Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maximum Likelihood Estimation with SAS Proc MIXED

ARTICLE *in* CROP SCIENCE · MARCH 2006

Impact Factor: 1.58 · DOI: 10.2135/cropsci2005.0191 · Source: OAI

1 AUTHOR:

James B Holland
North Carolina State University
127 PUBLICATIONS 5,867 CITATIONS

SEE PROFILE

Published online February 1, 2006

Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maximum Likelihood Estimation with SAS Proc MIXED

James B. Holland*

ABSTRACT

Plant breeders traditionally have estimated genotypic and phenotypic correlations between traits using the method of moments on the basis of a multivariate analysis of variance (MANOVA). Drawbacks of using the method of moments to estimate variance and covariance components include the possibility of obtaining estimates outside of parameter bounds, reduced estimation efficiency, and ignorance of the estimators' distributional properties when data are missing. An alternative approach that does not suffer these problems, but depends on the assumption of normally distributed random effects and large sample sizes, is restricted maximum likelihood (REML). This paper illustrates the use of Proc MIXED of the SAS system to implement REML estimation of genotypic and phenotypic correlations. Additionally, a method to obtain approximate parametric estimates of the sampling variances of the correlation estimates is presented. MANOVA and REML methods were compared with a real data set and with simulated data. The simulation study examined the effects of different correlation parameter values, genotypic and environmental sample sizes, and proportion of missing data on Type I and Type II error rates and on accuracy of confidence intervals. The two methods provided similar results when data were balanced or only 5% of data were missing. However, when 15 or 25% data were missing, the REML method generally performed better, resulting in higher power of detection of correlations and more accurate 95% confidence intervals. Samples of at least 75 genotypes and two environments are recommended to obtain accurate confidence intervals using the proposed method.

Genotypic correlations between traits indicate the direction and magnitude of correlated responses to selection, the relative efficiency of indirect selection, and permit calculation of optimal multiple trait selection indices (Falconer and Mackay, 1996). Plant breeders traditionally have estimated genotypic and phenotypic correlations between traits using the method of moments on the basis of a multivariate extension of ordinary least squares referred to as multivariate analysis of variance (MANOVA; Anderson, 1958; Mode and Robinson, 1959). Drawbacks of using MANOVA method of moments to estimate variance and covariance components include ignorance of the estimators' distributional properties when data are unbalanced and the possibility of obtaining estimates outside of parameter bounds (Liu et al., 1997). Furthermore, MANOVA method of moments can suffer a loss of efficiency when some trait data are missing, because data on other traits measured on the same experimental units, although available, are not used.

An alternative approach to using moments estimators of the variance and covariance components that compose the estimates of genotypic and phenotypic correlation estimates is restricted maximum likelihood (REML). REML is often more computationally intensive than MANOVA, but advances in computer processing speed have made REML computationally feasible on modern personal computers. Animal breeders and quantitative geneticists have implemented REML-based estimates of genotypic and phenotypic correlations using specialized software packages, such as AS-REML (Berry et al., 2002; Gilmour et al., 1999; Persson and Andersson, 2003), VCE (Conington et al., 2001; Legarra and Ugarte, 2001; Neumaier and Groeneveld, 1998), MTDFREML (Boldman et al., 1993; Bureau et al., 2001), or their own programs (Zhu and Weir, 1996). Some plant breeders, primarily tree breeders, also have used specialized software packages for estimating genotypic and phenotypic correlations (de Souza et al., 1998), but more generally, plant breeders, particularly crop breeders, use general statistical packages, including the SAS system. Proc MIXED of SAS is a component of a general use statistical software package that will provide REML estimates of variance and covariance components among model factors and permits fitting both fixed and random model effects in mixed models analyses (Littell et al., 1996). The dense-matrix computational methods used by SAS Proc MIXED make it slower than the aforementioned genetic-specific software, but Proc MIXED can handle a wide variety of experimental and treatment design combinations. Multivariate REML analysis can be implemented with Proc MIXED by treating the two variables as two repeated measurements of a single variable on each experimental unit, because Proc MIXED is well designed to handle longitudinal (repeated measures) analyses (Littell et al., 1996; Wright, 1998). Recently, Fry (2004) explicitly demonstrated the use of Proc MIXED for combined variance-covariance estimation of two traits in a quantitative genetics framework, but he did not discuss precision of the estimates. The advantages of REML estimation compared with MANOVA method of moments are that REML estimates of the variance and covariance components have known asymptotic distributional properties and efficiently use information from all experimental units when data are unbalanced (Meyer, 1985).

USDA-ARS Plant Science Research Unit, Dep. of Crop Science, Box 7620, North Carolina State University, Raleigh, NC 27695-7620. Received 4 Mar. 2005. *Corresponding author (James_Holland@ncsu.edu).

Published in Crop Sci. 46:642–654 (2006).
Crop Breeding, Genetics & Cytology
doi:10.2135/cropsci2005.0191
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: CI, confidence interval; GDD, growing degree days; GEI, genotype × environment interaction; MANOVA, multivariate analysis of variance; MCAR, missing completely at random; MPAR, missing plots at random; PH, plant height; REML, restricted maximum likelihood.

A drawback of REML-based approaches is that the sampling distributions of the correlation estimates are usually not available in closed form and are likely to be nonnormal (Liu et al., 1997). However, the asymptotic dispersion matrix of the covariance components that compose the correlation estimates is available from the second derivatives of the REML optimization; therefore, approximate standard errors can be obtained with the delta method (Holland et al., 2003; Mode and Robinson, 1959). These approximate standard errors are valid for very large sample sizes (Searle et al., 1992), but their reliability for smaller sample sizes is not known. REML estimators of treatment variance components behave poorly even for one-way treatment classification designs with very small sample sizes (e.g., fewer than ten treatments) and unbalanced data (Swallow and Monahan, 1984). Therefore, the question remains: how large a sample of genotypes, environments, and replications is needed to obtain accurate REML-based estimators of genotypic correlations and their standard errors for typical plant breeding experiments?

Holland et al. (2001) used multivariate REML to estimate genotypic and phenotypic correlations and their approximate standard errors for grain oil content and other agronomic traits in oat (*Avena sativa* L.). Zamudio and Wolfinger (2002) used a similar approach to estimate genetic covariances between measurements made at different ages on trees, but they analyzed each location separately and did not attempt an across-locations analysis. The multivariate REML approach is more straightforward than that used by Singh et al. (1997), and it also permits parametric estimation of the sampling variances of the parameters.

The objectives of this paper are to: (i) describe SAS code to obtain genotypic and phenotypic correlation estimates and their approximate standard errors using multivariate REML on data from multiple environment trials for typical plant breeding experiments, (ii) demonstrate the utility of this approach using experimental data, and (iii) compare the validity of approximate 95% confidence intervals of genotypic correlation estimates on the basis of REML and MANOVA method of moments-based estimation using simulated data sets with different levels of genetic correlations, sample sizes, and amounts and distribution of missing data.

MATERIALS AND METHODS

Statistical Model

SAS code for converting a data set from a typical MANOVA format to an appropriate format for multivariate REML analysis using SAS Proc MIXED and for estimating genotypic and phenotypic correlations and their standard errors is presented for a multiple environment trial design commonly used in plant breeding (Appendix A).

I consider the situation where correlations are estimated by randomly sampling genotypes (or families) from a reference population and evaluating them in randomized complete block designs replicated two or more times in two or more macroenvironments. This includes any one-way classification of genotypes. Family structures could include half-sib families or doubled haploid, recombinant inbred, or other random inbred line populations. This does not include mating designs with two or more classification levels (e.g., diallel, or North Carolina mating designs I, II, or III, Hallauer and Miranda, 1988), but the general approach outlined here can be modified for application to these designs (see examples in Appendices B and C).

The linear model for balanced data on one trait, Y_i is:

$$Y_{klmi} = \mu_i + E_{ki} + \mathrm{R(E)}_{kli} + G_{mi} + \mathrm{GE}_{kmi} + \varepsilon_{klmi},$$

where μ_i is the mean effect on trait i, E_{ki} is the effect of macroenvironment k on trait i, $\mathrm{R(E)}_{kli}$ is the effect of replication l within environment k on trait i, G_{mi} is the effect of genotype (or family) m on trait i, GE_{kmi} is the effect of the interaction between genotype m and environment k on trait i, and ε_{klmi} is the experimental error effect associated with genotype m and replication l within environment k on trait i.

Observations of traits i and j on the same plot have the following covariance:

$$Cov(Y_{klmi},Y_{klmj}) = \sigma_{Eij} + \sigma_{R(E)ij} + \sigma_{Gij} + \sigma_{GEij} + \sigma_{\varepsilon ij}.$$

Observations of traits i and j on the same genotype grown in different replications within the same environment have the following covariance:

$$Cov(Y_{klmi},Y_{kl'mj}) = \sigma_{Eij} + \sigma_{Gij} + \sigma_{GEij}.$$

Observations of traits i and j on the same genotype grown in different environments have the following covariance:

$$Cov(Y_{klmi},Y_{k'l'mj}) = \sigma_{Gij}.$$

The joint model for two traits, Y_i and Y_j, is:

$$\begin{bmatrix}\mathbf{y}_i\\ \mathbf{y}_\mathbf{j}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}_\mathbf{i}\\ \boldsymbol{\mu}_\mathbf{j}\end{bmatrix} + \begin{bmatrix}\mathbf{T}_i & 0\\ 0 & \mathbf{T}_j\end{bmatrix}\begin{bmatrix}\mathbf{E}_\mathbf{i}\\ \mathbf{E}_\mathbf{j}\end{bmatrix} + \begin{bmatrix}\mathbf{W}_i & 0\\ 0 & \mathbf{W}_j\end{bmatrix}\begin{bmatrix}\mathbf{r}_\mathbf{i}\\ \mathbf{r}_\mathbf{j}\end{bmatrix}$$
$$+ \begin{bmatrix}\mathbf{X}_i & 0\\ 0 & \mathbf{X}_j\end{bmatrix}\begin{bmatrix}\mathbf{g}_\mathbf{i}\\ \mathbf{g}_\mathbf{j}\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_i & 0\\ 0 & \mathbf{Z}_j\end{bmatrix}\begin{bmatrix}\mathbf{ge}_\mathbf{i}\\ \mathbf{ge}_\mathbf{j}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_\mathbf{i}\\ \boldsymbol{\varepsilon}_\mathbf{j}\end{bmatrix},$$

where $\mathbf{y}_\mathbf{i}$ and $\mathbf{y}_\mathbf{j}$ are $n \times 1$ vectors of phenotypic observations of the traits i and j, respectively, on the n total experimental units; $\boldsymbol{\mu}_\mathbf{i}$ and $\boldsymbol{\mu}_\mathbf{j}$, are $n \times 1$ vectors of trait mean effects; $\mathbf{E}_\mathbf{i}$ and $\mathbf{E}_\mathbf{j}$ are vectors of macro-environmental effects for the two traits, corresponding to e environments; $\mathbf{r}_\mathbf{i}$ and $\mathbf{r}_\mathbf{j}$ are vectors of block effects corresponding to r replications in each of e environments; $\mathbf{g}_\mathbf{i}$ and $\mathbf{g}_\mathbf{j}$ are vectors of genotype or family effects, corresponding to g genotypes; $\mathbf{ge}_\mathbf{i}$ and $\mathbf{ge}_\mathbf{j}$ are vectors of genotype × environment interaction effects; $\boldsymbol{\varepsilon}_\mathbf{i}$ and $\boldsymbol{\varepsilon}_\mathbf{i}$ are vectors of n experimental error effects for traits i and j, respectively; and $\mathbf{T}_i$, $\mathbf{T}_j$, $\mathbf{W}_i$, $\mathbf{W}_j$, $\mathbf{X}_i$, $\mathbf{X}_j$, $\mathbf{Z}_i$, and $\mathbf{Z}_j$ are incidence matrices. If data are balanced $\mathbf{T}_i = \mathbf{T}_j$, $\mathbf{W}_i = \mathbf{W}_j$, $\mathbf{X}_i = \mathbf{X}_j$, and $\mathbf{Z}_i, = \mathbf{Z}_j$. Missing data on either trait may cause some differences between the incidence matrices of the two traits, however.

