# Variation génétique de caractères mesurés dans plusieurs milieux. I. Estimation et test d'homogénéité des corrélations génétiques et intra-classe entre milieux

Christèle Robert-Granié, Jean Louis J. L. Foulley, Vincent Ducrocq

*Genet Sel Evol* (1995) 27, 111-123



**Original article**

# Genetic variation of traits measured in several environments. I. Estimation and testing of homogeneous genetic and intra-class correlations between environments

C Robert, JL Foulley, V Ducrocq

*Institut national de la recherche agronomique, station de génétique quantitative et appliquée, centre de recherche de Jouy-en-Josas, 78352 Jouy-en-Josas cedex, France*

(Received 23 March 1994; accepted 26 September 1994)

**Summary** – Estimation of between family (or genotype) components of (co)variance among environments, testing of homogeneity of genetic correlations between environments, and testing of homogeneity of both genetic and intra-class correlations between environments are investigated. The testing procedures are based on the ratio of maximized log-restricted likelihoods for the reduced (under each hypothesis of homogeneity) and saturated models, respectively. An expectation-maximization (EM) iterative algorithm is proposed for calculating restricted maximum likelihood (REML) estimates of the residual and between-family components of (co)variance. The EM formulae are applied to the multiple trait linear model for the saturated model and to the univariate linear model for the reduced models. The EM algorithm guarantees that (co)variance estimates remain within the parameter space. The procedures presented in this paper are illustrated with the analysis of 5 vegetative and reproductive traits recorded in an experiment on 20 full-sib families of black medic (*Medicago lupulina* L) tested in 3 environments.

**heteroskedasticity / genetic correlation / intra-class correlation / expectation-maximization / restricted maximum likelihood**

**Résumé – Variation génétique de caractères mesurés dans plusieurs milieux. I. Estimation et test d'homogénéité des corrélations génétiques et intra-classe entre milieux.** *Cet article étudie les problèmes d'estimation des composantes familiales de (co)variance entre milieux et les problèmes de test d'homogénéité, soit des corrélations génétiques entre milieux seules, soit des corrélations génétiques et des corrélations intra-classe entre milieux. Les procédures de test reposent sur le rapport de vraisemblances restreintes maximisées sous les modèles réduits (les différentes hypothèses d'homogénéité) et le modèle saturé. Un algorithme itératif d'espérance-maximisation (EM) est proposé pour calculer les estimations du maximum de vraisemblance restreinte (REML) des composantes résiduelles et familiales de variance-covariance. Les formules EM s'appliquent au modèle multica-*

*ractère pour le modèle saturé et à des modèles linéaires univariés pour les modèles réduits. Les formules EM garantissent l'appartenance des composantes de (co)variance estimées à l'espace des paramètres. Les procédures présentées dans cet article sont illustrées par l'analyse de 5 caractères végétatifs et reproductifs mesurés lors d'une expérience portant sur 20 familles de pleins frères testées dans 3 milieux différents chez la minette (*Medicago lupulina L*).*

**hétéroscédasticité / corrélation génétique / corrélation intra-classe / espérance-maximisation / maximum de vraisemblance restreinte**

## INTRODUCTION

Hypothesis testing of genetic parameters is of great concern when analyzing genotype × environment interaction experiments. For instance, Visscher (1992) investigated the statistical power of balanced sire × environment designs for detecting heterogeneity of phenotypic variance and intra-class correlation between environments. He assumed that the between-family correlation (henceforth referred to as 'genetic correlation') between environments was equal to 1 and consequently heterogeneity of variance components was only due to scaling. This assumption was relaxed by Foulley *et al* (1994), who considered estimation and testing procedures for homogeneous components of (co)variance between environments. In some cases, it may also be interesting to test less restrictive hypotheses, *eg*, constant genetic correlations between environments, and constant genetic and intra-class correlations between environments. The objective of this paper is to address this issue and to show how heteroskedastic linear mixed models can be useful for this objective.

## THEORY AND METHODS

### *The saturated model*

Let us assume that records are generated from a cross-classified layout. We will consider as in Falconer (1952) that expressions of the trait in different environments are those of genetically correlated traits, thus resulting in the following 'genotype × environment' multiple trait linear model:

$$y_{ijk} = \mu_i + b_{ij} + e_{ijk} \qquad [1]$$

where $y_{ijk}$ is the performance of the $k$th individual ($k = 1, 2, \ldots, n$) of the $j$th family ($j = 1, 2, \ldots, s$) evaluated in the $i$th environment ($i = 1, 2, \ldots, p$); $b_{ij}$ is the random effect of the $j$th family in the $i$th environment, assumed normally distributed such that $\mathrm{Var}(b_{ij}) = \sigma^2_{B_i}$, $\mathrm{Cov}(b_{ij}, b_{i'j}) = \sigma_{B_{ii}}$, for $i \neq i'$ and $\mathrm{Cov}(b_{ij}, b_{i'j'}) = 0$ for $j \neq j'$ and any $i$ and $i'$; and $e_{ijk}$ is a residual effect pertaining to the $k$th individual in the subclass $ij$, assumed normally and independently distributed with mean 0 and variance $\sigma^2_{W_i}$. Using vector notation, *ie* $\mathbf{y}_{jk} = \{y_{ijk}\}$, $\boldsymbol{\mu} = \{\mu_i\}$, $\mathbf{b}_j = \{b_{ij}\}$ and $\mathbf{e}_{jk} = \{e_{ijk}\}$ for $i = 1, 2, \ldots, p$, the model [1] can alternatively be written as: $\mathbf{y}_{jk} = \boldsymbol{\mu} + \mathbf{b}_j + \mathbf{e}_{jk}$, where $\mathbf{b}_j \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_B)$ and $\mathbf{e}_{jk} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_W)$ with $\boldsymbol{\Sigma}_B = \{\sigma_{B_{ii'}}\}$

representing the $(p \times p)$ matrix of between-family components of variance and covariance between environments and $\mathbf{\Sigma}_W = \text{diag}\{\sigma^2_{W_i}\}$ for the $(p \times p)$ diagonal matrix of residual components of variance.

