inference about some quantitative genetic characters via finite mixture models may be warranted in practice. For example, consider mastitis, an inflammation of the mammary gland of cows and goats associated with bacterial infection. The disease affects the dairy industry globally, and it has severe economic effects. Genetic variation in susceptibility to mastitis exists, and selection for increased resistance is feasible (Heringstad *et al.* 2000). However, recording of mastitis events is not routine in most nations, and milk somatic cell counts (SCC) have been used as a proxy in genetic evaluation of sires (using mixed-effects linear models), because an elevation of SCC is associated with mastitis. It is not obvious how the SCC information should be treated optimally in genetic evaluation. SCC is both an indicator of mastitis and a measure of response to infection. It is reasonable to expect that SCC observations taken on healthy and diseased animals display different distributions, which are "hidden" in the absence of disease recording. Finite mixture models have been suggested in this context by Detilleux and Leroy (2000), Ødegård *et al.* (2003, 2005), Gianola *et al.* (2004), and Boettcher *et al.* (2005).

This article explores quantitative genetics issues of continuous characters having a finite mixture of Gaussian components as statistical distribution. model introduces notation, gives a specification in which both genetic and residual effects follow mixtures, and derives pertinent marginal and conditional distributions. truncation selection gives formulas for assessing the effect of mass selection operating on a mixture. Next, covariance between relatives presents the partition of variance in a mixture, the calculation of covariance between relatives under the supposition of an additive genetic model, and illustrates the effect of heterogeneity on the offspring–parent regression. Genetic and phenotypic correlations between mixture and Gaussian traits, and between two mixture traits, are discussed in covariance structure. This article concludes with some

[1]*Corresponding author:* Department of Animal Sciences, 1675 Observatory Dr., Madison, WI 53706. E-mail: gianola@calshp.cals.wisc.edu

comments and with an APPENDIX, where some basic formulas of mixture distributions are presented.

MODEL

Suppose an observable random variable (y_i, phenotype of individual i) is drawn from the finite mixture of G_E Gaussian components,

$$y_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}^2_e, a_i \sim \sum_{k=1}^{G_E} P_{e_k} N(y_i \mid \mu_k + a_i, \sigma^2_{e_k}), \quad (1)$$

where $\mathbf{p}_e$ is a vector containing the mixing proportions P_{e_k} (summing to 1); $\boldsymbol{\mu}_e$ and $\boldsymbol{\sigma}^2_e$ are each $G_E \times 1$ vectors of means and variances with typical elements μ_k and $\sigma^2_{e_k}$, respectively; a_i is the genetic value of i, and $N(.|.,.)$ denotes a univariate normal density with appropriate mean and variance. As shown in the APPENDIX, the mean and variance of this conditional (given the genetic effect) distribution are

$$E(y_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}^2_e, a_i) = \sum_{k=1}^{G_E} P_{e_k}\mu_k + a_i, \quad (2)$$

and

$$\mathrm{Var}(y_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}^2_e, a_i) = \sum_{k=1}^{G_E} P_{e_k}(\sigma^2_{e_k} + \mu^2_k) - \left(\sum_{k=1}^{G_E} P_{e_k}\mu_k\right)^2 = \sigma^2_e, \quad (3)$$

respectively, where σ^2_e is the residual or "environmental" variance. Informally, $\sum_{k=1}^{G_E} P_{e_k}\mu^2_k - \left(\sum_{k=1}^{G_E} P_{e_k}\mu_k\right)^2$ is the part of the environmental variance contributed by population heterogeneity.

Assume that the genetic effect a_i is also drawn from the mixture with G_A components

$$a_i \mid \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_a \sim \sum_{m=1}^{G_A} P_{a_m} N(a_i \mid \alpha_m, \sigma^2_{a_m}), \quad (4)$$

where $\mathbf{p}_a = [P_{a_1}, \ldots, P_{a_{G_A}}]'$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{G_A}]'$, and $\boldsymbol{\sigma}^2_a = [\sigma^2_{a_1}, \ldots, \sigma^2_{a_{G_A}}]'$ are the vectors of mixing proportions, component means, and component variances, respectively. Then, $E(a_i \mid \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_a) = \sum_{m=1}^{G_A} P_{a_m}\alpha_m$, and

$$\mathrm{Var}(a_i \mid \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_a) = \sum_{m=1}^{G_A} P_{a_m}(\sigma^2_{a_m} + \alpha^2_m) - \left(\sum_{m=1}^{G_A} P_{a_m}\alpha_m\right)^2 = \sigma^2_a, \quad (5)$$

where σ^2_a is the genetic variance, and $\sum_{m=1}^{G_A} P_{a_m}\alpha^2_m - \left(\sum_{m=1}^{G_A} P_{a_m}\alpha_m\right)^2$ is interpretable as "variance between genetic means." In Gaussian linear models the distribution of the random genetic effects is often taken to be $N(a_i \mid 0, \sigma^2_a)$, where σ^2_a is the additive genetic variance, so it may be reasonable to introduce the restriction $\sum_{m=1}^{G_A} P_{a_m}\alpha_m = 0$ in the mixture (VERBEKE and LESAFFRE 1996). The joint density of a_i and y_i is obtained by multiplication of (1) and (4), yielding

$$p(y_i, a_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}^2_e, \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_a) = \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m} N(y_i \mid \mu_k + a_i, \sigma^2_{e_k}) N(a_i \mid \alpha_m, \sigma^2_{a_m}), \quad (6)$$

which is a finite mixture of $G_E \times G_A$ bivariate normal distributions, with mixing proportion $P_{e_k}P_{a_m}$ for the kmth component; note that $\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m} = \sum_{k=1}^{G_E} P_{e_k}\sum_{m=1}^{G_A} P_{a_m} = 1$. From standard Gaussian linear models theory, given the km component (let the indicator $\delta_{km} = 1$ denote such a situation),

$$\begin{bmatrix} y_i \\ a_i \end{bmatrix} \mid \mu_k, \alpha_m, \sigma^2_{e_k}, \sigma^2_{a_m}, \delta_{km} = 1 \sim N_2\left(\begin{bmatrix} y_i \\ a_i \end{bmatrix}, \begin{bmatrix} \mu_k + \alpha_m \\ \alpha_m \end{bmatrix}, \begin{bmatrix} \sigma^2_{e_k} + \sigma^2_{a_m} & \sigma^2_{a_m} \\ \sigma^2_{a_m} & \sigma^2_{a_m} \end{bmatrix}\right),$$

where $N_2(. \mid .,.)$ denotes a bivariate normal distribution. Further,

$$a_i \mid y_i, \mu_k, \alpha_m, \sigma^2_{e_k}, \sigma^2_{a_m}, \delta_{km} = 1 \sim N(a_i \mid \hat{a}_{km}, \sigma^2_{a_m}(1-b_{km})),$$

where

$$\hat{a}_{km} = \alpha_m + b_{km}(y_i - \mu_k - \alpha_m),$$

and

$$b_{km} = \frac{\sigma^2_{a_m}}{\sigma^2_{e_k} + \sigma^2_{a_m}}.$$

Under the standard additive genetic model of FISHER (1918), this regression of "genotype on phenotype" b_{km} is the heritability of the character under the kmth component of the bivariate mixture. The joint density (6) is also expressible as

$$p(y_i, a_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}^2_e, \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_a) = \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m} N(y_i \mid \mu_k + \alpha_m, \sigma^2_{e_k} + \sigma^2_{a_m}) \times N(a_i \mid \hat{a}_{km}, \sigma^2_{a_m}(1 - b_{km})). \quad (7)$$