Ideally, all effects except the means should be considered random, with zero means, independent bivariate normal distributions, and variance-covariance matrices given by:

$$V\begin{bmatrix}\mathbf{E}_\mathbf{i}\\ \mathbf{E}_\mathbf{j}\end{bmatrix} = \begin{bmatrix}\mathbf{I}\sigma^2_{Ei} & \mathbf{I}\sigma_{Eij}\\ \mathbf{I}\sigma_{Eij} & \mathbf{I}\sigma^2_{Ej}\end{bmatrix}$$

$$V[\,\mathbf{r}_\mathbf{i} \quad \mathbf{r}_\mathbf{j}\,]^T = \begin{bmatrix}\mathbf{I}\sigma^2_{Ri} & \mathbf{I}\sigma_{Rij}\\ \mathbf{I}\sigma_{Rij} & \mathbf{I}\sigma^2_{Rj}\end{bmatrix}$$

$$V[\,\mathbf{g}_\mathbf{i} \quad \mathbf{g}_\mathbf{j}\,]^T = \begin{bmatrix}\mathbf{I}\sigma^2_{Gi} & \mathbf{I}\sigma_{Gij}\\ \mathbf{I}\sigma_{Gij} & \mathbf{I}\sigma^2_{Gj}\end{bmatrix}$$

$$V[\mathbf{ge_i} \quad \mathbf{ge_j}]^T = \begin{bmatrix} \mathbf{I}\sigma^2_{GEi} & \mathbf{I}\sigma_{GEij} \\ \mathbf{I}\sigma_{GEij} & \mathbf{I}\sigma^2_{GEj} \end{bmatrix}$$

and

$$V[\boldsymbol{\varepsilon}_\mathbf{i} \quad \boldsymbol{\varepsilon}_\mathbf{j}]^T = \begin{bmatrix} \mathbf{I}\sigma^2_{\varepsilon i} & \mathbf{I}\sigma_{\varepsilon ij} \\ \mathbf{I}\sigma_{\varepsilon ij} & \mathbf{I}\sigma^2_{\varepsilon j} \end{bmatrix}.$$

In practice, however, environment main effects and replication main effects may have to be treated as fixed for computational ease. This does not affect results for balanced data but could lead to different results for unbalanced data. The differences in estimation of the key variance and covariance components of interest (genotype, genotype × environment, and error), however, are expected to be small (Piepho and Mohring, 2005).

Typically, the data set that one would use to analyze data with univariate Proc MIXED or multivariate Proc GLM would have a number of rows equal to the total number of experimental units or observations, and different traits would be recorded in different columns of the data set. For example, if two traits were recorded, the data set might appear as in Table 1.

The structure of the data sets must be modified to implement multivariate REML analysis (Wright, 1998). A new classification variable is created to indicate the name of the trait, and a single response variable (dependent variable) indicates the phenotypic value of each trait on each experimental unit. For example, the data set in Table 1 could be modified by introducing a variable called "Trait" that indicates if the response variable is Trait 1 Or Trait 2, and the response variable is named "Y" (Table 2).

REML estimation of the model variance and covariance components using this model is implemented with the SAS code described in Appendix A. Using the genotypic variance and covariance component estimates, the genotypic correlation between traits i and j is estimated as:

$$\hat{r}_{gij} = \frac{\hat{\sigma}_{Gij}}{\hat{\sigma}_{Gi}\hat{\sigma}_{Gj}},$$

where $\hat{\sigma}_{Gij}$ is the estimated genotypic covariance between traits i and j and $\hat{\sigma}_{Gi}$ is the estimated genotypic standard deviation for trait i.

Approximate sampling variances and standard errors for the genotypic correlation estimates can be obtained with the delta method, on the basis of a Taylor series expansion of up to second-order terms of the estimating functions (Holland et al., 2003; Lynch and Walsh, 1998; Mode and Robinson, 1959). The sampling variance of the estimate of the genotypic correlation for traits i and j is estimated as the matrix product:

$$\hat{V}(\hat{r}_g) \approx \begin{bmatrix} \frac{\partial \hat{r}_G}{\partial \hat{\sigma}^2_{Gi}} \\ \frac{\partial \hat{r}_G}{\partial \hat{\sigma}_{Gij}} \\ \frac{\partial \hat{r}_G}{\partial \hat{\sigma}^2_{Gj}} \end{bmatrix}^T \begin{bmatrix} V(\hat{\sigma}^2_{Gi}) & C(\hat{\sigma}^2_{Gi},\hat{\sigma}_{Gij}) & C(\hat{\sigma}^2_{Gi},\hat{\sigma}^2_{Gj}) \\ C(\hat{\sigma}^2_{Gi},\hat{\sigma}_{Gij}) & V(\hat{\sigma}_{Gij}) & C(\hat{\sigma}_{Gij},\hat{\sigma}^2_{Gj}) \\ C(\hat{\sigma}^2_{Gi},\hat{\sigma}^2_{Gj}) & C(\hat{\sigma}_{Gij},\hat{\sigma}^2_{Gj}) & V(\hat{\sigma}^2_{Gj}) \end{bmatrix} \begin{bmatrix} \frac{\partial \hat{r}_G}{\partial \hat{\sigma}^2_{Gi}} \\ \frac{\partial \hat{r}_G}{\partial \hat{\sigma}_{Gij}} \\ \frac{\partial \hat{r}_G}{\partial \hat{\sigma}^2_{Gj}} \end{bmatrix}$$

$$= (\hat{r}_g)^2 \begin{bmatrix} \frac{-1}{2\hat{\sigma}^2_{Gi}} \\ \frac{1}{\hat{\sigma}_{Gij}} \\ \frac{-1}{2\hat{\sigma}^2_{Gj}} \end{bmatrix}^T \begin{bmatrix} V(\hat{\sigma}^2_{Gi}) & C(\hat{\sigma}^2_{Gi},\hat{\sigma}_{Gij}) & C(\hat{\sigma}^2_{Gi},\hat{\sigma}^2_{Gj}) \\ C(\hat{\sigma}^2_{Gi},\hat{\sigma}_{Gij}) & V(\hat{\sigma}_{Gij}) & C(\hat{\sigma}_{Gij},\hat{\sigma}^2_{Gj}) \\ C(\hat{\sigma}^2_{Gi},\hat{\sigma}^2_{Gj}) & C(\hat{\sigma}_{Gij},\hat{\sigma}^2_{Gj}) & V(\hat{\sigma}^2_{Gj}) \end{bmatrix} \begin{bmatrix} \frac{-1}{2\hat{\sigma}^2_{Gi}} \\ \frac{1}{\hat{\sigma}_{Gij}} \\ \frac{-1}{2\hat{\sigma}^2_{Gj}} \end{bmatrix}$$

Table 1. Structure of an example data set prepared for SAS Proc MIXED univariate analysis of two variables separately or for Proc GLM multivariate analysis of variance. Data on growing degree days (GDD) to flowering and plant height (PH) were collected on 132 oat recombinant inbred lines evaluated in two replications in each of 3 yr. Data only for the first four plots in the first environment are shown. Plot three was dropped from the data set, because it included a check cultivar. Full data set is available as supplementary material accompanying the online version of this paper or at www4.ncsu.edu/~jholland/correlation/correlation.html.

Environment	Replication	Plot	Genotype	GDD	PH
96	1	1	OT131	1339.8	86
96	1	2	OT37	1454.4	96
96	1	4	OT89	1454.4	98
etc…					

Table 2. Structure of an example data set from a replicated, multiple-environment trial with two traits measured modified for analysis by multivariate REML analysis using SAS Proc MIXED. Data on growing degree days (GDD) to flowering and plant height (PH) were collected on 132 oat recombinant inbred lines evaluated in two replications in each of 3 yr. Data are identical to those in Table 1 but structured differently.

Environment	Replication	Plot	Genotype	Trait	Y
96	1	1	OT131	GDD	1339.8
96	1	1	OT131	PH	86
96	1	2	OT37	GDD	1454.4
96	1	2	OT37	PH	96
96	1	4	OT89	GDD	1454.4
96	1	4	OT89	PH	98
etc…					

Analysis of Experimental Data

Oat cultivars Ogle and TAM O-301, 132 recombinant inbred lines developed from their cross, and eight check cultivars were included as entries in replicated field trials, as described by Holland et al. (2002). For the purposes of estimating variance and covariance components in this study, parental and check cultivars were deleted from the data set. The experiment was conducted at the Agronomy and Agricultural Engineering Research Farm near Ames, IA, in years 1996, 1997, and 1998. The experimental design was a randomized complete block with two replications in each year. Plots were hills seeded with 30 seeds per plot and spaced 0.3 m apart on a grid arrangement. Heading date (date after planting on which the first nodes on half of the plants in the plot had emerged completely above the flag leaf) was measured on each plot. Days to heading were converted to growing degree days (GDD) to heading using the formula developed by Wiggans (1956, cited in Sorrells and Simmons, 1992). Mean daily maximum temperature at the research farm was recorded and heat units for each day were computed as the number of degrees above 4.4°C. Plant height (PH) was measured from soil level to the tips of the panicles. Three data points (less than 0.4% of the total) for PH were missing (all in the 1997 environment, but representing three different genotypes) and no data for GDD to heading were missing.

Data were analyzed by multivariate REML implemented in Proc MIXED of SAS as described above, which considered environment and replication effects as fixed, only for the purpose of computational tractability. The effect of fitting environments and replications as fixed was investigated by also performing univariate REML analyses of the two variables separately with environments and replications modeled as either fixed or random. The multivariate REML results were also compared with MANOVA analysis, using Proc GLM, and obtaining coefficients of expected mean squares with the "random" option.