### *Equivalent heteroskedastic univariate models for $H_0$*

### $H_0$: constant genetic correlation between environments

The null hypothesis ($H_0$) considered here consists of assuming homogeneous genetic correlation coefficients $\rho_{ii'} = (\sigma_{B_{ii'}}/\sigma_{B_i}\sigma_{B_{i'}})$ between environments ($\rho_{ii'} = \rho$, $\forall i, i'$ and $i \neq i'$) without making any assumption about the residual variances $\mathbf{\Sigma}_e = \text{diag}\{\sigma^2_{e_i}\}$. Until now, we were unable to solve the problem of estimating the corresponding parameters by maximum likelihood (ML) procedures under the multiple trait approach in [1] even for balanced cross-classified designs (Foulley *et al*, 1994). An alternative is to tackle this issue *via* the concept of equivalent models (Henderson, 1984). Actually, an equivalent model to [1] under $H_0$ and restricted to $\rho > 0$ can be written using the following 2-way univariate mixed model with interaction:

$$y_{ijk} = \mu + h_i + \sigma_{s_i} s^*_j + \lambda \sigma_{s_i} h s^*_{ij} + e_{ijk} \qquad [2]$$

where $\mu$ is the mean, $h_i$ is the fixed effect of the $i$th environment; $\sigma_{s_i} s^*_j$ is the random family $j$ contribution such that $s^*_j \sim \text{NID}(0,1)$ and $\sigma^2_{s_i}$ is the family variance for records in the $i$th environment; $\lambda \sigma_{s_i} h s^*_{ij}$ is the random family $\times$ environment interaction effect such that $hs^*_{ij} \sim \text{NID}(0,1)$ and $\lambda^2 \sigma^2_{s_i}$ is the interaction variance for records in the $i$th environment; and $e_{ijk}$ is the residual effect assumed $\text{NID}(0, \sigma^2_{e_i})$. Models [1] under $H_0$ (and for $\rho > 0$) and [2] generate the same number of estimable parameters and the equalities necessary to obtain the same variance covariance structures are:

$$\begin{aligned}
&\text{Cov}(y_{ijk}, y_{ijk'}) = \sigma^2_{B_i} = \sigma^2_{s_i}(1+\lambda^2) \\
&\text{Cov}(y_{ijk}, y_{i'jk'}) = \sigma_{B_{ii'}} = \rho \sigma_{B_i} \sigma_{B_{i'}} = \sigma_{s_i} \sigma_{s_{i'}} \\
&\text{Var}(y_{ijk}) = \sigma^2_{B_i} + \sigma^2_{W_i} = \sigma^2_{s_i}(1+\lambda^2) + \sigma^2_{e_i}
\end{aligned}$$

These are met given the following 3 one-to-one relationships:

$$\sigma^2_{B_i} = \sigma^2_{s_i}(1+\lambda^2), \forall i \qquad [3a]$$

$$\sigma^2_{W_i} = \sigma^2_{e_i}, \forall i \qquad [3b]$$

$$\rho = 1/(1+\lambda^2) \quad \text{for} \quad \rho > 0 \qquad [3c]$$

### $H_0$: constant genetic and intra-class correlations between environments

In this part, the null hypothesis ($H_0$) consists of assuming homogeneous genetic and intra-class correlations between environments (*ie*, $\rho_{ii'} = \sigma_{B_{ii'}}/\sigma_{B_i}\sigma_{B_{i'}} = \rho$ and $t = \sigma^2_{B_i}/(\sigma^2_{B_i} + \sigma^2_{W_i}) = t \ \forall i, i'$ and $i \neq i'$). The variance covariance structure of the

residual is always assumed to be diagonal and heteroskedastic ($\mathbf{\Sigma}_e = \text{diag}\{\sigma^2_{e_i}\}$). As in the case of the above hypothesis of constant genetic correlation between environments only, an equivalent model to [1] under $H_0$ and restricted to $\rho > 0$ can be written as:

$$y_{ijk} = \mu + h_i + \tau\sigma_{e_i}s^*_j + \omega\sigma_{e_i}hs^*_{ij} + e_{ijk} \quad [4]$$

where $\mu$ and $h_i$ are the mean and the fixed effects of the $i$th environment respectively; $\tau\sigma_{e_i}s^*_j$ is the random family $j$ effect such that $s^*_j \sim \text{NID}(0,1)$ and $\tau^2\sigma^2_{e_i}$ is the family variance in the $i$th environment; $\omega\sigma_{e_i}hs^*_{ij}$ is the random family $\times$ environment interaction effect such that $hs^*_{ij} \sim \text{NID}(0,1)$ and $\omega^2\sigma^2_{e_i}$ is the interaction variance in the $i$th environment and $e_{ijk}$ is the residual effect assumed $\text{NID}(0,\sigma^2_{e_i})$. In the same way, the relationships between models [1] under $H_0$ (and for $\rho > 0$) and [4] are:

$$\sigma^2_{B_i} = \sigma^2_{e_i}(\tau^2 + \omega^2), \forall i \quad [5a]$$

$$\sigma^2_{B_{ii'}} = \tau^2\sigma_{e_i}\sigma_{e_{i'}}, \forall i, i' \quad [5b]$$

$$\sigma^2_{W_i} = \sigma^2_{e_i}, \forall i \quad [5c]$$

$$\rho = \tau^2/(\tau^2 + \omega^2) \text{ for } \rho > 0 \quad [5d]$$

$$t = (\tau^2 + \omega^2)/(\tau^2 + \omega^2 + 1) \quad [5e]$$

Notice that under the univariate model [4], the null hypothesis is tantamount to assuming constant ($\tau^2 = \sigma^2_{s_i}/\sigma^2_{e_i}$ ; $\omega^2 = \sigma^2_{hs_i}/\sigma^2_{e_i}$) ratios of variances between environments.

## Testing procedure

The theory of the likelihood ratio test (LRT) can be applied as previously proposed by Foulley *et al* (1990, 1992), Shaw (1991) and Visscher (1992) among others. Let $H_0: \boldsymbol{\gamma} \in \Gamma_0$ be the null hypothesis and $H_1: \boldsymbol{\gamma} \in \Gamma - \Gamma_0$ its alternative, where $\boldsymbol{\gamma}$ is the vector of genetic and residual parameters, $\Gamma$ refers to the complete parameter space and $\Gamma_0$ a subset of it pertaining to $H_0$. The likelihood under the null hypothesis (one of the 2 described above) is obtained by constraining the ratio(s) to be constant and finding the maximum under this constraint. The magnitude of the difference between the value of the likelihood obtained under the null hypothesis and the maximum of the likelihood obtained under the saturated model indicates the strength of evidence against the null hypothesis. Under $H_0$, the statistic:

$$\delta(\mathbf{y}) = 2\text{Max}_{\Gamma}\text{L}(\boldsymbol{\gamma};\mathbf{y}) - 2\text{Max}_{\Gamma_0}\text{L}(\boldsymbol{\gamma};\mathbf{y}) \quad [6]$$

(where $\text{L}(\boldsymbol{\gamma};\mathbf{y})$ is the log-likelihood) is expected to be distributed as a chi-square with $r$ degrees of freedom given by the difference between the number of parameters specifying the saturated model and the number of parameters estimated under the null hypothesis. $H_0$ is rejected at the level $\alpha$ if $\delta \geqslant \delta_0$ where $\Pr[\chi^2_r \geqslant \delta_0] = \alpha$. Since the parameters involved here are variance components, the LRT that has desirable asymptotic properties is applied using restricted maximum likelihood

(REML) rather than ML estimators (Patterson and Thompson, 1971; Harville, 1974). Formulae to evaluate $-2\text{MaxL}(\boldsymbol{\gamma}; \mathbf{y})$ under this saturated model were given by Foulley *et al* (1994).

## An EM-REML algorithm for models [2] and [4]

Models [2] and [4] can be written more generally using matrix notation.

For model [2]: $$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sigma_{u_i}\mathbf{Z}_{1i}\mathbf{u}_1^* + \lambda\sigma_{u_i}\mathbf{Z}_{2i}\mathbf{u}_2^* + \mathbf{e}_i \quad [7]$$

For model [4]: $$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \tau\sigma_{e_i}\mathbf{Z}_{1i}\mathbf{u}_1^* + \omega\sigma_{e_i}\mathbf{Z}_{2i}\mathbf{u}_2^* + \mathbf{e}_i \quad [8]$$

where $\mathbf{y}_i$ is a $(n_i \times 1)$ vector of observations in environment $i$; $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of fixed effects with incidence matrix $\mathbf{X}_i$; $\mathbf{u}_1^* = \{s_j^*\}$ and $\mathbf{u}_2^* = \{hs_{ij}^*\}$ are 2 independent random normal components of the model (in this case, family and interaction effects respectively) with incidence matrices for standardized effects $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ respectively; $\sigma_{u_i}$ and $\sigma_{e_i}$ being the $u$-component and residual components of variances respectively, pertaining to stratum $i$, and $\mathbf{e}_i$ is the vector of residuals for stratum $i$ assumed $\text{N}(\mathbf{0}, \sigma_{e_i}^2\mathbf{I}_{n_i})$.