The marginal density of y_i is arrived at by integrating (7) over a_i, yielding

$$p(y_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}^2_e, \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_a) = \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m} N(y_i \mid \mu_k + \alpha_m, \sigma^2_{e_k} + \sigma^2_{a_m}). \quad (8)$$

This is a finite mixture of $G_E \times G_A$ univariate normal distributions with mixing proportions $P_{e_k}P_{a_m}$. From the

APPENDIX, the mean and variance of the phenotypic distribution are

$$E(y_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}^2_e, \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_a) = \sum_{k=1}^{G_E} P_{e_k}\mu_k + \sum_{m=1}^{G_A} P_{a_m}\alpha_m \quad (9)$$

and

$$\begin{aligned}
&\mathrm{Var}(y_{ij} \mid \mathbf{p}_e,\ \boldsymbol{\mu}_e,\ \boldsymbol{\sigma}^2_e,\ \mathbf{p}_a,\ \boldsymbol{\alpha},\ \boldsymbol{\sigma}^2_a)\\
&\quad = E_{a_i}[\mathrm{Var}(y_i \mid \mathbf{p}_e,\ \boldsymbol{\mu}_e,\ \boldsymbol{\sigma}^2_e,\ a_i)]\\
&\qquad + \mathrm{Var}_{a_i}[E(y_i \mid \mathbf{p}_e,\ \boldsymbol{\mu}_e,\ \boldsymbol{\sigma}^2_e,\ a_i)]\\
&\quad = \sum_{k=1}^{G_E} P_{e_k}(\sigma^2_{e_k} + \mu^2_k) - \left(\sum_{k=1}^{G_E} P_{e_k}\mu_k\right)^2\\
&\qquad + \sum_{m=1}^{G_A} P_{a_m}(\sigma^2_{a_m} + \alpha^2_m) - \left(\sum_{m=1}^{G_A} P_{a_m}\alpha_m\right)^2\\
&\quad = \sigma^2_e + \sigma^2_a. \qquad (10)
\end{aligned}$$

A standard problem in quantitative genetics is that of inferring genetic values from phenotypes. From (7) and (8), the density of the conditional distribution of a_i given y_i is

$$\begin{aligned}
&p(a_i \mid y_i,\ \mathbf{p}_e,\ \boldsymbol{\mu}_e,\ \boldsymbol{\sigma}^2_e,\ \mathbf{p}_a,\ \boldsymbol{\alpha},\ \boldsymbol{\sigma}^2_a)\\
&\quad = \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km} N(a_i \mid \hat{a}_{km},\ \sigma^2_{a_m}(1 - b_{km})), \qquad (11)
\end{aligned}$$

where

$$Q_{km} = \frac{P_{e_k}P_{a_m}N(y_i \mid \mu_k + \alpha_m,\ \sigma^2_{e_k} + \sigma^2_{a_m})}{\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{e_m}N(y_{ij} \mid \mu_k + \alpha_m,\ \sigma^2_{e_k} + \sigma^2_{a_m})}.$$

Hence, the conditional distribution of a_i given y_i is a mixture of the $G_E \times G_A$ normal distributions $N(a_i \mid \hat{a}_{km},\ \sigma^2_{a_m}(1 - b_{km}))$, where the mixing proportion is Q_{km}, the conditional probability that the datum is drawn from $N(y_i \mid \mu_k + \alpha_m,\ \sigma^2_{e_k} + \sigma^2_{a_m})$, given the observation y_i. The best predictor of genetic value is the conditional expectation function

$$\begin{aligned}
&E(a_i \mid y_i,\ \mathbf{p}_e,\ \boldsymbol{\mu}_e,\ \boldsymbol{\sigma}^2_e,\ \mathbf{p}_a,\ \boldsymbol{\alpha},\ \boldsymbol{\sigma}^2_a)\\
&\quad = \int a_i \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km} N(a_i \mid \hat{a}_{km},\ \sigma^2_{a_m}(1 - b_{km}))\,da_i\\
&\quad = \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km}\hat{a}_{km} \qquad (12)
\end{aligned}$$

(HENDERSON 1973; BULMER 1980; FERNANDO and GIANOLA 1986; SEARLE *et al.* 1992), which is a weighted average of the conditional expectations peculiar to each of the $G_E \times G_A$ components of mixture (11). This result is important: the regression of genotype on phenotype is not linear in y_i. Therefore, standard linear models give less than optimal predictions of genetic effects for traits distributed as mixtures. Further, using (39) in the APPENDIX, the variance of the conditional distribution is

$$\begin{aligned}
&\mathrm{Var}(a_i \mid y_i,\ \mathbf{p}_e,\ \boldsymbol{\mu}_e,\ \boldsymbol{\sigma}^2_e,\ \mathbf{p}_a,\ \boldsymbol{\alpha},\ \boldsymbol{\sigma}^2_a)\\
&\quad = \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km}[\sigma^2_{a_m}(1 - b_{km}) + \hat{a}^2_{km}] - \left(\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km}\hat{a}_{km}\right)^2. \qquad (13)
\end{aligned}$$

In the standard additive genetic linear model, the variance of the conditional distribution of genotypes given phenotypes is $\sigma^2_a(1 - h^2)$ (FALCONER 1989), where h^2 is the coefficient of heritability; this conditional variance is homogeneous and does not depend on the data. In a mixture model, however, the dispersion about the regression function is heteroscedastic and nonlinear on the phenotypic value. Hence, both point and interval predictions of genetic value in mixtures involve strikingly different formulas.

TRUNCATION SELECTION

Consider the standard truncation selection setting in which individuals kept as parents are such that $y_i > t$, with the proportion of individuals selected being $\Pr(y_i > t) = \gamma$. From (8), the distribution of phenotypic values within selected individuals has density

$$p_S(y_i) = \frac{\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m}N(y_i \mid \mu_k + \alpha_m,\ \sigma^2_{e_k} + \sigma^2_{a_m})}{\gamma},$$

$$y_i > t,$$

where

$$\begin{aligned}
\gamma &= \int_t^\infty \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m}N(y_i \mid \mu_k + \alpha_m,\ \sigma^2_{e_k} + \sigma^2_{a_m})\,dy_i\\
&= \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m}\left[1 - \Phi\left(\frac{t - \mu_k - \alpha_m}{\sqrt{\sigma^2_{e_k} + \sigma^2_{a_m}}}\right)\right] = \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m}\gamma_{km}. \qquad (14)
\end{aligned}$$

Above, γ_{km} is the proportion selected within the kmth mixture component and $\Phi(.)$ is the standard normal distribution function. The proportion selected γ is, thus, a weighted average of the individual component selection proportions γ_{km}. Since the threshold is fixed, the components that are most prevalent, have largest means, and are most variable will be influential.