Simulation Study

The properties of multivariate REML and MANOVA method of moments estimates of genotypic correlations were compared over a wide range of parameter settings using simulated data. The initial parameter settings were based on the results of the analysis of the oat data set described above; later, parameter settings were varied to cover a wide range of genotypic correlation values, by changing the genotypic covariance values. Random samples were generated for environmental, replication, genotypic, genotype × environment interaction, and experimental error effects using the VNORMAL routine in SAS Proc IML. This function was used to draw samples from bivariate random normal distributions with variances and covariances initially on the basis of the variance and covariance components estimates from the oat data set, rounded to the nearest integer (Table 3). Variance and covariance components for genotypic, genotype × environment interaction, and experimental error were taken from the multivariate REML analysis. This analysis did not provide variance and covariance component estimates for environments and replications (because they were considered fixed effects); therefore, the variance components for these effects for the simulation studies were taken from the MANOVA analysis, and their covariance components were arbitrarily set to make the environmental correlation 0.75 and the replication correlation −0.75.

For each parameter setting, 1000 data sets were constructed, each sampling eight environments, two replications per environment, and 250 genotypes. Environments and genotypes were dropped at random to form reduced data sets with 75 or 150 genotypes and two or four environments. Data sets with 5, 15, or 25% missing data were also constructed by sampling from the balanced data sets. Missing data were distributed in three different ways. First, missing data points were distributed at random among the 2*N* observations from the *N* experimental units. That is, if heading date data were missing on a plot, this did not affect the probability that the height data were also missing on the same plot. This structure of missing data was referred to as "Missing Completely at Random" (MCAR, Little and Rubin, 1987). Second, whole experimental units were eliminated at random, such that if heading date data were missing on a plot, then height data were also always missing on the same plot. This structure of missing data was referred to as "Missing Plots at Random" (MPAR). Third, half of the desired missing data points were assigned to plots at random, as in the MPAR method, then the other half of missing data points were assigned to remaining observations at random, as in the MCAR method. This structure of missing data was referred to as "50% MCAR."

Table 3. Variance, covariance, and correlation parameter settings for simulation study.

Source of variation	σ^2_{GDD}	$\sigma_{GDD,HT}$	σ^2_{HT}	correlation
Initial setting (based on analysis of oat RIL data), r_g = 0.33, r_p = 0.20:				
Environment	**7605**	**327**	**25**	
Replication(environment)	**3**	**−1.3**	**1**	
Genotype	**11001**	**116**	**11**	**0.3335**
Genotype × environment interaction	**2319**	**47**	**13**	
Error	**1452**	**20**	**31**	
Phenotype	**14772**	**183**	**55**	**0.2030**
r_g = 0.00, r_p = 0.07 (H_0 Case I): settings same as initial settings, except:				
Genotype	**11001**	**0**	**11**	**0.0**
Phenotype	**14772**	**67**	**55**	**0.0743**
r_g = 0.00, r_p = 0.00 (H_0 Case II): settings same as initial settings, except:				
Genotype	**11001**	**0**	**11**	**0.0**
Genotype × environment interaction	**2319**	**0**	**13**	
Error	**1452**	**0**	**31**	
Phenotype	**14772**	**0**	**55**	**0.0**
r_g = −0.33, r_p = −0.05: settings same as initial settings, except:				
Genotype	**11001**	**−116**	**11**	**−0.3335**
Phenotype	**14772**	**−49**	**55**	**−0.0544**
r_g = 0.05, r_p = 0.09: settings same as initial settings, except:				
Genotype	**11001**	**17**	**11**	**0.0489**
Phenotype	**14772**	**84**	**55**	**0.0932**
r_g = 0.15, r_p = 0.13: settings same as initial settings, except:				
Genotype	**11001**	**52**	**11**	**0.1495**
Phenotype	**14772**	**119**	**55**	**0.1320**
r_g = 0.60, r_p = 0.31: settings same as initial settings, except:				
Genotype	**11001**	**209**	**11**	**0.6008**
Phenotype	**14772**	**276**	**55**	**0.3062**
r_g = 0.90, r_p = 0.42: settings same as initial settings, except:				
Genotype	**11001**	**313**	**11**	**0.8998**
Phenotype	**14772**	**380**	**55**	**0.4216**

Each simulation data set was analyzed in two ways: first, by multivariate REML (Proc MIXED), and, second, by MANOVA method of moments (Proc GLM) in SAS version 8.2. Genotypic and phenotypic correlations and their standard errors were estimated with the two methods. Correlation and standard error estimates for the MANOVA method were based on the method of moments (Mode and Robinson, 1959), using actual coefficients of expected mean squares (which varied among data sets) obtained with the "random" statement in GLM to estimate variance and covariance components. For each simulation, an approximate 95% confidence interval (CI) was estimated for the correlation estimates as the estimates plus or minus 1.96 times their estimated standard error (Lynch and Walsh, 1998). Correlation estimates were declared significantly different than zero if the approximate 95% CI did not include zero. If a genotypic variance component was estimated to be zero, the correlation coefficient and its standard error were considered to be zero.

Data sets with 2, 4, or 8 environments and 75, 150, or 250 genotypes were created for balanced data with true values of r_g and r_p set at 0.33 and 0.20 (similar to the real data set), 0.00 and 0.07 (no genetic covariance but genotype × environment and error covariances present, all other variances and covariances identical to initial settings), and 0.00 and 0.00 (no covariances between genotypic, genotype × environment, or error effects), respectively (Table 3). Data sets with these same parameter settings and sample sizes of two environments and 75 genotypes or four environments and 250 genotypes were created with 5, 15, or 25% missing data, and with missing data distributed as MCAR, MPAR, or 50% MCAR, and analyzed. To determine the effects of even smaller sample sizes, data sets with the original parameter settings and sample sizes of two environments and 50, 25, or 10 genotypes were created with 0, 5, 15, or 25% of data MCAR. Finally, to observe the effects of different true values of the genotypic correlation, data sets with sample sizes of two environments and 75 genotypes or four environments and 250 genotypes, 0, 5, 15, or 25% missing data, with missing data distributed as MCAR, MPAR, or 50% MCAR were created by changing only the genotypic covariance to set the genetic correlation parameter at −0.33, 0.05, 0.15, 0.60, or 0.90 (Table 3).

Type I error (false-positive) rates for the two methods were evaluated by analyzing the data sets in which the true correlation values were zero and determining the proportion of analyses in which the 95% CI for a correlation did not overlap zero. This was done in two ways for the genotypic correlation. First, the genotypic covariance was set to zero, but the GEI and error covariances were maintained at the original values, resulting in true values of r_g = 0.00 and r_p = 0.07

(Case I, Table 3). Case I was appropriate for estimating the Type I error rate for genotypic correlations but not for phenotypic correlations. Second, the genotypic, GEI, and error covariances were all set to zero, resulting in true values of zero for both genotypic and phenotypic correlations (Case II, Table 3).

In all, 193 combinations of parameter and missing data settings were tested, resulting in 193 000 simulated data sets and, since each data set was analyzed two ways, 386 000 analyses.

RESULTS

SAS codes

A general form for the SAS code is presented in Appendix A and codes specific for different experimental designs are available at as supplementary material accompanying the online version of this paper or at www4.ncsu.edu/~jholland/correlation/correlation.html. SAS codes for mating designs with two or more classification levels (e.g., North Carolina mating designs I, II) are presented in Appendices B and C.

Analysis of Experimental Data

Considering environments and replications as fixed or random effects in the multivariate REML analysis of the experimental oat data had little impact on the estimates of genotypes, genotype × environment interaction (GEI), and error variance components. The variance components estimates and their standard error estimates differed by less than 1% of their values when environments and replications were changed from random to fixed effects in Proc MIXED (Table 4). Similarly, the MANOVA method of moments estimates of variance and covariance components and their standard errors were very similar to the estimates from the multivariate REML estimates (Table 4). The heritability and genetic correlation estimates were also very similar between the different methods (Table 4). These results were expected because the data set was nearly balanced (less than 1% missing data). The heritability of GDD was high, from 0.74 on a plot-basis to 0.92 on an entry mean-basis, whereas the heritability of PH was low to moderate, from 0.20 on a plot-basis to 0.54 on an entry mean-basis. The estimated genotypic and phenotypic correlations were 0.33 and 0.20, respectively.

Analysis of Simulated Data Sets—Balanced Data

The relative effects of increasing genotypic and environmental sample sizes were first investigated in balanced simulation data sets. For each simulation data set, the approximate standard errors of correlation estimates were estimated using the delta method. Approximate 95% confidence intervals (CIs) for the correlation estimates were constructed as the estimates plus or minus 1.96 times their approximate standard errors. The accuracy of these CIs was tested by determining the proportion of analyses in which the true parameter value fell within the estimated CI. For most sample sizes, the CIs included the true correlation value in slightly less than 95% of samples, at worse including the true param-

Table 4. MANOVA, multivariate REML†, and univariate REML‡ estimates of variance and covariance components, heritabilities, and genotypic and phenotypic correlations (and their standard errors) of growing degree days (GDD) to flowering and plant height (PH) from real data set of 132 oat recombinant inbred lines grown in three environments.

	GDD			GDD, PH		PH		
	Variance component estimates, $\hat{\sigma}^2_{GDD}$			Covariance component estimates, $\hat{\sigma}_{GDD.PH}$		Variance component estimates, $\hat{\sigma}^2_{PH}$		
Source of variation	MANOVA	Multivariate REML	Univariate REML	MANOVA	Multivariate REML	MANOVA	Multivariate REML	Univariate REML
	GDD^2			GDD cm		cm^2		
Genotype	11076.76 (1486.42)	11001.00 (1487.31)	11001.00 (1487.34)	117.36 (45.38)	115.65 (45.41)	11.23 (2.67)	11.19(2 .67)	11.20 (2.68)
GEI	2329.25 (271.91)	2319.04 (271.47)	2319.38 (271.54)	47.37 (19.24)	46.76 (19.21)	12.83 (2.70)	12.74 (2.70)	12.77 (2.71)
Error	1454.69 (104.04)	1452.36 (103.74)	1452.51 (103.75)	20.38 (10.73)	20.88 (10.74)	30.66 (2.19)	30.72 (2.20)	30.70 (2.20)
Phenotype	14860.70	14772.61	14772.89	185.12	183.29	54.72	54.65	54.67
	Heritability estimates on a plot-basis			Genotypic correlation estimates		Heritability estimates on a plot-basis		
	0.75 (0.03)	0.74 (0.03)	0.74 (0.03)	0.33 (0.11)	0.33 (0.11)	0.21(0.04)	0.20(0.04)	0.20 (0.04)
	Heritability estimates on an entry mean-basis			Phenotypic correlation estimate		Heritability estimates on an entry mean-basis		
	0.92 (0.01)	0.92 (0.01)	0.92 (0.01)	0.21 (0.05)	0.20 (0.05)	0.54 (0.07)	0.54 (0.07)	0.54 (0.07)

† Multivariate REML model considered environments and replications to be fixed effects and genotypes, genotype × environment interactions, and error effects to be random.

‡ Univariate REML model considered all effects (environments, replications genotypes, genotype × environment interactions, and errors) to be random.