The 'expectation-maximization' (EM) approach is a very efficient concept in ML estimation (Dempster *et al*, 1977) and this algorithm is frequently advocated for estimating variance components in linear models (Quaas, 1992). The generalized EM procedure to compute REML estimators of dispersion parameters, as described by Foulley and Quaas (1994) for one-way heteroskedastic mixed models, can be applied here. Letting $\mathbf{u}^* = (\mathbf{u}_1^{*'}, \mathbf{u}_2^{*'})'$, $\boldsymbol{\sigma}_u^2 = \{\sigma_{u_i}^2\}$, $\boldsymbol{\sigma}_e^2 = \{\sigma_{e_i}^2\}$, $\boldsymbol{\gamma}_1 = (\sigma_u^{2'}, \sigma_e^{2'}, \lambda)'$ and $\boldsymbol{\gamma}_2 = (\tau, \omega, \sigma_e^{2'})'$ being the 2 sets of estimable parameters for the models [7] and [8] respectively (later on denoted as $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1$ or $\boldsymbol{\gamma} = \boldsymbol{\gamma}_2$), the $E$ step consists of computing the function $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) = E_c^{[t]}\left[\ln p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}^*, \boldsymbol{\gamma})\right]$ where the expectation between brackets is taken with respect to the distribution of $\boldsymbol{\beta}$, $\mathbf{u}^*$ given $\mathbf{y}$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{[t]}$, $\boldsymbol{\gamma}^{[t]}$ being the current estimate of $\boldsymbol{\gamma}$ at iteration $[t]$. The M step consists of selecting the next value $\boldsymbol{\gamma}^{[t+1]}$ of $\boldsymbol{\gamma}$ by maximizing $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]})$ with respect to $\boldsymbol{\gamma}$. This EM-REML algorithm can also be derived using Bayesian arguments (Foulley *et al*, 1987; Foulley and Gianola, 1989). For models [7] and [8], the function to be maximized:

$$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) = \text{Const} - (1/2)\sum_{i=1}^{p} n_i \ln(\sigma_{e_i}^2) - (1/2)\sum_{i=1}^{p} \sigma_{e_i}^{-2} E_c^{[t]}\left[\mathbf{e}_i'\mathbf{e}_i\right] \quad [9]$$

For model [7], the differentiation of expression [9] with respect to $\lambda$, $\sigma_{u_i}^2$ and $\sigma_{e_i}^2$ yields:

$$\partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) / \partial\lambda = -(1/2)\sum_{i=1}^{p} \sigma_{e_i}^{-2}\left[\partial E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i) / \partial\lambda\right] \quad [10a]$$

$$\partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) / \partial\sigma_{u_i}^2 = -(1/2)\sigma_{e_i}^{-2}\left[\partial E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i) / \partial\sigma_{u_i}^2\right] \quad [10b]$$

$$\partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) / \partial\sigma_{e_i}^2 = -\left[n_i / 2\sigma_{e_i}^2\right] + \left[E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i) / 2\sigma_{e_i}^4\right] \quad [10c]$$

For model [8], differentiating the function [9] with respect to $\tau$, $\omega$ and $\sigma^2_{e_i}$, we get:

$$\partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) / \partial\tau = -(1/2)\sum_{i=1}^{p}\sigma_{e_i}^{-2}\left[\partial E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i) / \partial\tau\right] \quad [11a]$$

$$\partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) / \partial\omega = -(1/2)\sum_{i=1}^{p}\sigma_{e_i}^{-2}\left[\partial E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i) / \partial\omega\right] \quad [11b]$$

$$\partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) / \partial\sigma^2_{e_i} = -\left[n_i / 2\sigma^2_{e_i}\right] - (1/2)\left[\partial\left\{\sigma_{e_i}^{-2}E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i)\right\} / \partial\sigma^2_{e_i}\right] \quad [11c]$$

The corresponding system $\partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) / \partial\boldsymbol{\gamma} = 0$ cannot simply be written as a linear system, as in the case with a saturated model, because the interaction variance in model [7] is proportional to the family variance in environment $i$, and the interaction and family variances in model [8] are proportional to the residual variance in environment $i$. A convenient way of solving it is to use the method of 'cyclic ascent' (Zangwill, 1969). For instance, let us consider model [7]. The different steps to implement in this procedure starting with $\lambda^{[t]}$, $\sigma^{2[t]}_{u_i}$ and $\sigma^{2[t]}_{e_i}$ are as follows: (1) solve [10a] = 0 with respect to $\lambda$; (2) substitute the solution $\lambda^{[t,1]}$ to $\lambda$ back into $E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i)$ of [10b] = 0; (3) solve that equation; (4) substitute $\lambda^{[t,1]}$ and $\sigma^{2[t,1]}_{u_i}$ to $\lambda$ and $\sigma^2_{u_i}$ back into $E_c^{[t]}(\mathbf{e}_i'\mathbf{e}_i)$ of [10c] = 0; (5) solve for $\sigma^2_{e_i}$; and (6) return to [10a], [10b] and [10c] for a second inner cycle yielding $\lambda^{[t,2]}$, $\sigma^{2[t,2]}_{u_i}$ and $\sigma^{2[t,2]}_{e_i}$ and continue to $\lambda^{[t,c]}$, $\sigma^{2[t,c]}_{u_i}$ and $\sigma^{2[t,c]}_{e_i}$ (convergence at iteration $c$). Finally, take $\lambda^{[t+1]} = \lambda^{[t,c]}$, $\sigma^{2[t+1]}_{u_i} = \sigma^{2[t,c]}_{u_i}$ and $\sigma^{2[t+1]}_{e_i} = \sigma^{2[t,c]}_{e_i}$. In practice, it may be advantageous to reduce the number of inner iterations even down to only one.