The mean value of selected individuals is

$$\begin{aligned}
E_S(y_i) &= \frac{\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m}\int_t^\infty y_i N(y_i \mid \mu_k + \alpha_m,\ \sigma^2_{e_k} + \sigma^2_{a_m})\,dy_i}{\gamma}\\
&= \frac{\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m}\gamma_{km}(\mu_k + \alpha_m + i_{km}\sqrt{\sigma^2_{e_k} + \sigma^2_{a_m}})}{\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} P_{e_k}P_{a_m}\gamma_{km}}\\
&= \sum_{k=1}^{G_E}\sum_{m=1}^{G_A} v_{km}(\mu_k + \alpha_m + i_{km}\sqrt{\sigma^2_{e_k} + \sigma^2_{a_m}}), \qquad (15)
\end{aligned}$$

where i_{km} is the selection intensity factor under the kmth component (Falconer 1989) and

$$v_{km} = \frac{P_{e_k} P_{a_m} \gamma_{km}}{\sum_{k=1}^{G_E} \sum_{m=1}^{G_A} P_{e_k} P_{a_m} \gamma_{km}}$$

are relative weights summing to 1. The phenotypic superiority of selected individuals or selection differential (S) is given by the difference between (15) and (9). Further, the mean genetic value of selected parents is

$$E_S(a_i) = E_y[E(a_i \mid y) \mid y_i > t].$$

Employing (12),

$$E_S(a_i) = E_y\left[\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km}\hat{a}_{km} \mid y_i > t\right].$$

This expression cannot be evaluated analytically, because it is a highly nonlinear function of the phenotypic values. However, it can be approximated by Monte Carlo procedures, *e.g.*, by drawing samples from the bivariate mixture (7). Accept the draws in which $y > t$, calculate $\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km}\hat{a}_{km}$ for each accepted y, and then average this quantity over the samples kept. Finally, the genetic superiority of accepted parents over the unselected population is

$$\begin{aligned}\Delta_a &= E_S(a_i) - E(a_i \mid \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}_a^2)\\ &= E_y\left[\sum_{k=1}^{G_E}\sum_{m=1}^{G_A} Q_{km}\hat{a}_{km} \mid y_i > t\right] - \sum_{m=1}^{G_A} P_{a_m}\alpha_m.\end{aligned}$$

The expected fraction of the selection differential that is realized can be assessed as Δ_a/S, and this will differ from what could be expected from the regression of offspring on midparent, because of nonlinearity (see the following section).

Effects of truncation selection upon a heterogeneous population, *i.e.*, a mixture, have been studied extensively in quantitative genetics. For example, Hill (1974) and Bijma and Woolliams (1999) gave formulas for prediction of response suitable for age-structured populations or for overlapping generations. Also, Latter (1965), Lande (1976), and Kimura and Crow (1978) addressed consequences of truncation selection when there is some grouping structure in a population, *e.g.*, caused by genes of large effects. To illustrate, suppose that a genetic mixture derives from a major locus with two alleles. Prior to selection, the mixing proportions (frequencies) of the three genotypes are, in our notation, P_{a_1}, P_{a_2}, and P_{a_3}. Also, suppose that the environmental distribution is zero-mean normal, with variance σ_e^2, independent of the genetic distribution, and that the polygenic genetic variance is homoscedastic and equal to σ_a^2 (equivalently, the within major genotype genetic variance is constant). Using (14), the overall selection proportion is $\gamma = \sum_{m=1}^{3} P_{a_m}\gamma_m$, and the genotypic frequencies after selection become $P_{a_m}^* = P_{a_m}\gamma_m/\gamma$. After selection, employing (15), the phenotypic distribution has mean value

$$\mu^* = \sum_{m=1}^{3} P_{a_m}^*\left(\alpha_m + i_m\sqrt{\sigma_e^2 + \sigma_a^2}\right).$$

Falconer (1989) gives approximate expressions for relative fitness of genotypes, *e.g.*, γ_2/γ_1. Note that, under these assumptions, the phenotypic distribution remains a mixture, irrespective of the number of cycles of selection. The means and variance change, however, due to the modification of the $P_{a_m}^*$ frequencies produced by selection.

COVARIANCE BETWEEN RELATIVES

General: The fraction of variance attributable to additive genetic effects (usual definition of heritability) is location invariant for a Gaussian trait, *i.e.*, it does not involve mean values. In a mixture, "heritability" becomes

$$\tau^2 = \frac{\sum_{m=1}^{G_A} P_{a_m}(\sigma_{a_m}^2 + \alpha_m^2) - (\sum_{m=1}^{G_A} P_{a_m}\alpha_m)^2}{\sum_{m=1}^{G_A} P_{a_m}(\sigma_{a_m}^2 + \alpha_m^2) - (\sum_{m=1}^{G_A} P_{a_m}\alpha_m)^2 + \sum_{k=1}^{G_E} P_{e_k}(\sigma_{e_k}^2 + \mu_k^2) - (\sum_{k=1}^{G_E} P_{e_k}\mu_k)^2}. \tag{16}$$

The partition of variance depends on component-specific variances $\left(\sigma_{a_m}^2 \text{ and } \sigma_{e_k}^2\right)$, on mixing proportions (P_{a_m} and P_{e_k}), and on mean values (μ_k and α_m) as well. In the simpler case in which the genetic distribution is the homogeneous process $N\left(a_i \mid 0, \sigma_a^2\right)$, heritability becomes

$$\tau^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sum_{k=1}^{G_E} P_{e_k}(\sigma_{e_k}^2 + \mu_k^2) - (\sum_{k=1}^{G_E} P_{e_k}\mu_k)^2}, \tag{17}$$

and this is expected to be lower than in a homogeneous population because fixed effects contribute to variance. If the residual variance is homoscedastic across mixture components, this reduces further to

$$\tau^2 = \frac{h^2}{1 + (\sum_{k=1}^{G_E} P_{e_k}\mu_k^2 - (\sum_{k=1}^{G_E} P_{e_k}\mu_k)^2)/(\sigma_a^2 + \sigma_e^2)}, \tag{18}$$

where $h^2 = \sigma_a^2/\left(\sigma_a^2 + \sigma_e^2\right)$. Heterogeneity in means reduces heritability in a mixture (τ^2), and the standard h^2 is obtained only when the sampling model invokes a draw from a single component distribution.

The covariance between phenotypes of related individuals i and i' is

$$\begin{aligned}&\mathrm{Cov}(y_i, y_{i'} \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}_e^2, \mathbf{p}_a, \boldsymbol{\alpha}, \boldsymbol{\sigma}_a^2)\\ &\quad= E_{a_i,a_{i'}}[\mathrm{Cov}(y_i, y_{i'} \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}_e^2, a_i, a_{i'})]\\ &\qquad+ \mathrm{Cov}_{a_i,a_{i'}}[E(y_i \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}_e^2, a_i), E(y_{i'} \mid \mathbf{p}_e, \boldsymbol{\mu}_e, \boldsymbol{\sigma}_e^2, a_{i'})]\\ &\quad= \mathrm{Cov}_{a_i,a_{i'}}\left[\sum_{k=1}^{G_E} P_{e_k}\mu_k + a_i, \sum_{k=1}^{G_E} P_{e_k}\mu_k + a_{i'}\right] = \mathrm{Cov}(a_i, a_{i'}),\end{aligned} \tag{19}$$

after assuming that phenotypes (given the additive genetic values a_i, $a_{i'}$) are conditionally independent. To develop the covariance between genetic values further, we assume that these are distributed as the bivariate mixture

$$\begin{bmatrix} a_i \\ a_{i'} \end{bmatrix} \mid \mathbf{p}_{\mathrm{a}}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2_{\mathrm{a}} \sim \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} N_2\left(\begin{bmatrix} a_i \\ a_{i'} \end{bmatrix} \mid \begin{bmatrix} \alpha_m \\ \alpha_m \end{bmatrix}, \begin{bmatrix} 1 & A_{ii'} \\ A_{ii'} & 1 \end{bmatrix} \sigma^2_{\mathrm{a}_m} \right),$$