Table 5. Power to detect a significant genotypic correlation, proportion of samples in which the true genotypic correlation value was within the estimated 95% confidence interval (95% CI coverage), and Type I error rates for REML and MANOVA methods of estimation of the genotypic correlation and different sample sizes of environments (N_e) and genotypes (N_g), based on 1000 simulated data sets with no missing data. Power and coverage were based on simulations with parameter settings $r_g = 0.33$ and $r_p = 0.20$. Type I error rates were based on two simulations (Cases I and II) where the true value of the genotypic correlation was zero and the null hypothesis H_0: $r_g = 0$ was tested. Parameter settings were $r_g = 0$ and $r_p = 0.07$ for Case I and $r_g = 0$ and $r_p = 0$ for Case II.

		Power		95% CI coverage†		Case I Type I error rate†		Case II Type I error rate†	
N_e	N_g	REML	MANOVA	REML	MANOVA	REML	MANOVA	REML	MANOVA
2	75	0.441	0.449	0.945	0.943	0.048	0.053	0.048	0.053
2	150	0.698	0.702	0.945	0.945	0.048	0.048	0.052	0.052
2	250	0.880	0.882	0.956	0.953	0.038	0.040	0.061	0.062
4	75	0.654	0.661	0.931	0.929	0.061	0.066	0.081	0.082
4	150	0.887	0.887	0.938	0.936	0.052	0.053	0.065	0.065
4	250	0.980	0.981	0.943	0.939	0.051	0.052	0.051	0.052
8	75	0.752	0.762	0.935	0.932	0.058	0.063	0.074	0.080
8	150	0.956	0.958	0.946	0.944	0.043	0.044	0.055	0.057
8	250	0.998	0.998	0.945	0.944	0.052	0.054	0.039	0.040

† Standard error for 95% CI coverage and Type I error rate estimates ≈ $\sqrt{\frac{0.95 \times 0.05}{1000}} = 0.007$.

eter value 93% of the time (Table 5). There were no significant differences between REML and MANOVA methods for 95% CI coverages (differences of less than 1%, Table 5). Power of the significance test of the genotypic correlation ranged from less than 50% in the smaller sample size to almost 100% in the largest sample size (Table 5). Increasing sample size had a large effect on power of the test, whereas the analysis method had little or no effect.

Two cases were studied to investigate Type I error (false-positive) rates for the two methods. Case I had true values $r_g = 0.00$ and $r_p = 0.07$ and Case II had true values $r_g = 0.00$ and $r_p = 0.00$ (Table 3). The Type I error rates for both cases and both analyses tended to be slightly greater than the expected 0.05, with a maximum of 0.082. With the largest sample size, however, the Type I error rates ranged from 0.039 to 0.054 (Table 5). The Type I error rates for the REML method were always equal to or smaller than those for the MANOVA method. Except for the largest sample size, the Type I error rate of the genotypic correlation significance test was higher in Case II than Case I, by as much as 0.023 (Table 5). Therefore, hereafter, only Case II Type I error rates are reported.

Analysis of simulated data sets– missing data

For the case of 75 genotypes and two environments, the distribution of REML and MANOVA estimates appeared identical when data were balanced, but with 25% of data missing completely at random (MCAR), the spread of the distribution increased, and more so for MANOVA than for REML estimates (Fig. 1). Furthermore, there was a spike in the frequency of estimates of value zero from the MANOVA method when data were missing, due to obtaining negative estimates of genetic

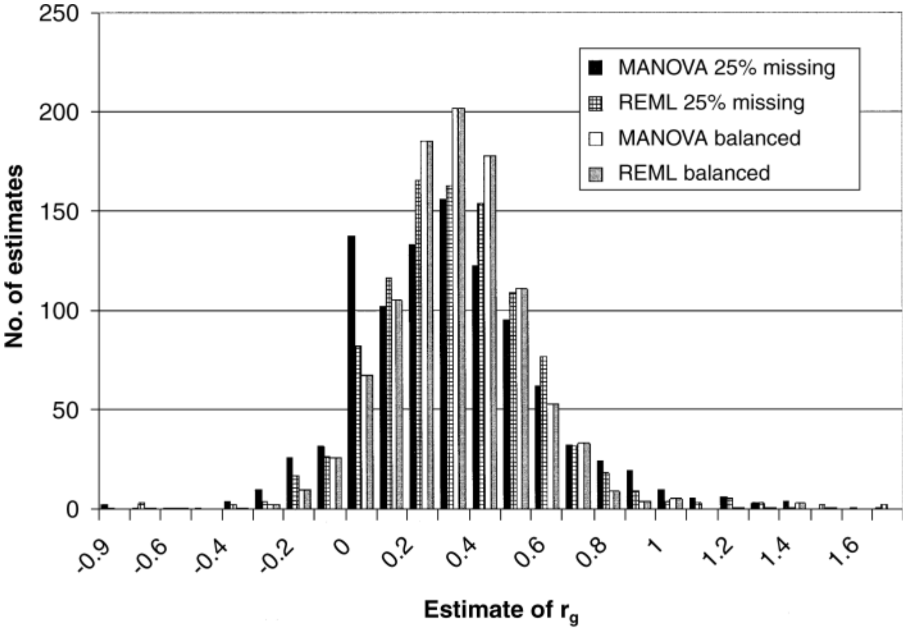

Fig. 1. Distribution of genotypic correlation estimates from 1000 simulations from data sets with two environments and 75 genotypes, with either balanced data or 25% data missing completely at random (MCAR), and analyzed either with MANOVA or REML.

variances of one of the variance components (Fig. 1). Genotypic correlation estimates greater than one (outside of the parameter space) were obtained with both methods, but were most frequent with MANOVA and 25% missing data (Fig. 1).

Power of the significance test of $\hat{r}_g$ was similar between methods, except in the samples of 75 genotypes and two environments with larger amounts of missing data distributed as 50% or 100% MCAR (Table 6), where REML was clearly superior. With this sample size, REML 95% CIs had better coverage of the true value of the correlation coefficient than MANOVA in all cases (Table 6). The poorest performance of MANOVA occurred with the 75 genotype and two environment sample size with 25% missing data distributed as 100% MCAR. In this case, the MANOVA 95% CIs included the true value only 90.7% of the time (compared with 94.1% for REML) and the power of the test for a significant genetic correlation was 24.9% (compared with 31.6% for REML). MANOVA estimation improved in the sample of four environments and 250 genotypes, although there were some cases where Type I error rates of MANOVA estimates were significantly greater than 0.05 (Table 6).

The results with samples of 75 genotypes and two environments indicated that approximate 95% CIs of REML, but not MANOVA, estimators were close to the stated coverage even with 25% missing data. To determine what minimum sample size of genotypes was required for the 95% CIs of the REML estimators to remain close to the stated coverage, simulations were conducted with samples of two environments and 50, 25, and 10 genotypes with the true value of r_g set at 0.33 and 0, 5, 15, or 25% of data MCAR (results not shown). Even with balanced data, samples of 50 genotypes were not adequate to obtain accurate coverage by the estimated CIs: the REML CIs had 92.4% coverage, and the MANOVA CIs had 91.7% coverage. Coverage was worse with more missing data (REML CIs had 89.3% and MANOVA CIs had 84.4% coverage with 25% missing data) or with smaller sample sizes (CIs estimated from both methods had coverage of only about 80% with samples of 25 genotypes and no missing data). An important factor causing poor coverage with the smaller sample sizes, even with balanced data, was the increased chance of obtaining a zero or negative variance component estimate, resulting in a zero estimate of the genotypic correlation. This caused serious deviations from normality of the empirical distribution of the correlation estimates (with a frequency spike at zero) and, consequently, a poor performance by the delta method estimators, which assume normality of the estimator's distribution.

The effect of missing data was checked for data sets of two environments and 75 genotypes and of four environments and 250 genotypes while varying true parameter values for the genotypic covariance and correlation. For the situation where the parameter settings were identical to the original settings, but the genotypic correlation was −0.33, the results were similar to the case $r_g = +0.33$ (Table 7). Thus, the behavior of positive and negative genetic correlation estimates were not identical but similar and with consistent trends with missing data.

Across all parameter settings, MANOVA and REML methods performed similarly when data were balanced, but, with one exception, REML always had better power and 95% CI coverage when 25% of data were missing completely at random (Table 6). The only ex-

Table 6. Power to detect a significant genotypic correlation, proportion of samples in which the true genotypic correlation value was within the estimated 95% confidence interval (95% CI coverage), and Type I error rates for REML and MANOVA methods of estimation of the genotypic correlation and different sample sizes of environments (N_e) and genotypes (N_g), based on 1000 simulated data sets with 0, 5, 15, or 25% missing data distributed as missing plots at random (MPAR), missing completely at random (MCAR), or as 50% MCAR and 50 MPAR (50/50). Power and coverage were based on simulations with parameter settings $r_g = 0.33$ and $r_p = 0.20$. Type I error rates were based on parameter settings of $r_g = 0$ and $r_p = 0$ (equivalent to Case II in Table 5).

		Power		95% CI coverage†		Type I error rate†	
% missing data	Distribution of missing data	REML	MANOVA	REML	MANOVA	REML	MANOVA
Sample size: 2 environments, 75 genotypes							
0	NA‡	0.441	0.449	0.945	0.943	0.048	0.053
5	MPAR	0.420	0.437	0.940	0.938	0.056	0.053
5	50/50	0.422	0.423	0.949	0.945	0.054	0.054
5	MCAR	0.415	0.418	0.947	0.935	0.052	0.055
15	MPAR	0.393	0.407	0.946	0.942	0.055	0.049
15	50/50	0.379	0.360	0.938	0.927	0.054	0.058
15	MCAR	0.374	0.333	0.944	0.938	0.044	0.041
25	MPAR	0.352	0.358	0.939	0.934	0.042	0.039
25	50/50	0.340	0.301	0.926	0.918	0.045	0.057
25	MCAR	0.316	0.249	0.941	0.907	0.038	0.039
Sample size: 4 environments, 250 genotypes							
0	NA	0.980	0.981	0.943	0.939	0.051	0.052
5	MPAR	0.977	0.977	0.943	0.941	0.055	0.053
5	50/50	0.976	0.975	0.941	0.941	0.054	0.060
5	MCAR	0.977	0.973	0.935	0.937	0.060	0.066
15	MPAR	0.974	0.973	0.943	0.941	0.054	0.062
15	50/50	0.973	0.971	0.947	0.939	0.050	0.053
15	MCAR	0.970	0.963	0.942	0.944	0.055	0.059
25	MPAR	0.964	0.965	0.934	0.937	0.058	0.068
25	50/50	0.966	0.960	0.945	0.937	0.062	0.067
25	MCAR	0.966	0.940	0.942	0.942	0.051	0.062

† Standard error for 95% CI coverage and Type I error rate estimates $\approx \sqrt{\frac{0.95 \times 0.05}{1000}} = 0.007$.

‡ Not applicable.