For model [7], the algorithm can be summarized as:

$$\lambda^{[t,l+1]} = \frac{\sum_{i=1}^{p}\left(\sigma^{[t,l]}_{u_i} / \sigma^{2[t,l]}_{e_i}\right)E_c^{[t]}\left[\mathbf{u}_2^{*'}\mathbf{Z}_{2i}'\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \sigma^{[t,l]}_{u_i}\mathbf{Z}_{1i}\mathbf{u}_1^*\right)\right]}{\sum_{i=1}^{p}\left(\sigma^{2[t,l]}_{u_i} / \sigma^{2[t,l]}_{e_i}\right)E_c^{[t]}\left[\mathbf{u}_2^{*'}\mathbf{Z}_{2i}'\mathbf{Z}_{2i}\mathbf{u}_2^*\right]} \quad [12a]$$

$$\sigma^{[t,l+1]}_{u_i} = \frac{E_c^{[t]}\left[\left(\mathbf{u}_2^{*'}\mathbf{Z}_{2i}'\lambda^{[t,l+1]} + \mathbf{u}_1^{*'}\mathbf{Z}_{1i}'\right)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right]}{E_c^{[t]}\left[\left(\mathbf{u}_2^{*'}\mathbf{Z}_{2i}'\lambda^{[t,l+1]} + \mathbf{u}_1^{*'}\mathbf{Z}_{1i}'\right)\left(\mathbf{Z}_{1i}\mathbf{u}_1^* + \lambda^{[t,l+1]}\mathbf{Z}_{2i}\mathbf{u}_2^*\right)\right]} \quad [12b]$$

$$\sigma^{2[t,l+1]}_{e_i} = \frac{E_c^{[t]}\left[\mathbf{e}_i^{'[t,l+1]}\mathbf{e}_i^{[t,l+1]}\right]}{n_i} \quad [12c]$$

with $\mathbf{e}_i^{[t,1+l]} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \sigma^{[t,l+1]}_{u_i}\left[\mathbf{Z}_{1i}\mathbf{u}_1^* + \lambda^{[t,l+1]}\mathbf{Z}_{2i}\mathbf{u}_2^*\right]$.

Similarly for model [8], we obtain the following algorithm:

$$\tau^{[t,l+1]} = \frac{\sum_{i=1}^{p} \left(1/\sigma_{e_i}^{[t,l]}\right) E_c^{[t]} \left[\mathbf{u}_1^{*'} \mathbf{Z}_{1i}' \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \omega^{[t,l]}\sigma_{e_i}^{[t,l]}\mathbf{Z}_{2i}\mathbf{u}_2^*\right)\right]}{\sum_{i=1}^{p} E_c^{[t]} \left[\mathbf{u}_1^{*'} \mathbf{Z}_{1i}' \mathbf{Z}_{1i} \mathbf{u}_1^*\right]} \quad [13a]$$

$$\omega^{[t,l+1]} = \frac{\sum_{i=1}^{p} \left(1/\sigma_{e_i}^{[t,l]}\right) E_c^{[t]} \left[\mathbf{u}_2^{*'} \mathbf{Z}_{2i}' \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \tau^{[t,l+1]}\sigma_{e_i}^{[t,l]}\mathbf{Z}_{1i}\mathbf{u}_1^*\right)\right]}{\sum_{i=1}^{p} E_c^{[t]} \left[\mathbf{u}_2^{*'} \mathbf{Z}_{2i}' \mathbf{Z}_{2i} \mathbf{u}_2^*\right]} \quad [13b]$$

$$\sigma_{e_i}^{[t,l+1]} = \frac{-E_c^{[t]} \left[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'(\tau^{[t,l+1]}\mathbf{Z}_{1i}\mathbf{u}_1^* + \omega^{[t,l+1]}\mathbf{Z}_{2i}\mathbf{u}_2^*)\right]}{2n_i} +$$

$$\frac{\sqrt{\left\{E_c^{[t]} \left[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})' \left(\tau^{[t,l+1]}\mathbf{Z}_{1i}\mathbf{u}_1^* + \omega^{[t,l+1]}\mathbf{Z}_{2i}\mathbf{u}_2^*\right)\right]\right\}^2 + 4n_i E_c^{[t]} \left[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right]}}{2n_i} \quad [13c]$$

with $\mathbf{e}_i^{[t,l+1]} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \sigma_{e_i}^{[t,l+1]} \left[\tau^{[t,l+1]}\mathbf{Z}_{1i}\mathbf{u}_1^* + \omega^{[t,l+1]}\mathbf{Z}_{2i}\mathbf{u}_2^*\right]$.

$E_c^{[t]}(.)$ can be expressed as the sum of a quadratic form and the trace of parts of the inverse coefficient matrix of the mixed model equations (as described in Foulley and Quaas, 1994). Note also that simple forms of [12b] and [13c] involve the standard deviation and not the variance component, as explained in Foulley and Quaas (1994).