where $N_2(.,.)$ denotes a bivariate normal distribution and A'_{ii} is the additive relationship between the two individuals, assumed constant across all components of the mixture; inbreeding is supposed to be nil. It follows directly that each of the genetic values has $\sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} N\left(a_i \mid \alpha_m, \sigma^2_{\mathrm{a}_m}\right)$ as marginal distribution. Then

$$\begin{aligned} \mathrm{Cov}(a_i, a_{i'}) &= \int\int a_i a_{i'} \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} N_2 \\ &\quad \times \left(\begin{bmatrix} a_i \\ a_{i'} \end{bmatrix} \mid \begin{bmatrix} \alpha_m \\ \alpha_m \end{bmatrix}, \begin{bmatrix} 1 & A_{ii'} \\ A_{ii'} & 1 \end{bmatrix} \sigma^2_{\mathrm{a}_m} \right) \\ &\quad - \left(\sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \alpha_m \right)^2 \\ &= \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} (A_{ii'} \sigma^2_{\mathrm{a}_m} + \alpha^2_m) - \left(\sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \alpha_m \right)^2 \\ &= A_{ii'} \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \sigma^2_{\mathrm{a}_m} + \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \alpha^2_m - \left(\sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \alpha_m \right)^2. \end{aligned} \tag{20}$$

This reduces to $A_{ii'} \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \sigma^2_{\mathrm{a}_m}$ if $\alpha_m = 0$ for every m and to the standard $A_{ii'} \sigma^2_{\mathrm{a}}$ in the absence of heterogeneity in the distribution of genetic effects.

Regression of offspring on parent: Using (10) and (20), the standard formula for the regression of the phenotypic value of a progeny (O) on that of a parent (P) (with $A_{ii'} = \frac{1}{2}$) gives

$$\begin{aligned} \beta_{\mathrm{OP}} &= \frac{\mathrm{Cov}(y_{\mathrm{O}}, y_{\mathrm{P}})}{\mathrm{Var}(y_{\mathrm{P}})} \\ &= \frac{(1/2) \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \sigma^2_{\mathrm{a}_m} + \sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \alpha^2_m - \left(\sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \alpha_m\right)^2}{\sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} (\sigma^2_{\mathrm{a}_m} + \alpha^2_m) - \left(\sum_{m=1}^{G_{\mathrm{A}}} P_{\mathrm{a}_m} \alpha_m\right)^2 + \sum_{k=1}^{G_{\mathrm{E}}} P_{\mathrm{e}_k} (\sigma^2_{\mathrm{e}_k} + \mu^2_k) - \left(\sum_{k=1}^{G_{\mathrm{E}}} P_{\mathrm{e}_k} \mu_k\right)^2}. \end{aligned} \tag{21}$$

If the distribution of genetic effects is homogeneous, this simplifies to

$$\beta_{\mathrm{OP}} = \frac{(1/2)\sigma^2_{\mathrm{a}}}{\sigma^2_{\mathrm{a}} + \sum_{k=1}^{G_{\mathrm{E}}} P_{\mathrm{e}_k} (\sigma^2_{\mathrm{e}_k} + \mu^2_k) - \left(\sum_{k=1}^{G_{\mathrm{E}}} P_{\mathrm{e}_k} \mu_k\right)^2}. \tag{22}$$

The consequences of (21) and (22) are clear: if there is heterogeneity in the distribution either of sampling model residuals or of genetic effects, then β_{OP} is affected by the mixing proportions and by the means μ_k. To illustrate, suppose that the genetic distribution is homogeneous; let $G_{\mathrm{E}} = 2$, take $\mu_1 = 0$ as "origin," $\mu_2 = \Delta\sigma^2_{\mathrm{e}_1}$, and $\sigma^2_{\mathrm{e}_1} = \sigma^2_{\mathrm{e}_2} = 1$. Then (22) is expressible as

$$\beta_{\mathrm{OP}} = \frac{(1/2)\sigma^2_{\mathrm{a}}}{\sigma^2_{\mathrm{a}} + 1 + P_{\mathrm{e}}(1 - P_{\mathrm{e}})\Delta^2}.$$

When $P_{\mathrm{e}} = 1$, the formula gives half of heritability, which is a standard result (Falconer 1989). The function is symmetric with respect to P_{e}; since $P_{\mathrm{e}}(1 - P_{\mathrm{e}})$ is maximum at $P_{\mathrm{e}} = \frac{1}{2}$, the regression is minimum at this value. As an example, consider the offspring–parent regression as a function of P_{e} for four situations with different additive genetic variance (σ^2_{a}) and distances between means (Δ) in the two distributions of the mixture: (1) $\sigma^2_{\mathrm{a}} = 1, \Delta = 1$; (2) $\sigma^2_{\mathrm{a}} = 1, \Delta = 2$; (3) $\sigma^2_{\mathrm{a}} = 0.10, \Delta = 1$; and (4) $\sigma^2_{\mathrm{a}} = 0.10, \Delta = 2$. Situations 1 and 2 correspond to a trait with a heritability of 0.50 under homogeneity, while 3 and 4 are for a lowly heritable trait ($h^2 \approx 0.09$). In 1 and 2, the regression β decreases from 0.25 to ~0.22 and 0.17, respectively, representing relative decreases in heritability of 12 and 32%. The relative decreases in heritability are 18 and 47% in cases 3 and 4, respectively. In brief, heritability in heterogeneous or admixed populations depends on the mixing proportion, on the mean difference between mixture components, and on the "homogeneous situation" heritability.

COVARIANCE STRUCTURE

Correlations with a Gaussian trait: Correlations between a mixture trait and a normally distributed character may be of interest. For example, the mixture trait could be SCC in dairy cattle, with several component distributions corresponding to different unknown statuses of mammary gland disease. The Gaussian trait could be milk yield of a cow. Is the genetic correlation between the two traits affected by heterogeneity of somatic cell count?

Let the model for the Gaussian trait w be

$$w_i = \mu_w + a_{wi} + e_{wi}, \tag{23}$$

where μ_w is the mean of the trait, $a_{wi} \sim N\left(0, \sigma^2_{\mathrm{a}_w}\right)$ is the additive genetic value of individual i for trait w, and $e_{wi} \sim N\left(0, \sigma^2_{\mathrm{e}_w}\right)$ is a residual effect, independent of a_{wi}; $\sigma^2_{\mathrm{a}_w}$ and $\sigma^2_{\mathrm{e}_w}$ are genetic and residual components of variance, respectively. The phenotypic distribution is, thus, $w_i \sim N\left(\mu_w, \sigma^2_{\mathrm{a}_w} + \sigma^2_{\mathrm{e}_w}\right)$.