Table 7. Power to detect a significant genotypic correlation and proportion of samples in which the true genotypic correlation value was within the estimated 95% confidence interval (95% CI coverage) for REML and MANOVA methods of estimation of the genotypic correlation, based on 1000 simulated data sets with parameter settings varying from $r_g = -0.33$ to $r_g = 0.90$, different sample sizes of environments (N_e) and genotypes (N_g), and either balanced data or 25% data missing completely at random (MCAR).

		Genotypic correlation			
		Power		95% CI coverage†	
True value of r_g	% of data missing	REML	MANOVA	REML	MANOVA
Sample size: 2 environments, 75 genotypes					
−0.33	0	0.407	0.414	0.952	0.950
−0.33	25	0.307	0.201	0.944	0.901
0.05	0	0.044	0.045	0.950	0.948
0.05	25	0.032	0.029	0.939	0.897
0.15	0	0.137	0.138	0.949	0.948
0.15	25	0.109	0.085	0.932	0.905
0.33	0	0.441	0.449	0.945	0.943
0.33	25	0.316	0.249	0.941	0.907
0.60	0	0.879	0.881	0.956	0.954
0.60	25	0.765	0.605	0.951	0.910
0.90	0	0.963	0.965	0.971	0.969
0.90	25	0.896	0.772	0.952	0.902
Sample size: 4 environments, 250 genotypes					
−0.33	0	0.988	0.988	0.935	0.934
−0.33	25	0.976	0.950	0.931	0.944
0.05	0	0.076	0.077	0.950	0.948
0.05	25	0.069	0.080	0.941	0.944
0.15	0	0.426	0.431	0.936	0.934
0.15	25	0.380	0.363	0.936	0.927
0.33	0	0.980	0.981	0.943	0.939
0.33	25	0.966	0.940	0.942	0.942
0.60	0	1.000	1.000	0.940	0.940
0.60	25	1.000	1.000	0.937	0.941
0.90	0	1.000	1.000	0.942	0.942
0.90	25	1.000	1.000	0.951	0.960

† Standard error for 95% CI coverage estimates ≈ $\sqrt{\frac{0.95 \times 0.05}{1000}} = 0.007$.

ception was when the true value of the genetic correlation was 0.05 and the larger sample size was considered.

DISCUSSION

The results of the simulation study suggest that MANOVA and REML estimation of genotypic and phenotypic correlations are approximately equal if data are balanced (as expected) and do not differ dramatically unless the amount of missing data is more than 5%. The greatest differences between the two methods occurred when missing data were distributed randomly among observations rather than as pairs of missing values from common experimental units. This occurred because the MANOVA method eliminates all experimental units that lack values for both phenotypes from the analysis, whereas REML can use the information from the experimental units that have data for only one trait or the other to estimate the variance components for those traits. As the amount of missing data exceeds 5%, particularly with smaller samples sizes and data missing completely at random (MCAR), the REML method can be recommended over the MANOVA method because it had greater power of detection and more accurate 95% CIs for both genotypic and phenotypic correlations. However, sample sizes of 50 genotypes were not sufficient to obtain accurate CIs with either method. Therefore, a minimum sample size of 75 genotypes and two environments is recommended to obtain accurate 95% CIs using the methods proposed in this paper.

Power to detect genotypic correlations tended to be low unless genotypic sample sizes were 150 or more and the true parameter value was greater than 0.15. Power of detection was always greater for phenotypic than genotypic correlations, and power greater than 70% was observed for phenotypic correlations of value 0.09 if a sample of 250 genotypes tested in four environments was used (results not shown). Detection of a significant phenotypic correlation, however, does not imply that there is also a nonzero genotypic correlation (e.g., the case of $r_g = 0.00$ and $r_p = 0.07$ in this simulation study). Estimates of phenotypic correlations can be useful in determining the relationship between phenotypic values of different traits, but they do not reflect expected correlated changes that may occur because of selection on one of the traits.

A drawback to the use of multivariate REML is that it is often more computationally and memory intensive than MANOVA. With complex models and multiple traits to be analyzed, the number of model parameters may be too large for current typical personal computers to handle. For this reason, the sample SAS code provided (see supplementary material accompanying the online version of this paper or ww4.ncsu.edu/~jholland/correlation/correlation.html) treats environment and replication as fixed factors. This should not be of concern if users are not interested in estimating the variance components or predicting the effects of these factors (as is the case when estimating genotypic and phenotypic correlations, see also Piepho and Mohring, 2005) and switching environments and replications from random to fixed in the univariate analyses of the real data set had negligible impact on the results (Table 4). Whereas the Proc MIXED multivariate analysis of the oat data set did not converge correctly after many hours of execution time when all components were treated as random, the Proc MIXED REML analysis was actually substantially faster than the GLM MANOVA analysis for the largest simulated data sets when environments and replications were considered fixed (e.g., 3.5 min for REML vs. 25 min on average for GLM analysis of the eight environment, 250 genotype data sets on a Pentium III computer).

Convergence of the multivariate REML model can also be hindered if the two variables have greatly different scales. In such cases, memory demands of the program also can be reduced by centering the data from both variables (SAS Institute Inc., 1999). To reduce the number of iterations required to reach a solution and to improve chances for correct convergence, users can first use MANOVA to obtain initial estimates of the covariance components, then supply these as starting parameter estimates to SAS Proc MIXED with the PARMS statement (SAS Institute Inc., 1999). Finally, more than two traits can be analyzed at one time (Zamudio and Wolfinger, 2002), but additional traits added to the model will quickly increase computing demands. The example data sets presented in this paper and SAS codes for analyzing various experimental designs, including single-and multiple-environment trials, and randomized complete block and incomplete block designs, are available as supplementary material accompanying the online version of this paper or at www4.ncsu.edu/~jholland/correlation/correlation.html.

Alternative estimates of precision of the correlation estimates should be possible with resampling techniques such as bootstrapping (Efron, 1982) or with Bayesian techniques (Shoemaker et al., 1999). Liu et al. (1997) investigated the distribution of REML estimates of genotypic correlations and found that parametric bootstrapping produced estimates of sample variances close to their known values. They did not include approximate parametric estimates of the sampling variance, such as delta method estimates, in their investigation, however, so it is not known how they compare to bootstrap estimates of variances of REML-based correlation estimates. Furthermore, they studied a relatively simple completely randomized design in one environment, in which missing data do not cause unbalance in the design. Similarly, nonparametric bootstrapping (Xie and Mosjidis, 1999) and jackknifing (Roff and Preziosi, 1994) can provide accurate estimates of genetic correlations and their sampling distributions, but the utility of these methods has only been demonstrated with single-environment data. Appropriate resampling schemes for more complicated data structures, such as multiple-environment trials, remain largely uninvestigated. Until appropriate resampling or Bayesian techniques are designed and proven useful, the approximate standard errors presented in this study should be adequate for most applications where large numbers of genotypes are evaluated.

ACKNOWLEDGMENTS

I thank Drs. Javier Betrán, Cavell Brownie, Cuauhtemoc Cervantes-Martinez, Lauren McIntyre, John McLaughlin, and Russell Wolfinger and several anonymous reviewers for helpful comments on previous versions of this paper. I thank Sandy Donaghy for helpful tips on SAS code for data simulation. I thank Dr. Steve Szalma for use of his computers to execute the simulation study.

APPENDIX A

SAS code for converting a data set with different traits in separate columns to a “longitudinal” data set and for conducting multivariate REML analysis on the longitudinal data set on the basis of a one-way classification of genotypes. This example uses four traits called trait1, trait2, trait3, and trait4.

```
  * Create the data set in standard format;
data one;
input env rep plot geno trait1 trait2 trait3 trait4;
  * Use a “cards” statement to read in data. This is not shown in this example;
  * Create the “tall” data set from standard data set format;
data tall; set one;
trait = “trait1”;
y = trait1;
output;
trait = trait2;
y = “trait2”;
output;
trait = trait3;
y = “trait3”;
output;
trait = trait4;
y = “trait4”;
output;
drop trait1 trait2 trait3 trait4;
run;
  * Create a macro to perform multivariate reml analysis;
%macro correlation(TraitI, TraitJ);
  * Select the two traits to be analyzed from the tall dataset;
data subset; set tall; if trait = “&TraitI” or if trait = “&TraitJ”;
  * Perform multivariate REML estimation of variance and covariance components, using the “asycov” option of proc mixed to obtain the asymptotic variance-covariance matrix of the estimates;
proc mixed data = subset asycov;
class trait env rep geno;
  * Treat environments and replications as fixed effects to speed computation of the variance and covariance estimates of interest. Note that the F-tests associated with these factors are testing the hypothesis that there are no significant differences among environments or among replications for the two traits combined. Such hypothesis are generally not of real interest, instead tests of main effects of environments and replications, if they are of interest, should be conducted on each trait separately with univariate analyses;
model y = env(trait) rep(env*trait);
  * Treat genotypes (“geno”) and genotype × environment interactions (“geno*env”) as random effects and estimate their variance and covariance components for the two traits with the following codes;
random trait/subject = geno type = un;
random trait/subject = geno*env type = un;
  * Model the residual error term (“rep*geno(env)”) to allow for covariances between error effects on the two traits measured on the same plot, but not between different plots;
repeated trait/sub = rep*geno(env) type = un;
  * Output the estimates of variance and covariance components to a data set called “estmat” and output the asymptotic variance-covariance matrix of those estimates to a data set called “covmat”;
ods output covparms = estmat asycov = covmat;
run;
  * Read variance and covariance estimates (“estmat” data set, to be read into a vector called “e”) and their variance-covariance matrix (“covmat” data set, to be read into a matrix called “cov”) into proc iml to estimate correlations and their standard errors using the delta method;
proc iml;
use estmat; read all into e;
use covmat; read all into cov;
  * Obtain the “C” matrix by removing the extra first column of the “cov” matrix;
C = cov(|1:nrow(cov), 2:ncol(cov)|);
```

```
  *Obtain genotypic covariance (CovG) and variance components (VG1 and VG2) from the elements of the "e" vector;

CovG = e(|2,1|); VG1 = e(|1,1|); VG2 = e(|3,1|);

  *Obtain phenotypic covariance (CovP) and variance components (VP1 and VP2) from the elements of the "e" vector;

CovP = CovG + e(|5,1|) + e(|8,1|);
VP1 = VG1 + e(|4,1|) + e(|7,1|); VP2 = VG2 + e(|6,1|) +
  e(|9,1|);

  *Create a module called "correl" that will estimate genotypic and phenotypic correlations and their standard errors;

start correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG, RP,
  SERG, SERP);
RG = CovG/sqrt(VG1*VG2);

  *Make the derivative vector for rg, note that the order of the rows and columns of the variance covariance matrix is VG1, CovG, VG2, VGE1, CovGE, VGE2, VError1, CovError, VError2;

dg = (−1/(2*VG1))//(1/CovG)//(−1/(2*VG2))//0//0//0//0//0//0;

  *Compute the variance of the estimate of the genotypic correlation using the delta method, then take its square root to obtain the standard error of the genotypic correlation estimate ("serg");

varrg = (RG**2)*dg′*C*dg; serg = sqrt(varrg);
RP = CovP/sqrt(VP1*VP2);

  *Make the derivate vector for rp;

d1p = −1/(2*VP1); d2p = 1/CovP; d3p = −1/(2*VP2);
dp = d1p//d2p//d3p//d1p//d2p//d3p//d1p//d2p//d3p;

  *Compute the variance of the estimate of the phenotypic correlation using the delta method, then take its square root to obtain the standard error of the phenotypic correlation estimate ("serp");

varrp = (RP**2)*dp′*C*dp; serp = sqrt(varrp);
finish correl;

  *Run the "correl" module and display the results;

call correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG, RP,
  SERG, SERP);
print "Genotypic Correlation Between &TraitI and &TraitJ";
print RG serg;
print "Phenotypic Correlation Between &TraitI and &TraitJ ";
print RP serp;
quit;
run;

  *End the macro;

%mend correl;

  *Invoke the correlation macro for each pair of traits. In this example, there are four traits and six pairs of traits to be analyzed;

%correlation(Trait1, Trait2);
%correlation(Trait1, Trait3);
%correlation(Trait1, Trait4);
%correlation(Trait2, Trait3);
%correlation(Trait2, Trait4);
%correlation(Trait3, Trait4);
run;
```