## ILLUSTRATION

The procedures presented in this paper are illustrated with the analysis of an experiment carried out on 20 full-sib families of black medic (*Medicago lupulina* L) tested in 3 different environments (harvesting, control and competition treatments). The experimental design was described in detail by Hébert (1991). There were 2 replicates per environment and the 20 genotypes were randomly allocated to each replicate (Foulley *et al*, 1994). As an illustrative example, we consider 5 vegetative and reproductive traits out of the 36 traits which have been recorded. Table I presents the estimation of genetic and residual parameters under the saturated model. Table II presents the result of the estimation of (co)variance components under the reduced (hypothesis of homogeneity of genetic correlations between environments) model and the likelihood ratio test of this reduced model against the saturated model. Similarly, table III presents similar results but in which the reduced model considered represents the hypothesis of homogeneity of genetic and intra-class correlations between environments. Table III also presents the

**Table I.** Estimation of genetic and residual parameters between environments[a] under the saturated model.

| | *Environment* | *[1] Days to flowering (d)* | *[2] Days to 1st ripe pod (d)* | *[3] Dry matter weight (g)* | *[4] Dry matter weight/max plant size ($dgcm^{-1}$)* | *[5] Pod weight/total weight (%)* |
|---|---|---|---|---|---|---|
| Residual components | 11 | 13.94 | 11.69 | 156.04 | 0.83 | 27.28 |
| | 22 | 39.94 | 21.59 | 512.03 | 2.99 | 10.94 |
| | 33 | 15.51 | 8.02 | 46.35 | 0.26 | 19.59 |
| Between-family components | 11 | 52.55 | 43.68 | 337.85 | 2.01 | 96.57 |
| | 22 | 100.46 | 37.20 | 1 154.86 | 5.73 | 83.82 |
| | 33 | 99.63 | 35.49 | 193.75 | 1.07 | 70.33 |
| | 12 | 69.47 | 33.45 | 556.39 | 3.07 | 81.02 |
| | 13 | 68.47 | 34.83 | 207.15 | 1.12 | 81.98 |
| | 23 | 99.98 | 35.00 | 467.30 | 2.40 | 72.18 |
| Genetic correlation | 12 | 0.96 | 0.83 | 0.89 | 0.90 | 0.90 |
| | 13 | 0.95 | 0.88 | 0.81 | 0.76 | 0.99 |
| | 23 | 0.99 | 0.96 | 0.99 | 0.97 | 0.94 |
| Intra-class correlations | 11 | 0.79 | 0.79 | 0.68 | 0.71 | 0.78 |
| | 22 | 0.72 | 0.63 | 0.69 | 0.66 | 0.88 |
| | 33 | 0.87 | 0.82 | 0.81 | 0.80 | 0.78 |
| $-2L^b$ | | 540.51 | 487.78 | 774.74 | 169.89 | 534.41 |

[a] 1, 2, 3 = 'harvesting', 'control', 'competition' environments, respectively; [b] −2log-likelihood.

likelihood ratio test of the reduced model ($H_0$: homogeneity of genetic and intra-class correlations between environments) against the reduced model of table II ($H_1$: homogeneity of genetic correlations between environments only).

Convergence of the EM-REML procedure was measured as the norm of the vector of changes in genetic parameters between iterations. A norm less than $10^{-6}$ was obtained after 150 iterations (the number of inner iterations was only one) and the computing time was less than 10 CPU seconds per trait (on an IBM 3090-17T computer).

The results in table II suggest that differences among genetic correlations are not statistically significant (except perhaps for trait [4] with $P$-value of 0.07). $P$-values for vegetative and reproductive yields traits represented here by traits [1], [2] and [3] were very high, indicating a lack of heterogeneity in genetic correlations between environments. It seems that the overall correlation under the reduced model (table II) is much larger than a simple average of the 3 estimates under the saturated model. These results are due to one pair of environments with a genetic correlation of 0.99, which pushes the overall correlation also to 0.99. In table III (tests 1 or 2), $P$-values also indicate that there are no significant differences between ratios of variances between environments, indicating a homogeneity in genetic and intra-class variation between environments.

It can be concluded that the harvesting and competition environments do not generate a meaningful level of stress as compared to the control environment for the expression of genetic and intra-class variation of all traits analyzed. These results can be due to the small sample size (only 40 records per environment). Since genetic correlations between environments were very high and close to one, it is interesting to test for these traits the assumption of these correlations being equal to one. We have thus tested the model under the hypothesis of constant genetic correlations and equal to one (according to the procedure described in Foulley and Quaas (1994)) against the reduced model (hypothesis of homogeneity of genetic correlations). $P$-values for all traits analyzed (except for trait [2] where the $P$-value was equal to 0.1) were very high and indicated that these correlations did not differ from one.

## DISCUSSION AND CONCLUSION

This paper clearly illustrates the value of univariate heteroskedastic models (Foulley *et al*, 1990, 1992; Gianola *et al*, 1992; San Cristobal *et al*, 1993) to tackle problems of estimation and hypothesis testing of genetic parameters arising in genotype × environment data structures. It was shown that under each null hypothesis, constant genetic correlations between environments and constant genetic and intra-class correlations between environments, multiple trait and univariate linear models generated the same number of estimable parameters and that there were one-to-one relationships between both models. However, it should be noticed that strictly speaking the univariate linear model under $H_0$ (either hypothesis) is defined only under $\rho > 0$ because negative variances are by definition not possible. Caution must thus be exercised in applying the univariate linear model as an equivalent multiple trait linear model. This last model is obviously more flexible, as previously pointed out by Mallard *et al* (1983).