Suppose that the distribution of mixture trait y has two components at each of the genetic and residual levels; *i.e.*, $G_{\mathrm{A}} = G_{\mathrm{E}} = 2$. Assume further that

$$\begin{bmatrix} a_{1i} \\ a_{2i} \\ a_{wi} \\ e_{1i} \\ e_{2i} \\ e_{wi} \end{bmatrix} \sim N_6\left(\begin{bmatrix} \alpha \\ \alpha \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\mathrm{a}_1} & \sigma_{\mathrm{a}_{12}} & \sigma_{\mathrm{a}_{1w}} & 0 & 0 & 0 \\ & \sigma^2_{\mathrm{a}_2} & \sigma_{\mathrm{a}_{2w}} & 0 & 0 & 0 \\ & & \sigma^2_{\mathrm{a}_w} & 0 & 0 & 0 \\ & & & \sigma^2_{\mathrm{e}_1} & 0 & \sigma_{\mathrm{e}_{1w}} \\ & & & & \sigma^2_{\mathrm{e}_2} & \sigma_{\mathrm{e}_{2w}} \\ \text{Symmetric} & & & & & \sigma^2_{\mathrm{e}_w} \end{bmatrix} \right). \tag{24}$$

The distribution of each of the two genetic effects entering into the mixture for trait y (a_1, a_2) is centered at α, but has component-specific variances. These two genetic effects may be imperfectly correlated, with genetic covariance $\sigma_{a_{12}}$; for instance, different genes affecting somatic cell count are expressed under the unknown "mastitis" and "no-mastitis" disease conditions generating the mixture. Also, the genetic covariance with the Gaussian trait may be specific to each of the two components. The component-specific residual effects (e_1, e_2) are heteroscedastic but uncorrelated, as it is not possible to observe "disease" and "no disease" in the same individual at the same time. However, a residual correlation with the Gaussian trait is allowed and assumed peculiar to each component distribution.

Recall from (4) that $a_i \,|\, P_{a_1}, P_{a_2}, \alpha, \sigma^2_{a_1}, \sigma^2_{a_2} \sim \sum_{m=1}^{2} P_{a_m} N\left(a_i \,|\, \alpha, \sigma^2_{a_m}\right)$. Now, let ∂_m take the value 1 when the draw is from component m and 0 otherwise. The joint density under m is

$$p(a_{wi}, a_{mi} \,|\, \partial_m = 1) = \begin{bmatrix} a_{mi} \\ a_{wi} \end{bmatrix} \sim N_2\left(\begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{a_m} & \sigma_{a_{mw}} \\ \sigma_{a_{mw}} & \sigma^2_{a_w} \end{bmatrix}\right), \quad m = 1, 2,$$

so that, unconditionally

$$p(a_{wi}, a_i) = \sum_{m=1}^{2} P_{a_m} N_2\left(\begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{a_m} & \sigma_{a_{mw}} \\ \sigma_{a_{mw}} & \sigma^2_{a_w} \end{bmatrix}\right).$$

Then

$$\begin{aligned} \mathrm{Cov}(a_w, a_i) &= \iint a_w a_i p(a_w, a_i)\, da_w da_i \\ &= \sum_{m=1}^{2} P_{a_m} \iint a_w a_i N_2\left(\begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{a_m} & \sigma_{a_{mw}} \\ \sigma_{a_{mw}} & \sigma^2_{a_w} \end{bmatrix}\right) da_w da_i \\ &= \sum_{m=1}^{2} P_{a_m} \sigma_{a_{mw}} = P_{a_1} \sigma_{a_{1w}} + (1 - P_{a_1}) \sigma_{a_{2w}}. \end{aligned} \tag{25}$$

Using (5) and (25), and taking $\alpha = 0$, the genetic correlation between the mixture-distributed trait and the Gaussian character is

$$\rho_{a,a_w} = \frac{P_{a_1} \sigma_{a_{1w}} + (1 - P_{a_1}) \sigma_{a_{2w}}}{\sigma_{a_w} \sqrt{\sum_{m=1}^{2} P_{a_m} \sigma^2_{a_m}}}. \tag{26}$$

This reduces to the standard $\sigma_{a_{yw}} / \left(\sigma_{a_y} \sigma_{a_w}\right)$ when the distribution of genetic effects for trait y is homogeneous and to

$$\rho_{a_y,a_w} = \frac{\sigma_{a_{yw}}}{\sigma_{a_w} \sqrt{\left[\sum_{m=1}^{2} P_{a_m} \sigma^2_{a_m}\right]}}, \tag{27}$$

when $\sigma_{a_{1w}} = \sigma_{a_{2w}}$.

The effect of P_{a_m} on genetic correlation (27) is illustrated next. Let $\lambda = \sigma^2_{a_2} / \sigma^2_{a_1}$ be a heteroscedasticity factor, where $\sigma^2_{a_1}$, the genetic variance under the first component of the mixture, is viewed as "baseline" genetic variance, *i.e.*, a measure of variability in the absence of heterogeneity. Then

$$\begin{aligned} \rho_{a_y,a_w} &= \frac{\sigma_{a_{yw}}}{\sigma_{a_w} \sigma_{a_1} \sqrt{P_{a_1} + (1 - P_{a_1})\lambda}} \\ &= \frac{\rho_{\mathrm{homo}}}{\sqrt{P_{a_1} + (1 - P_{a_1})\lambda}}, \end{aligned} \tag{28}$$

where ρ_{homo} is the genetic correlation in the absence of a mixture and

$$[P_{a_1} + (1 - P_{a_1})\lambda]^{-(1/2)}$$

is the factor by which ρ_{homo} is modified by heterogeneity. Since the sign of ρ_{a_y,a_w} is invariant with respect to P_{a_1}, it suffices to examine function (28) only under positive values of ρ_{homo}. Figure 1 displays the relationship between the genetic correlation (28) and P_{a_1} for two values of ρ_{homo} (0.7 and 0.3) and of λ (1.5 and 2). As P_{a_1} increases, the proportion of the component with larger genetic variance ($m = 2$) decreases. The genetic correlation increases monotonically with P_{a_1} and more rapidly so at the largest value of genetic heteroscedasticity. Suppose that w is total lactation milk yield in dairy cows and that y is SCC, a mixture trait resulting from the fact that some cows have mastitis (~20–40%). There is evidence (*e.g.*, Heringstad *et al.* 2006) that the genetic variance of somatic cell count is ~2.5 times larger in healthy than in diseased (clinical cases) cows. Under the assumptions leading to (28), our model predicts that the genetic correlation between milk yield and somatic cell score would decrease as the frequency of mastitis in the population decreases. Similar algebra and

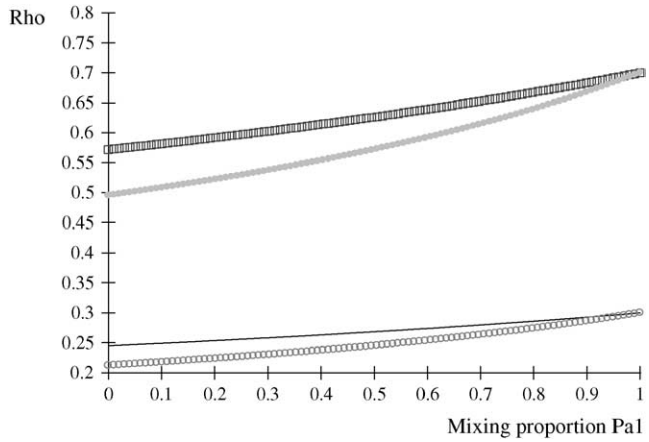

Figure 1.—Genetic correlation (Rho) between a Gaussian character and a mixture trait for a two-component mixture, as a function of the mixing proportion (P_{a_1}), for different combinations of ρ_{homo}, genetic correlation in absence of mixture, and λ, heteroscedasticity factor. From top to bottom: (1) $\rho_{\mathrm{homo}} = 0.7$, $\lambda = 1.5$ (open squares); (2) $\rho_{\mathrm{homo}} = 0.7$, $\lambda = 2$ (dotted line); (3) $\rho_{\mathrm{homo}} = 0.3$, $\lambda = 1.5$ (solid line); (4) $\rho_{\mathrm{homo}} = 0.3$, $\lambda = 2$ (open circles).

considerations hold for the environmental correlation between traits.