APPENDIX B

SAS code for conducting multivariate REML analysis based on a nested mating design (design I), to obtain estimates of the additive genetic and phenotypic correlations between four traits, called trait1, trait2, trait3, and trait4. The experimental design is a sets within replications repeated over environments, following Hallauer and Miranda (1988, p. 79) without information on within-plot variation. Use of initial parameters with the PARMS statement in Proc MIXED (on the basis of a previous MANOVA analysis) is highly recommended to aid convergence of this complex model.

```
  *Create "tall" dataset, following the example in appendix A;
  *Create a macro to perform multivariate reml analysis;

%macro design1(TraitI, TraitJ);

  *Select the two traits to be analyzed from the tall dataset;

data subset; set tall; if trait = "&TraitI" or trait = "&TraitJ";

  *Perform multivariate REML estimation of variance and covariance components, using the "asycov" option of proc mixed to obtain the asymptotic variance-covariance matrix of the estimates;

proc mixed data = subset asycov;
class trait env rep set male female;

  *Treat environments, replications, and sets as fixed effects to speed computation of the variance and covariance estimates of interest;
model y = env(trait) rep(env*trait) set(rep*env*trait);

  *Treat male within set, female within male and set, male × environment interaction within set, and female × environment interaction within male and set as random effects and estimate their variance and covariance components for the two traits with the following codes;

random trait/subject = male(set) type = un;
random trait/subject = female(male*set) type = un;
random trait/subject = male*env(set) type = un;
random trait/subject = female*env(male*set) type = un;

  * Model the residual error term to allow covariances between error effects on the two traits measured on the same plot or on the same male group, but not between different plots or different male groups. Note that the residual error term is a compound term of the plot-to-plot error variance ("rep*female(male*set*env)") and the interaction of male groups with replications ("rep*male(set*env)"). The (co)variances of these two terms must be added later to obtain the total error co(variance);

random trait/subject = rep*male(set*env) type = un;
repeated trait/sub = rep*female(male*set*env) type = un;

  * Output the estimates of variance and covariance components to a data set called "estmat" and output the asymptotic variance-covariance matrix of those estimates to a data set called "covmat";

ods output covparms = estmat asycov = covmat;
run;

  * Read variance and covariance estimates ("estmat" data set, to be read into a vector called "e") and their variance-covariance matrix ("covmat" data set, to be read into a matrix called "cov") into proc iml to estimate correlations and their standard errors using the delta method;

proc iml;
```

```
use estmat; read all into e;
use covmat; read all into cov;

   *Obtain the "C" matrix by removing the extra first column of the "cov" matrix;

C = cov(|1:nrow(cov), 2:ncol(cov)|);

   *Obtain male covariance (CovM = 1/4 CovA) and variance components (VM1 and VM2 = 1/4 additive variances) from the elements of the "e" vector;

CovM = e(|2,1|); VM1 = e(|1,1|); VM2 = e(|3,1|);

   *Obtain phenotypic covariance (CovP) and variance components (VP1 and VP2) from the elements of the "e" vector;

CovP = CovM + e(|5,1|) + e(|8,1|) + e(|11,1|) + e(|14,1|) +
   e(|17,1|);
VP1 = VM1 + e(|4,1|) + e(|7,1|) + e(|10,1|) + e(|13,1|) +
   e(|16,1|);
VP2 = VM2 + e(|6,1|) + e(|9,1|) + e(|12,1|) + e(|15,1|) +
   e(|18,1|);

   *Create a module called "correl" that will estimate genotypic and phenotypic correlations and their standard errors;

start correl(C, CovM, VM1, VM2, CovP, VP1, VP2, RG, RP,
   SERG, SERP); RG = CovM/sqrt(VM1*VM2);

   *Make the derivative vector for rg, note that the order of the rows and columns of the variance covariance matrix is VM1, CovM, VM2, VF(M)1, CovF(M), VF(M)2, VME1, CovME, VME2, VF(M)E1, CovF(M)E, VF(M)2E, VRME1, CovRME, VRME2, VRFME1, CovRFME, VRFME2;

dg = (−1/(2*VM1))//(1/CovM)//(−1/(2*VM2))//0//0//0//0//0//
   0//0//0//0//0//0//0//0//0//0;

   *Compute the variance of the estimate of the genotypic correlation using the delta method, then take its square root to obtain the standard error of the genotypic correlation estimate ("serg");

varrg = (RG**2)*dg′*C*dg; serg = sqrt(varrg);
RP = CovP/sqrt(VP1*VP2);

   *Make the derivate vector for rp;

d1p = −1/(2*VP1); d2p = 1/CovP; d3p = −1/(2*VP2);
dp = d1p//d2p//d3p//d1p//d2p//d3p//d1p//d2p//d3p//d1p//d2p//
   d3p//d1p//d2p//d3p//d1p//d2p//d3p;

   *Compute the variance of the estimate of the phenotypic correlation using the delta method, then take its square root to obtain the standard error of the phenotypic correlation estimate ("serp");

varrp = (RP**2)*dp′*C*dp; serp = sqrt(varrp);
finish correl;

   *Run the "correl" module and display the results;

call correl(C, CovM, VM1, VM2, CovP, VP1, VP2, RG, RP,
   SERG, SERP);
print "Additive Genetic Correlation Between &TraitI and
   &TraitJ";
print RG serg;
print "Phenotypic Correlation Between &TraitI and &TraitJ ";
print RP serp;
quit;
run;

   *End the macro;

%mend;

   *Invoke the correlation macro for each pair of traits. In this example, there are four traits and six pairs of traits to be analyzed;

%design1(Trait1, Trait2);
%design1(Trait1, Trait3);
%design1(Trait1, Trait4);
%design1(Trait2, Trait3);
%design1(Trait2, Trait4);
%design1(Trait3, Trait4);
run;
```

APPENDIX C

SAS code for conducting multivariate REML analysis based on a cross-classified mating design (design II), to obtain estimates of the additive genetic and phenotypic correlations between four traits, called trait1, trait2, trait3, and trait4. The experimental design is a replications within sets design repeated over environments, following Hallauer and Miranda (1988, p. 70). The additive genetic correlation is estimated from the male covariance and covariance components; other estimators are also possible. Use of initial parameters with the PARMS statement in Proc MIXED (on the basis of a previous MANOVA analysis) is highly recommended to aid convergence of this complex model.

```
   *Create a macro to perform multivariate reml analysis;

%macro design2(TraitI, TraitJ);

   *Select the two traits to be analyzed from the tall dataset;
data subset; set tall; if trait = "&TraitI" or trait = "&TraitJ";

   *Perform multivariate REML estimation of variance and covariance components, using the "asycov" option of proc mixed to obtain the asymptotic variance-covariance matrix of the estimates;

proc mixed data = subset asycov;
class trait env rep set male female;

   *Treat environments, replications, and sets as fixed effects to speed computation of the variance and covariance estimates of interest;

model y = env(trait) set(trait) set*env(trait) rep(set*env*trait);

   *Treat male within set, female within set, female × male within set and their respective interactions with environment as random effects and estimate their variance and covariance components for the two traits with the following codes;

random trait/subject = male(set) type = un;
random trait/subject = female(set) type = un;
random trait/subject = male*female(set) type = un;
random trait/subject = male*env(set) type = un;
random trait/subject = female*env(set) type = un;
random trait/subject = male*female*env(set) type = un;

   *Model the residual error term to covariances between error effects on the two traits measured on the same plot or on the same male group, but not between different plots or different male groups. Note that the residual error term is a compound term of the plot-to-plot error variance
   ("rep*female*male(set*env)"), the interaction of male groups with replications ("rep*male(set*env)"), and the interaction of female groups with replications ("rep*female(set*env)"). The following code combines rep*female*male(set*env) and rep*female(set*env) into a compound term, rep*female(male*set*env). The (co)variances of this term and rep*male(set*env) must be added later to obtain the total error co(variance);

random trait/subject = rep*male(set*env) type = un;
repeated trait/sub = rep*female(male*set*env) type = un;

   * Output the estimates of variance and covariance components to a data set called "estmat" and output the asymptotic variance-covariance matrix of those estimates to a data set called "covmat";

ods output covparms = estmat asycov = covmat;
run;

   * Read variance and covariance estimates ("estmat" data set, to be read into a vector called "e") and their variance-covariance matrix ("covmat" data set, to be read into a matrix called "cov") into proc iml to estimate correlations and their standard errors using the delta method;

proc iml;
use estmat; read all into e;
use covmat; read all into cov;

   * Obtain the "C" matrix by removing the extra first column of the "cov" matrix;

C = cov(|1:nrow(cov), 2:ncol(cov)|);

   * Obtain male covariance (CovM = 1/4 CovA) and variance components (VM1 and VM2 = 1/4 additive variances) from the elements of the "e" vector;

CovM = e(|2,1|); VM1 = e(|1,1|); VM2 = e(|3,1|);
   * Obtain phenotypic covariance (CovP) and variance components (VP1 and VP2) from the elements of the "e" vector;

CovP = CovM + e(|5,1|) + e(|8,1|) + e(|11,1|) + e(|14,1|) +
   e(|17,1|) + e(|20,1|) + e(|23,1|) + e(|26,1|);
VP1 = VM1 + e(|4,1|) + e(|7,1|) + e(|10,1|) + e(|13,1|) +
   e(|16,1|) + e(|19,1|) + e(|22,1|) + e(|25,1|);
VP2 = VM2 + e(|6,1|) + e(|9,1|) + e(|12,1|) + e(|15,1|) +
   e(|18,1|) + e(|21,1|) + e(|24,1|) + e(|27,1|);

   * Create a module called "correl" that will estimate genotypic and phenotypic correlations and their standard errors;

start correl(C, CovM, VM1, VM2, CovP, VP1, VP2, RG, RP,
   SERG, SERP);
RG = CovM/sqrt(VM1*VM2);

   * Make the derivative vector for rg, note that the order of the rows and columns of the variance covariance matrix is VM1, CovM, VM2, VF(M)1, CovF(M), VF(M)2, VME1, CovME, VME2, VF(M)E1, CovF(M)E, VF(M)2E, VRME1, CovRME, VRME2, VRFME1, CovRFME, VRFME2;

dg = (−1/(2*VM1))//(1/CovM)//
(−1/(2*VM2))//0//0//0//0//0//0//0//0//0//0//0//0//0//0//0//0//0//0//
   0//0//0//0//0//0;

   * Compute the variance of the estimate of the genotypic correlation using the delta method, then take its square root to obtain the standard error of the genotypic correlation estimate ("serg");

varrg = (RG**2)*dg′*C*dg; serg = sqrt(varrg);
RP = CovP/sqrt(VP1*VP2);
* Make the derivate vector for rp;
d1p = −1/(2*VP1); d2p = 1/CovP; d3p = −1/(2*VP2);
dp = d1p//d2p//d3p//d1p//d2p//d3p//d1p//d2p//d3p//d1p//d2p//
   d3p//d1p//d2p//d3p//d1p//d2p//d3p//
d1p//d2p//d3p//d1p//d2p//d3p//d1p//d2p//d3p;

   * Compute the variance of the estimate of the phenotypic correlation using the delta method, then take its square root to obtain the standard error of the phenotypic correlation estimate ("serp");

varrp = (RP**2)*dp‘*C*dp; serp = sqrt(varrp);
finish correl;

   * Run the "correl" module and display the results;

call correl(C, CovM, VM1, VM2, CovP, VP1, VP2, RG, RP,
   SERG, SERP);
print "Additive Genetic Correlation Between &TraitI and
   &TraitJ";
print RG serg;
print "Phenotypic Correlation Between &TraitI and &TraitJ ";
print RP serp;
quit;
run;

   * End the macro;

%mend;

   * Invoke the correlation macro for each pair of traits. In this example, there are four traits and six pairs of traits to be analyzed;

%design2(Trait1, Trait2);
%design2(Trait1, Trait3);
%design2(Trait1, Trait4);
%design2(Trait2, Trait3);
%design2(Trait2, Trait4);
%design2(Trait3, Trait4);
run;
```