**Table II.** Estimation of variance and covariance components and genetic correlations between environments[a] under the reduced (constant genetic correlation) model and test of this model against the saturated model.

| | *Environment* | *[1] Days to flowering (d)* | *[2] Days to 1st ripe pod (d)* | *[3] Dry matter weight (g)* | *[4] Dry matter weight/max plant size ($dgcm^{-1}$)* | *[5] Pod weight/total weight (%)* |
|---|---|---|---|---|---|---|
| Residual components | 11 | 17.52 | 12.62 | 233.40 | 1.13 | 31.52 |
| | 22 | 40.08 | 21.28 | 513.41 | 2.99 | 14.11 |
| | 33 | 16.34 | 7.65 | 51.36 | 0.29 | 20.87 |
| Between-family components | 11 | 48.88 | 42.97 | 258.44 | 1.71 | 92.20 |
| | 22 | 100.31 | 37.50 | 1 153.40 | 5.70 | 80.64 |
| | 33 | 98.77 | 35.69 | 188.60 | 1.04 | 68.96 |
| | 12 | 70.02 | 35.97 | 545.97 | 2.95 | 84.63 |
| | 13 | 69.48 | 35.09 | 220.78 | 1.26 | 78.27 |
| | 23 | 99.54 | 32.78 | 466.40 | 2.30 | 73.20 |
| Genetic correlation between environments | $\rho$ | 0.99 | 0.90 | 0.99 | 0.94 | 0.98 |
| Intra-class correlations | 11 | 0.74 | 0.77 | 0.53 | 0.60 | 0.75 |
| | 22 | 0.71 | 0.64 | 0.69 | 0.66 | 0.85 |
| | 33 | 0.86 | 0.82 | 0.79 | 0.78 | 0.77 |
| $-2L$[b] | | 541.69 | 489.24 | 778.19 | 175.11 | 539.13 |
| *Test* | | | | | | |
| $\delta$[c] | | 1.18 | 1.46 | 3.45 | 5.22 | 4.72 |
| $P$-value[d] | | 0.55 | 0.48 | 0.18 | 0.07 | 0.09 |

[a] 1, 2, 3 = 'harvesting', 'control', 'competition' environments respectively; [b] −2log-likelihood; [c] likelihood ratio statistic; [d] degrees of freedom = 2.

**Table III.** Estimation of variance and covariance components, genetic and intra-class correlations between environments[a] under the reduced (constant ratios) model and test of this model against the saturated model (*test 1*) or against the reduced model of Table II (*test 2*).

| | *Environment* | *[1] Days to flowering (d)* | *[2] Days to 1st ripe pod (d)* | *[3] Dry matter weight (g)* | *[4] Dry matter weight/max plant size ($dgcm^{-1}$)* | *[5] Pod weight/total weight (%)* |
|---|---|---|---|---|---|---|
| Residual components | 11 | 15.48 | 13.51 | 165.65 | 0.93 | 27.77 |
| | 22 | 33.75 | 16.22 | 531.47 | 2.82 | 20.54 |
| | 33 | 25.22 | 9.98 | 72.60 | 0.40 | 19.86 |
| Between-family components | 11 | 51.68 | 40.90 | 354.08 | 2.01 | 98.02 |
| | 22 | 112.64 | 49.13 | 1 136.06 | 6.10 | 72.47 |
| | 33 | 84.16 | 30.21 | 155.19 | 0.86 | 70.10 |
| | 12 | 76.14 | 39.53 | 618.77 | 3.29 | 84.28 |
| | 13 | 65.81 | 31.00 | 228.70 | 1.24 | 82.89 |
| | 23 | 97.17 | 33.98 | 409.65 | 2.16 | 71.28 |
| Genetic correlation | $\rho$ | 1.00 | 0.88 | 0.98 | 0.94 | 1.00 |
| Intra-class correlations | $t^2$ | 0.77 | 0.75 | 0.68 | 0.68 | 0.78 |
| $-2L$[b] | | 545.31 | 492.06 | 781.53 | 176.78 | 539.95 |
| *Test 1* | | | | | | |
| $\delta$[c] | | 4.80 | 4.28 | 6.79 | 6.89 | 5.54 |
| $P$-value[d] | | 0.31 | 0.37 | 0.15 | 0.14 | 0.24 |
| *Test 2* | | | | | | |
| $\delta$[c] | | 3.62 | 2.82 | 3.34 | 1.67 | 0.82 |
| $P$-value[e] | | 0.16 | 0.24 | 0.19 | 0.43 | 0.66 |

[a] 1, 2, 3 = 'harvesting', 'control', 'competition' environments respectively; [b] −2log-likelihood; [c] likelihood ratio statistic; [d] degree of freedom = 4; [e] degree of freedom = 2.

The EM algorithm seems a natural choice for the estimation of variance components in univariate linear models but methods other than EM (ECME, Liu and Rubin, 1994; Newton Raphson; quasi-Newton method based on average information, Johnson and Thompson, 1994; derivative-free, Meyer, 1989) can be used to solve this problem. The EM-REML approach presented in this paper is quite flexible. It can accommodate any structure of fixed effects and nondiagonal patterns of the variance-covariance matrices of $\mathbf{u}_1^*$ and $\mathbf{u}_2^*$, $\text{Var}(\mathbf{u}_1^*) = \mathbf{A}_1$ and $\text{Var}(\mathbf{u}_2^*) = \mathbf{A}_2$, *ie* for the particular model in [2] (Foulley and Henderson, 1989) $\text{Var}(\mathbf{s}^*) = \mathbf{A}\sigma^2_{s_i}$, $\text{Var}(\mathbf{hs}^*) = \mathbf{I}_p \otimes \mathbf{A}\sigma^2_{hs_i}$ with $\sigma^2_{hs_i} = \lambda^2\sigma^2_{s_i}$, $\mathbf{s}^* = \{s_j^*\}$, $\mathbf{hs}^* = \{hs_{ij}^*\}$ and $\mathbf{A}$ is the additive genetic relationship matrix.