Consider again the joint distribution (24), and write

$$y_i = \delta_e\mu_1 + (1-\delta_e)\mu_2 + \delta_a a_{1i} + (1-\delta_a)a_{2i} + \delta_e e_{1i} + (1-\delta_e)e_{2i}, \quad (29)$$

where the random variable δ_e takes the value 1 or 0 with probabilities P_e and $1 - P_e$, respectively; δ_a is another binary variable taking the values 1 or 0 with probabilities P_a and $(1 - P_a)$, respectively, and distributed independently of δ_e. These two binary variates are assumed to be independent of a_{wi} and e_{wi} entering into the model for w_i in (23). Then,

$$E(w_i \mid \delta_e, \delta_a) = E(w_i) = \mu_w,$$
$$E(y_i \mid \delta_e, \delta_a) = \delta_e\mu_1 + (1-\delta_e)\mu_2 + \alpha,$$

and

$$\begin{aligned}\mathrm{Cov}(y_i, w_i \mid \delta_e, \delta_a) &= \mathrm{Cov}[\delta_a a_{1i} + (1-\delta_a)a_{2i} + \delta_e e_{1i} \\ &\quad + (1-\delta_e)e_{2i},\ a_{wi} + e_{wi} \mid \delta_e, \delta_a] \\ &= \delta_a\sigma_{a_{1w}} + (1-\delta_a)\sigma_{a_{2w}} + \delta_e\sigma_{e_{1w}} + (1-\delta_e)\sigma_{e_{2w}}.\end{aligned}$$

The phenotypic covariance between y and w is, therefore,

$$\begin{aligned}\mathrm{Cov}(y_i,\ w_i) &= E_{\delta_e,\delta_a}[\delta_a\sigma_{a_{1w}} + (1-\delta_a)\sigma_{a_{2w}} \\ &\quad + \delta_e\sigma_{e_{1w}} + (1-\delta_e)\sigma_{e_{2w}}] \\ &\quad + \mathrm{Cov}_{\delta_e,\delta_a}[\delta_e\mu_1 + (1-\delta_e)\mu_2 + \alpha,\ \mu_w].\end{aligned}$$

Since the second term is null

$$\mathrm{Cov}(y_i,\ w_i) = P_a\sigma_{a_{1w}} + (1-P_a)\sigma_{a_{2w}} + P_e\sigma_{e_{1w}} + (1-P_e)\sigma_{e_{2w}}. \quad (30)$$

Above, $P_a\sigma_{a_{1w}} + (1 - P_a)\sigma_{a_{2w}}$ is the genetic covariance, as in (25), and $P_e\sigma_{e_{1w}} + (1 - P_e)\sigma_{e_{2w}}$ is the residual covariance. Collecting (30), (10), plus the fact that $w_i \sim N(\mu_w,\ \sigma^2_{a_w} + \sigma^2_{e_w})$, and assuming that $\alpha = 0$, yields as phenotypic correlation

$$\rho_{yw} = \frac{P_a\sigma_{a_{1w}} + (1-P_a)\sigma_{a_{2w}} + P_e\sigma_{e_{1w}} + (1-P_e)\sigma_{e_{2w}}}{\sigma_w\sqrt{\sum_{m=1}^{2} P_{a_m}\sigma^2_{a_m} + \sum_{k=1}^{2} P_{e_k}(\sigma^2_{e_k} + \mu^2_k) - (\sum_{k=1}^{2} P_{e_k}\mu_k)^2}}. \quad (31)$$

The phenotypic correlation depends not only on the underlying components of variance and covariance, but also on the mixing proportions and population means.

If the genetic distribution is homoscedastic $(\sigma^2_{a_1} = \sigma^2_{a_2} = \sigma^2_{a_y}; \sigma_{a_{1w}} = \sigma_{a_{2w}} = \sigma_{a_{yw}})$, and heterogeneity is at the level of the sampling model only, but with $\sigma_{e_{1w}} = \sigma_{e_{2w}} = \sigma_{e_{yw}}$ and $\sigma^2_{e_1} = \sigma^2_{e_2} = \sigma^2_{e_y}$, the phenotypic correlation becomes

$$\rho_{yw} = \frac{\rho_{\mathrm{homo}}}{\sqrt{1 + (\sum_{k=1}^{2} P_{e_k}\mu_k^2 - (\sum_{k=1}^{2} P_{e_k}\mu_k)^2)/\sigma^2_y}}, \quad (32)$$

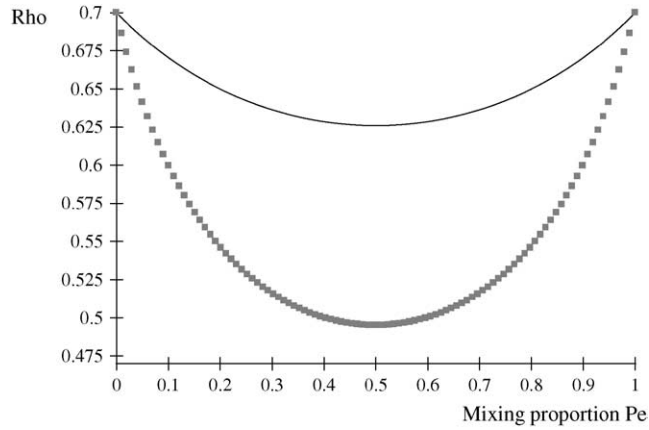

FIGURE 2.—Phenotypic correlation (Rho) between a Gaussian character and a mixture trait for a two-component mixture, as a function of the mixing proportion (P_e). (1) $\rho_{\mathrm{homo}} = 0.7$, $\Delta = 1$ (solid line); (2) $\rho_{\mathrm{homo}} = 0.7$, $\Delta = 2$ (line with squares). ρ_{homo}, phenotypic correlation in the absence of a mixture; Δ, difference between means of the two distributions.

where $\rho_{\mathrm{homo}} = (\sigma_{a_{yw}} + \sigma_{e_{1w}})/(\sigma_w\sigma_y)$ is the phenotypic correlation in the absence of a mixture for the residual distribution and $\sigma^2_y = \sigma^2_{a_y} + \sigma^2_{e_y}$. To illustrate, take $\mu_1 = 0$ as origin, $\mu_2 = \Delta$, and $\sigma_y = 1$. Then

$$\rho_{yw} = \frac{\rho_{\mathrm{homo}}}{\sqrt{1 + P_e(1-P_e)\Delta^2}}. \quad (33)$$

The phenotypic correlation has a minimum at $P_e = 0.5$ if ρ_{homo} is positive; however, it is maximum at this value of the mixing proportion if ρ_{homo} is negative. Effects of P_e and of Δ on the genetic correlation are shown in Figure 2, for $\rho_{\mathrm{homo}} = 0.7$ and $\Delta = 1$ and 2. The function is symmetric and steeper as Δ increases. For $\Delta = 2$, the correlation decreases from 0.7 (for $P_e = 0$ or 1) to a minimum of $\sim$0.50. The curves are inverted if ρ_{homo} is negative. In short, if the value of $P_e(1 - P_e)$ is used to measure admixture in the residual distribution, the phenotypic correlation decreases with admixture if it is positive in a homogeneous population. On the other hand, ρ_{yw} increases with admixture if negative under the homogeneous situation.