REFERENCES

Anderson, T.W. 1958. An introduction to multivariate statistical analysis. John Wiley & Sons, New York.

Berry, D.P., F. Buckley, P. Dillon, R.D. Evans, M. Rath, and R.F. Veerkamp. 2002. Genetic parameters for level and change of body condition score and body weight in dairy cows. J. Dairy Sci. 85: 2030–2039.

Boldman, K.G., L.A. Kriese, L.D. Van Vleck, C.P. Van Tassel, and S.D. Kachman. 1993. A manual for the use of MTDFREML. USDA-ARS, Clay Center, NE.

Bureau, F., C. Michaux, J. Coghe, U. Uystepruyst, P.L. Leroy, and P. Lekeux. 2001. Spirometric performance in Belgian Blue calves: II. Analysis of environmental factors and estimation of genetic parameters. J. Anim. Sci. 79:1162–1165.

Conington, J., S.C. Bishop, B. Grundy, A. Waterhouse, and G. Simm. 2001. Multi-trait selection indexes for sustainable UK hill sheep production. Anim. Sci. 73:413–423.

de Souza, V.A.B., D.H. Byrne, and J.F. Taylor. 1998. Heritability, genetic and phenotpyic correlations, and predicted selection response of quantitative traits in peach: I. An analysis of several reproductive traits. J. Am. Soc. Hortic. Sci. 123:598–603.

Efron, B. 1982. The jackknife, the bootstrap, and other resampling plans. Soc. Industr. Appl. Math., Philadelphia.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics, 4th ed. Longman Technical, Essex, UK.

Fry, J.D. 2004. Estimation of genetic variances and covariances by restricted maximum likelihood using PROC MIXED. p. 11–34. *In* A.M. Saxton (ed.) Genetic analysis of complex traits using SAS. SAS Institute, Cary, NC.

Gilmour, A.R., B.R. Cullis, S.J. Wellham, and R. Thompson. 1999. ASREML reference manual. NSW Agriculture biometric bulletin no. 3. NSW Agriculture, Orange, NSW, Australia.

Hallauer, A.R., and J.B. Miranda. 1988. Quantitative genetics in maize breeding. 2nd ed. Iowa State Univ. Press, Ames.

Holland, J.B., K.J. Frey, and E.G. Hammond. 2001. Correlated response of fatty acid composition and grain quality and agronomic traitsto nine cycles of recurrent selection for increased oil content in oat. Euphytica 122:69–79.

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martinez. 2003. Estimating and interpreting heritability for plant breeding: An update. p. 9–111. *In* J. Janick (ed.) Plant breeding reviews. Vol. 22. Wiley, New York.

Holland, J.B., V.A. Portyanko, D.L. Hoffman, and M. Lee. 2002. Genomic regions controlling vernalization and photoperiod responses in oat. Theor. Appl. Genet. 105:113–126.

Legarra, A., and E. Ugarte. 2001. Genetic parameters of milk traits in Laxta dairy sheep. Anim. Sci. 73:407–412.

Littell, R.C., G.A. Milliken, W.A. Stroup, and R.D. Wolfinger. 1996. SAS system for mixed models. SAS Institute Inc., Cary, NC.

Little, R.J.A., and D.R. Rubin. 1987. Statistical analysis with missing data. John Wiley & Sons, New York.

Liu, B.H., S.J. Knapp, and D. Birkes. 1997. Sampling distributions, biases, variances, and confidence intervals for genetic correlations. Theor. Appl. Genet. 94:8–19.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, MA.

Meyer, K. 1985. Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. Biometrics 41:153–165.

Mode, C.J., and H.F. Robinson. 1959. Pleiotropism and the genetic variance and covariance. Biometrics 15:518–537.

Neumaier, A., and E. Groeneveld. 1998. Restricted maximum likelihood estimation of covariances in sparse linear models. Genet. Sel. Evol. (Paris) 30:3–26.

Persson, T., and B. Andersson. 2003. Genetic variance and covariance patterns of growth and survival in northern *Pinus sylvestris.* Scand. J. For. Res. 18:332–343.

Piepho, H.P., and J. Mohring. 2005. Best linear unbiased prediction of cultivar effects for subdivided target regions. Crop Sci. 45: 1151–1159.

Roff, D.A., and R. Preziosi. 1994. The estimation of the genetic correlation: The use of the jackknife. Heredity 73:544–548.

SAS Institute Inc. 1999. SAS/STAT user's guide, Vers. 8. SAS Institute Inc., Cay, NC.

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. John Wiley & Sons, New York.

Shoemaker, J.S., I.S. Painter, and B.S. Weir. 1999. Bayesian statistics in genetics. Trends Genet. 15:354–358.

Singh, M., S. Ceccarelli, and S. Grando. 1997. Precision of the genotypic correlation estimated from variety trials conducted in incomplete block designs. Theor. Appl. Genet. 95:1044–1048.

Sorrells, M.E., and S.R. Simmons. 1992. Influence of environment on the development and adaptation of oat. p. 115–163. *In* H.G. Marshall and M.E. Sorrells (ed.) Oat science and technology. ASA, Madison, WI.

Swallow, W.H., and J.F. Monahan. 1984. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. Technometrics 26:47–57.

Wright, S.P. 1998. Multivariate analysis using the MIXED procedure. p. 1238–1242. Proc. 38th Annual SAS Users Group International Conference, Nashville, TN. SAS Institute, Cary, NC.

Xie, C., and J.A. Mosjidis. 1999. Influence of sample size on precision of genetic correlations in red clover. Crop Sci. 39:863–867.

Zamudio, F., and R.D. Wolfinger. 2002. Growth increments and stability over time in fast-growing forest tree species. Can. J. For. Res. 32:942–953.

Zhu, J., and B.S. Weir. 1996. Mixed model approaches for diallel analysis based on a bio-model. Genet. Res. (Cambridge) 68:233–240.

Supplement 1 To “Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maximum Likelihood Estimation with SAS PROC Mixed” by J.B. Holland.

SAS code to estimate genotypic and phenotypic correlations and their standard errors from a multiple environment experiment with incomplete block designs within each environment and a one-way classification of genotypes.

```
*Create “tall” dataset, following the example in appendix A;

*Create a macro to perform multivariate reml analysis;
 %macro correlation(TraitI, TraitJ);

*Select the two traits to be analyzed from the tall dataset;
 data subset; set tall; if trait = “&TraitI” or if trait =
 “&TraitJ”;

*Perform multivariate REML estimation of variance and covariance components, using the “asycov” option of proc mixed to obtain the asymptotic variance-covariance matrix of the estimates;
 proc mixed data=subset asycov;
    class trait env rep genblock;

*Treat environments and replications as fixed effects to speed computation of the variance and covariance estimates of interest. Note that the F-tests associated with these factors are testing the hypothesis that there are no significant differences among environments or among replications for the two traits combined. Such hypothesis are generally not of real interest, instead tests of main effects of environments and replications, if they are of interest, should be conducted on each trait separately with univariate analyses;
    model y = env(trait) rep(env*trait) block(rep*env*trait);

*Treat genotypes (“geno”) and genotype-by-environment interactions (“geno*env”) as random effects and estimate their variance and covariance components for the two traits with the following codes;
    random trait / subject = geno type = un;
     random trait / subject = geno*env type = un;

*Model the residual error term (“rep*geno(env)”) to allow for covariances between error effects on the two traits measured on the same plot, but not between different plots;
    repeated trait/ sub = rep*geno(env) type = un;

*Output the estimates of variance and covariance components to a data set called “estmat” and output the asymptotic variance-covariance matrix of those estimates to a data set called “covmat”;
    ods output covparms = estmat asycov = covmat;
 run;

*Read variance and covariance estimates (“estmat” data set, to be read into a vector called “e”) and their variance-covariance matrix (“covmat” data set, to be read into a matrix called “cov”) into proc iml to estimate correlations and their standard errors using the delta method;
 proc iml;
    use estmat; read all into e;
    use covmat; read all into cov;

*Obtain the “C” matrix by removing the extra first column of the “cov” matrix;
    C = cov(|1:nrow(cov), 2:ncol(cov)|);

*Obtain genotypic covariance (CovG) and variance components (VG1 and VG2) from the elements of the “e” vector;
    CovG = e(|2,1|); VG1 = e(|1,1|); VG2 = e(|3,1|);

* Obtain phenotypic covariance (CovP) and variance components (VP1 and VP2) from the elements of the “e” vector;
     CovP = CovG + e(|5,1|) + e(|8,1|);
     VP1 = VG1 + e(|4,1|) + e(|7,1|); VP2 = VG2 +
      e(|6,1|) + e(|9,1|);

*Create a module called “correl” that will estimate genotypic and phenotypic correlations and their standard errors;
 start correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG,
 RP, SERG, SERP);
    RG = CovG/sqrt(VG1*VG2);

*Make the derivative vector for rg, note that the order of the rows and columns of the variance covariance matrix is VG1, CovG, VG2, VGE1, CovGE, VGE2, VError1, CovError, VError2;
    dg = (−1/(2*VG1))//(1/CovG)//(−1/(2*VG2))//0//
    0//0//0//0//0;

*Compute the variance of the estimate of the genotypic correlation using the delta method, then take its square root to obtain the standard error of the genotypic correlation estimate (“serg”);
    varrg = (RG**2)*dg′*C*dg; serg = sqrt(varrg);
    RP = CovP/sqrt(VP1*VP2);

*Make the derivate vector for rp;
    d1p = −1/(2*VP1); d2p = 1/CovP; d3p = −1/(2*VP2);
    dp = d1p//d2p//d3p//d1p//d2p//d3p//d1p//d2p//d3p;

*Compute the variance of the estimate of the phenotypic correlation using the delta method, then take its square root to obtain the standard error of the phenotypic correlation estimate (“serp”);
    varrp = (RP**2)*dp′*C*dp; serp = sqrt(varrp);
 finish correl;

*Run the “correl” module and display the results;
 call correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG,
 RP, SERG, SERP);
 print "Genotypic Correlation Between &TraitI and
 &TraitJ";
 print RG serg;
 print "Phenotypic Correlation Between &TraitI and
 &TraitJ ";
 print RP serp;
 quit;
 run;
```

```
* End the macro;
  %mend correl;

* Invoke the correlation macro for each pair of traits. In this example, there are four traits and six pairs of traits to be analyzed;
  %correlation(Trait1, Trait2);
  %correlation(Trait1, Trait3);
  %correlation(Trait1, Trait4);
  %correlation(Trait2, Trait3);
  %correlation(Trait2, Trait4);
  %correlation(Trait3, Trait4);
  run;
```