Evidently, the approaches presented in this paper apply to an unbalanced structure of data and to additional nuisance fixed effects cross-classified with family effects, using the formulae defined in [12abc] and [13abc]. These algorithms can also be utilized for the homoskedastic case by just taking $i$ equal to 1 in the previous formulae. This means that several EM-REML algorithms are presently available to calculate REML estimates of variance components under the standard homoskedastic linear model: (i) the classical EM algorithm based on sufficient statistics; (ii) the related EM of EM-type algorithms (Henderson, 1973; Harville, 1974; Callanan, 1985); and (iii) the generalized EM algorithms proposed by Foulley and Quaas (1994) for models parameterized either with variance components or as in this paper. But additional work is needed to compare the performance of these different algorithms.

Finally, the null hypothesis of constant intra-class correlations without making any assumption on genetic correlations between environments remains to be considered. This problem requires a special treatment as far as the parameterization of the model is concerned and will be reported in a separate article.

## ACKNOWLEDGMENTS

The authors wish to thank I Olivieri and D Hébert (INRA-Montpellier) for providing the data set and raising the issue of estimation and testing genetic parameters in experiments of this kind.

## REFERENCES

Callanan TP (1985) Restricted maximum likelihood estimation of variance components: computational aspects. PhD thesis, Iowa State University, Ames, USA

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data *via* the EM algorithm. *J R Statist Soc B* 39, 1-38

Falconer DS (1952) The problem of environment and selection. *Am Nat* 86, 293-298

Foulley JL, Im S, Gianola D, Hoeschele I (1987) Empirical Bayes estimation of parameters for $n$ polygenic binary traits. *Genet Sel Evol* 19, 197-224

Foulley JL, Gianola D (1989) A simple algorithm for computing marginal maximum likelihood estimates of variance components and its relation to EM. 47th ISI Meeting Paris, Bull Inst Int Stat, Paris, France, vol I, 337-338

Foulley JL, Gianola D, San Cristobal M, Im S (1990) A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. *J Dairy Sci* 73, 1612-1624

Foulley JL, San Cristobal M, Gianola D, Im S (1992) Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Comput Stat Data Anal* 13, 291-305

Foulley JL, Quaas RL (1994) Statistical analysis of heterogeneous variances in Gaussian linear mixed models. *Proc 5th World Congress Genet Appl Livest Prod*, Univ Guelph, Guelph, ON, Canada, 18, 341-348

Foulley JL, Hébert D, Quaas RL (1994) Inference on homogeneity of between-family components of variance and covariance among environments in balanced cross-classified designs. *Genet Sel Evol* 26, 117-136

Gianola D, Foulley JL, Fernando RL, Henderson CR, Weigel KA (1992) Estimation of heterogeneous variances using empirical Bayes methods: theoretical considerations. *J Dairy Sci* 75, 2805-2823

Harville DA (1974) Bayesian inference for variance components using only error constrats. *Biometrika* 61, 393-408

Hébert D (1991) Plasticité phénotypique et interaction génotype milieu chez *Medicago lupulina*. Thèse de Doctorat en Sciences. Université des Sciences et Techniques du Languedoc, Montpellier, France

Henderson CR (1973) Sire evaluation and genetic trends. *In: Proc Anim Breed Genet Symp in honor of Dr J Lush.* Amer Soc Anim Sci, Amer Dairy Sci Assoc, 10-41, Champaign, IL, USA

Henderson CR (1984) *Applications of Linear Models in Animal Breeding.* Univ Guelph, Guelph, ON, Canada

Johnson DL, Thompson R (1994) Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and a quasi-Newton procedure. *Proc 5th World Congress Genet Appl Livest Prod*, Univ Guelph, Guelph, ON, Canada, 18, 410-413

Liu C, Rubin DB (1994) Applications of the ECME algorithm and the Gibbs sampler to general linear mixed models. *XVIIth International Biometric conference*, McMaster Univ, Hamilton, ON, Canada, vol 1, 97-107

Mallard J, Masson JP, Douaire M (1983) Interaction génotype × milieu et modèle mixte: I-Modélisation. *Genet Sel Evol* 15, 379-394

Meyer K (1989) Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet Sel Evol* 21, 317-340

Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. *Biometrika* 58, 545-554

Quaas RL (1992) *REML Notebook.* Mimeo, Dept Anim Sci, Cornell Univ, Ithaca, NY, USA

San Cristobal M, Foulley JL, Manfredi E (1993) Inference about multiplicative heteroskedastic components of variance in a mixed linear Gaussian model with an application to beef cattle breeding. *Genet Sel Evol* 25, 3-30

Shaw RG (1991) The comparison of quantitative genetic parameters between populations. *Evolution* 45, 143-151

Visscher PM (1992) On the power of likelihood ratio test for detecting heterogeneity of intra-class correlations and variances in balanced half-sib designs. *J Dairy Sci* 73, 1320-1330

Zangwill (1969) *Non-linear Programming: A Unified Approach.* Prentice-Hall, Englewood Cliffs, NJ, USA