Correlations between two mixture-distributed characters: Suppose now that measurements are available for traits y and z. For simplicity, it is assumed that the joint distribution of y and z arises from a two-component mixture of bivariate normal distributions at the level of the sampling model (that is, given the genetic effects) and from a two-component bivariate normal mixture of genetic effects. Given the independently distributed binary indicator variables δ_e and δ_a, one can write

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} = \begin{bmatrix} \delta_e\mu_{1y} + (1-\delta_e)\mu_{2y} + \delta_a a_{1i,y} \\ \quad + (1-\delta_a)a_{2i,y} + \delta_e e_{1i,y} + (1-\delta_e)e_{2i,y} \\ \delta_e\mu_{1z} + (1-\delta_e)\mu_{2z} + \delta_a a_{1i,z} \\ \quad + (1-\delta_a)a_{2i,z} + \delta_e e_{1i,z} + (1-\delta_e)e_{2i,z} \end{bmatrix}. \quad (34)$$

Then

$$\begin{aligned}\mathrm{Cov}(y_i,\, z_i) = {} & E_{\delta_e,\delta_a}\mathrm{Cov}(y,\, z \mid \delta_e,\, \delta_a) \\ & + \mathrm{Cov}_{\delta_e,\delta_a}[E(y \mid \delta_e,\, \delta_a),\, E(z \mid \delta_e,\, \delta_a)].\end{aligned} \tag{35}$$

Assuming independence between genetic and environmental effects in the distributions, it follows that

$$\begin{aligned}\mathrm{Cov}(y,\, z \mid \delta_e,\, \delta_a) = {} & \delta_a^2\sigma_{a_{1,yz}} + 2\delta_a(1-\delta_a)\sigma_{a_{12,yz}} \\ & + (1-\delta_a)^2\sigma_{a_{2,yz}} \\ & + \delta_e^2\sigma_{e_{1,yz}} + 2\delta_e(1-\delta_e)\sigma_{e_{12,yz}} \\ & + (1-\delta_e)^2\sigma_{e_{2,yz}},\end{aligned}$$

where $\sigma_{a_{m,yz}}$ and $\sigma_{e_{m,yz}}$ $(m = 1, 2)$ are the additive and environmental components of covariance between y and z under m, and $\sigma_{a_{12,yz}}$ and $\sigma_{e_{12,yz}}$ are potential cross-mixture covariances. Further, under the assumption of a null mean of the component-specific genetic and environmental distributions, as well as of independence of the binary indicator variables δ_e and δ_a,

$$E(y \mid \delta_e,\, \delta_a) = \delta_e\mu_{1y} + (1-\delta_e)\mu_{2y},$$

and

$$E(z \mid \delta_e,\, \delta_a) = \delta_e\mu_{1z} + (1-\delta_e)\mu_{2z}.$$

Since δ_e and δ_a are Bernoulli, $E(\delta_e) = P_e$, $\mathrm{Var}(\delta_e) = P_e(1-P_e)$, $E(\delta_a) = P_a$, and $\mathrm{Var}(\delta_a) = P_a(1-P_a)$. Then

$$\begin{aligned}& E_{\delta_e,\delta_a}\mathrm{Cov}(y,\, z \mid \delta_e,\, \delta_a) \\ & \quad = P_a\sigma_{a_{1,yz}} + (1-P_a)\sigma_{a_{2,yz}} + P_e\sigma_{e_{1,yz}} + (1-P_e)\sigma_{e_{2,yz}}\end{aligned}$$

and

$$\begin{aligned}& \mathrm{Cov}_{\delta_e,\delta_a}[E(y \mid \delta_e,\, \delta_a),\, E(z \mid \delta_e,\, \delta_a)] \\ & \quad = \mathrm{Cov}_{\delta_e,\delta_a}[\delta_e\mu_{1y} + (1-\delta_e)\mu_{2y},\, \delta_e\mu_{1z} + (1-\delta_e)\mu_{2z}] \\ & \quad = P_e(1-P_e)\mu_{1y}\mu_{1z} + P_e(1-P_e)\mu_{2y}\mu_{2z} \\ & \qquad - P_e(1-P_e)(\mu_{1y}\mu_{2z} + \mu_{2y}\mu_{1z}).\end{aligned}$$

Employing the preceding results in (35),

$$\begin{aligned}& \mathrm{Cov}(y_i,\, z_i) \\ & \quad = P_a\sigma_{a_{1,yz}} + (1-P_a)\sigma_{a_{2,yz}} + P_e\sigma_{e_{1,yz}} + (1-P_e)\sigma_{e_{2,yz}} \\ & \qquad + P_e(1-P_e)[\mu_{1y}\mu_{1z} + \mu_{2y}\mu_{2z} - (\mu_{1y}\mu_{2z} + \mu_{2y}\mu_{1z})].\end{aligned} \tag{36}$$

Under the assumption that the underlying genetic distributions have zero means, the genetic correlation between the two mixture traits is

$$\rho_{a_{yz}} = \frac{P_a\sigma_{a_{1,yz}} + (1-P_a)\sigma_{a_{2,yz}}}{\sqrt{\sum_{m=1}^{2} P_{a_m}\sigma^2_{a_{m,y}} \sum_{m=1}^{2} P_{a_m}\sigma^2_{a_{m,z}}}}.$$

The environmental correlation takes the form

$$\rho_{e_{yz}} = \frac{\sum_{k=1}^{2} P_{e_m}\sigma_{e_{m,yz}} + P_e(1-P_e)[\mu_{1y}\mu_{1z} + \mu_{2y}\mu_{2z} - (\mu_{1y}\mu_{2z} + \mu_{2y}\mu_{1z})]}{\sqrt{[\sum_{k=1}^{2} P_{e_k}\mu^2_{ky} - (\sum_{k=1}^{2} P_{e_k}\mu_{ky})^2][\sum_{k=1}^{2} P_{e_k}\mu^2_{kz} - (\sum_{k=1}^{2} P_{e_k}\mu_{kz})^2]}},$$

and the phenotypic correlation follows from (36) and (10), after setting all α'*s* to 0.

CONCLUSION

Some basic results of standard theory of quantitative genetics under additive inheritance were extended to finite mixture models with Gaussian components. It was found that, if there is heterogeneity in a population at either the genetic or the environmental levels, then genetic parameters based on theory treating distributions as homogeneous can lead to misleading interpretations. Some peculiarities of mixture characters are: heritability depends on the mean values of the populations, the offspring–parent regression is nonlinear, and genetic or phenotypic correlations cannot be interpreted devoid of the mixture proportions and of the parameters of the component distributions. For example, nonlinearity of the offspring–parent regression was studied by Robertson (1977) and Gimelfarb (1986) under dominance and by Im and Gianola (1988) for binary traits. Gimelfarb (1986) gave conditions under which the regression would be nearly linear under dominance, *e.g.*, a large number of loci affecting the trait or mild dominance and gene frequencies not far from 0.5. Our results illustrate that nonlinearity can also arise due to heterogeneity at the environmental level due to, for instance, omitting relevant covariates in a linear model for quantitative genetic analysis.

Clearly, standard models for quantitative traits can lead to erroneous results if fitted to heterogeneous data. If a mixture is suspected, two suitable methods for inferring unknown mixture parameters are maximum-likelihood and Bayesian analyses. Procedures for likelihood- or posterior-based inference applied to mixtures are discussed extensively in Titterington *et al.* (1985) and McLachlan and Peel (2000), including situations in which the component distributions are not normal, *e.g.*, skewed survival processes. Implementations suitable for fitting different types of quantitative genetic mixture models have been described and applied by Ødegård *et al.* (2003, 2005), Gianola *et al.* (2004), and Boettcher *et al.* (2005). Prediction of breeding values is discussed in Gianola (2005). A suitable software for the analysis of mixtures with random effects is available in a forthcoming update of Version 6.0 of the DMU package (Madsen and Jensen 2002).