Supplement 2 to "Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maximum Likelihood Estimation with SAS PROC Mixed" by J.B. Holland.

SAS code to estimate genotypic and phenotypic correlations and their standard errors from a single environment experiment with an incomplete block design and a one-way classification of genotypes.

```
* Create "tall" dataset, following the example in appendix A;

* Create a macro to perform multivariate reml analysis;
  %macro correlation(TraitI, TraitJ);

* Select the two traits to be analyzed from the tall dataset;
  data subset; set tall; if trait = "&TraitI" or if trait = "&TraitJ";

* Perform multivariate REML estimation of variance and covariance components, using the "asycov" option of proc mixed to obtain the asymptotic variance-covariance matrix of the estimates;
  proc mixed data=subset asycov;
  class trait rep block geno;

* Treat environments and replications as fixed effects to speed computation of the variance and covariance estimates of interest. Note that the F-tests associated with these factors are testing the hypothesis that there are no significant differences among environments or among replications for the two traits combined. Such hypothesis are generally not of real interest, instead tests of main effects of environments and replications, if they are of interest, should be conducted on each trait separately with univariate analyses;
     model y = rep(trait) block(rep*trait);

* Treat genotypes ("geno") and genotype-by-environment interactions ("geno*env") as random effects and estimate their variance and covariance components for the two traits with the following codes;
     random trait/subject = geno type = un;

* Model the residual error term ("rep*geno(env)") to allow for covariances between error effects on the two traits measured on the same plot, but not between different plots;
      repeated trait/ sub = rep*geno type = un;

* Output the estimates of variance and covariance components to a data set called "estmat" and output the asymptotic variance-covariance matrix of those estimates to a data set called "covmat";
     ods output covparms = estmat asycov = covmat;
  run;

* Read variance and covariance estimates ("estmat" data set, to be read into a vector called "e") and their variance-covariance matrix ("covmat" data set, to be read into a matrix called "cov") into proc iml to estimate correlations and their standard errors using the delta method;
  proc iml;
     use estmat; read all into e;
     use covmat; read all into cov;

* Obtain the "C" matrix by removing the extra first column of the "cov" matrix;
     C = cov(|1:nrow(cov), 2:ncol(cov)|);

* Obtain genotypic covariance (CovG) and variance components (VG1 and VG2) from the elements of the "e" vector;
     CovG = e(|2,1|); VG1 = e(|1,1|); VG2 = e(|3,1|);

* Obtain phenotypic covariance (CovP) and variance components (VP1 and VP2) from the elements of the "e" vector;
     CovP = CovG + e(|5,1|);
     VP1 = VG1 + e(|4,1|); VP2 = VG2 + e(|6,1|);

* Create a module called "correl" that will estimate genotypic and phenotypic correlationsand their standard errors;
  start correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG, RP, SERG, SERP);
     RG = CovG/sqrt(VG1*VG2);

* Make the derivative vector for rg, note that the order of the rows and columns of the variance covariance matrix is VG1, CovG, VG2, VGE1, CovGE, VGE2, VError1, CovError, VError2;
     dg = (−1/(2*VG1))//(1/CovG)//(−1/(2*VG2))//0//0//0;

* Compute the variance of the estimate of the genotypic correlation using the delta method, then take its square root to obtain the standard error of the genotypic correlation estimate ("serg");
     varrg = (RG**2)*dg‘*C*dg; serg = sqrt(varrg);
     RP = CovP/sqrt(VP1*VP2);

* Make the derivate vector for rp;
     d1p = −1/(2*VP1); d2p = 1/CovP; d3p = −1/(2*VP2);
     dp = d1p//d2p//d3p//d1p//d2p//d3p;

* Compute the variance of the estimate of the phenotypic correlation using the delta method, then take its square root to obtain the standard error of the phenotypic correlation estimate ("serp");
     varrp = (RP**2)*dp′*C*dp; serp = sqrt(varrp);
  finish correl;

* Run the "correl" module and display the results;
  call correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG,
     RP, SERG, SERP);
  print "Genotypic Correlation Between &TraitI and &TraitJ";
  print RG serg;
  print "Phenotypic Correlation Between &TraitI and &TraitJ ";
  print RP serp;
  quit;
  run;

* End the macro;
%mend correl;

* Invoke the correlation macro for each pair of traits. In this example, there are four traits and six pairs of traits to be analyzed;
  %correlation(Trait1, Trait2);
  %correlation(Trait1, Trait3);
  %correlation(Trait1, Trait4);
  %correlation(Trait2, Trait3);
  %correlation(Trait2, Trait4);
  %correlation(Trait3, Trait4);
  run;
```

Supplement 3 to "Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maximum Likelihood Estimation with SAS PROC Mixed" by J.B. Holland.

SAS code to estimate genotypic and phenotypic correlations and their standard errors from a single environment experiment with a randomized complete block design and a one-way classification of genotypes.

```
* Create "tall" dataset, following the example in appendix A;

* Create a macro to perform multivariate reml analysis;
  %macro correlation(TraitI, TraitJ);

* Select the two traits to be analyzed from the tall dataset;
  data subset; set tall; if trait = "&TraitI" or if trait = "&TraitJ";

* Perform multivariate REML estimation of variance and covariance components, using the "asycov" option of proc mixed to obtain the asymptotic variance-covariance matrix of the estimates;
  proc mixed data=subset asycov;
  class trait rep geno;

* Treat environments and replications as fixed effects to speed computation of the variance and covariance estimates of interest. Note that the F-tests associated with these factors are testing the hypothesis that there are no significant differences among environments or among replications for the two traits combined. Such hypothesis are generally not of real interest, instead tests of main effects of environments and replications, if they are of interest, should be conducted on each trait separately with univariate analyses;
    model y = rep(trait);

* Treat genotypes ("geno") and genotype-by-environment interactions ("geno*env") as random effects and estimate their variance and covariance components for the two traits with the following codes;
    random trait/subject = geno type = un;

* Model the residual error term ("rep*geno(env)") to allow for covariances between error effects on the two traits measured on the same plot, but not between different plots;
    repeated trait/ sub = rep*geno type = un;

* Output the estimates of variance and covariance components to a data set called "estmat" and output the asymptotic variance-covariance matrix of those estimates to a data set called "covmat";
    ods output covparms = estmat asycov = covmat;
  run;

* Read variance and covariance estimates ("estmat" data set, to be read into a vector called "e") and their variance-covariance matrix ("covmat" data set, to be read into a matrix called "cov") into proc iml to estimate correlations and their standard errors using the delta method;
  proc iml;
    use estmat; read all into e;
    use covmat; read all into cov;
* Obtain the "C" matrix by removing the extra first column of the "cov" matrix;
    C = cov(|1:nrow(cov), 2:ncol(cov)|);
* Obtain genotypic covariance (CovG) and variance components (VG1 and VG2) from the elements of the "e" vector;
    CovG = e(|2,1|); VG1 = e(|1,1|); VG2 = e(|3,1|);
* Obtain phenotypic covariance (CovP) and variance components (VP1 and VP2) from the elements of the "e" vector;
    CovP = CovG + e(|5,1|);
    VP1 = VG1 + e(|4,1|); VP2 = VG2 + e(|6,1|);
* Create a module called "correl" that will estimate genotypic and phenotypic correlations and their standard errors;
  start correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG, RP, SERG, SERP);
    RG = CovG/sqrt(VG1*VG2);
* Make the derivative vector for rg, note that the order of the rows and columns of the variance covariance matrix is VG1, CovG, VG2, VGE1, CovGE, VGE2, VError1, CovError, VError2;
    dg = (−1/(2*VG1))//(1/CovG)//(−1/(2*VG2))//0//0//0;
* Compute the variance of the estimate of the genotypic correlation using the delta method, then take its square root to obtain the standard error of the genotypic correlation estimate ("serg");
    varrg = (RG**2)*dg′*C*dg; serg = sqrt(varrg);
    RP = CovP/sqrt(VP1*VP2);
* Make the derivate vector for rp;
    d1p = −1/(2*VP1); d2p = 1/CovP; d3p = −1/(2*VP2);
    dp = d1p//d2p//d3p//d1p//d2p//d3p;
```

```
*Compute the variance of the estimate of the phenotypic correlation using the delta method, then take
 its square root to obtain the standard error of the phenotypic correlation estimate (“serp”);
    varrp = (RP**2)*dp‘*C*dp; serp = sqrt(varrp);
 finish correl;

*Run the “correl” module and display the results;
 call correl(C, CovG, VG1, VG2, CovP, VP1, VP2, RG,
 RP, SERG, SERP);
 print "Genotypic Correlation Between &TraitI and
 &TraitJ";
 print RG serg;
 print "Phenotypic Correlation Between &TraitI and
 &TraitJ ";
 print RP serp;
 quit;
 run;

*End the macro;
 %mend correl;

*Invoke the correlation macro for each pair of traits. In this example, there are four traits and six pairs of traits
 to be analyzed;
 %correlation(Trait1, Trait2);
 %correlation(Trait1, Trait3);
 %correlation(Trait1, Trait4);
 %correlation(Trait2, Trait3);
 %correlation(Trait2, Trait4);
 %correlation(Trait3, Trait4);
 run;
```