An important issue is how many components should be fitted in a mixture model. Testing for the number of components is a difficult matter, and there may not be a one-to-one correspondence between the number of components fitted and the number of heterogeneous groups (McLachlan and Peel 2000). For example, several groups may be hidden behind an apparently bimodal distrbution, due to limited sample size. Also, if some of the component distributions are skewed, typically the number of components needed for a good

fit is larger than the number of groups causing heterogeneity. Probably the most elegant procedures for inferring the number of components needed are Bayesian implementations via the reversible-jump algorithm (Richardson and Green 1997), but computations can be extremely taxing.

Research was supported by the Wisconsin Agriculture Experiment Station and by grants National Research Initiatives Competitive Grants Program/U.S. Department of Agriculture 2003-35205-12833, National Science Foundation (NSF) DEB-0089742, and NSF DMS-044371. Financial support from the Babcock Institute for International Dairy Research and Development, University of Wisconsin, Madison, is acknowledged.

LITERATURE CITED

Bijma, P., and J. A. Woolliams, 1999 Prediction of genetic contributions and generation intervals in populations with overlapping generations under selection. Genetics **151:** 1197–1210.

Boettcher, P. J., P. Moroni, G. Pisoni and D. Gianola, 2005 Application of a finite mixture model to somatic cell scores of Italian goats. J. Dairy Sci. **88:** 2209–2216.

Bulmer, M. G., 1980 *The Mathematical Theory of Quantitative Genetics.* Clarendon Press, Oxford.

Detilleux, J., and P. L. Leroy, 2000 Application of a mixed normal mixture model to the estimation of mastitis-related parameters. J. Dairy Sci. **83:** 2341–2349.

Fernando, R. L., and D. Gianola, 1986 Optimal properties of the conditional mean as a selection criterion. Theor. Appl. Genet. **72:** 822–825.

Gianola, D., 2005 Prediction of random effects in finite mixture models with Gaussian components. J. Anim. Breed. Genet. **122:** 145–160.

Gianola, D., J. Ødegård, B. Heringstad, G. Klemetsdal, D. Sorensen *et al.*, 2004 Mixture model for inferring susceptibility to mastitis in dairy cattle: a procedure for likelihood-based inference. Genet. Sel. Evol. **36:** 3–27.

Falconer, D. S., 1989 *Introduction to Quantitative Genetics.* Longman, Burnt Mill, Harlow, UK.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. **52:** 399–433.

Gimelfarb, A., 1986 Offspring-parent regression: How linear is it? Biometrics **42:** 67–71.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

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

Heringstad, B., G. Klemetsdal and J. Ruane, 2000 Selection for mastitis resistance in dairy cattle: a review with focus on the situation in the Nordic countries. Livest. Prod. Sci. **64:** 95–106.

Heringstad, B., D. Gianola, Y. M. Chang, J. Ødegård and G. Klemetsdal, 2006 Genetic associations between clinical mastitis and somatic cell score in early first lactation cows. J. Dairy Sci. **89:** 2236–2244.

Hill, W. G., 1974 Prediction and evaluation of response to selection with overlapping generations. Anim. Prod. **18:** 117–139.

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

Kimura, M., and J. F. Crow, 1978 Effect of overall phenotypic selection on genetic change at individual loci. Proc. Natl. Acad. Sci. USA **75:** 6168–6171.

Lande, R., 1976 Natural selection and random genetic drift in phenotypic evolution. Evolution **30:** 314–334.

Latter, B. D. H., 1965 The response to artificial selection due to autosomal genes of large effect. Aust. J. Biol. Sci. **18:** 585–598.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer, Sunderland, MA.

Madsen, P., and J. Jensen, 2002 *A User's Guide to DMU. A Package for Analysing Mixed Models*, Version 6, Release 4.3. Danish Institute of Agricultural Sciences, Tjele, Denmark.

McLachlan, G., and D. Peel, 2000 *Finite Mixture Models.* Wiley, New York.

Ødegård, J., J. Jensen, P. Madsen, D. Gianola, G. Klemetsdal *et al.*, 2003 Mixture models for detection of mastitis in dairy cattle using test-day somatic cell scores: a Bayesian approach via Gibbs sampling. J. Dairy Sci. **86:** 3694–3703.

Ødegård, J., P. Madsen, D. Gianola, G. Klemetsdal, J. Jensen *et al.*, 2005 A Bayesian threshold-normal mixture model for analysis of a continuous mastitis-related trait. J. Dairy Sci. **88:** 2652–2659.

Pearson, K., 1894 Contributions to the mathematical theory of evolution. Philos. Trans. R. Soc. A **185:** 71–110.

Richardson, S., and P. J. Green, 1997 On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Stat. Soc. B **59:** 731–792.

Robertson, A., 1977 The non-linearity of offspring-parent regression, pp. 297–303 in *Proceedings of the International Conference on Quantitative Genetics*, edited by E. Pollak, O. Kempthorne and T. B. Bailey, Jr. Iowa State University Press, Ames, IA.

Searle, S. R., G. Casella and C. E. McCulloch, 1992 *Variance Components.* Wiley, New York.

Titterington, D. M., A. F. M. Smith and U. E. Makov, 1985 *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester, UK.

Verbeke, G., and E. Lesaffre, 1996 A linear mixed effects model with heterogeneity in the random-effects population. J. Am. Stat. Assoc. **91:** 217–221.

Communicating editor: B. J. Walsh

APPENDIX

The first and second moments, and the variance of a finite mixture of K Gaussian distributions, with parameters $\boldsymbol{\theta} = [P_1, \ldots, P_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2]'$, where the mixture proportions P_k are such that $\sum_{k=1}^{K} P_k = 1$, are

$$E(y \mid \boldsymbol{\theta}) = \int y \left[\sum_{k=1}^{K} P_k N(y \mid \mu_k, \sigma_k^2) \right] dy = \sum_{k=1}^{K} P_k \mu_k, \quad \text{(A1)}$$

$$E(y^2 \mid \boldsymbol{\theta}) = \int y^2 \left[\sum_{k=1}^{K} P_k N(y \mid \mu_k, \sigma_k^2) \right] dy = \sum_{k=1}^{K} P_k (\mu_k^2 + \sigma_k^2), \quad \text{(A2)}$$

and

$$\mathrm{Var}(y \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} P_k \sigma_k^2 + \sum_{k=1}^{K} P_k \mu_k^2 - \left(\sum_{k=1}^{K} P_k \mu_k \right)^2. \quad \text{(A3)}$$

The first term in (A3) can be interpreted as an average variance, while $\sum_{k=1}^{K} P_k \mu_k^2 - \left(\sum_{k=1}^{K} P_k \mu_k \right)^2$ measures dispersion between group means or heterogeneity; if the μ's are equal, this term is null. The variance of the mixture depends not only on individual component variances, but also on group means. If the components have homogeneous variance σ^2,

$$\mathrm{Var}(y \mid \boldsymbol{\theta}) = \sigma^2 + \sum_{k=1}^{K} P_k \mu_k^2 - \left(\sum_{k=1}^{K} P_k \mu_k \right)^2. \quad \text{(A4)}$$