

Use of canonical variates in genetic divergence studies

Daniel Furtado Ferreira¹ and Gabriel Dehon Sampaio Peçanha Rezende²

ABSTRACT

Exact correlation coefficients between a canonical variate and measured traits were derived to evaluate genetic divergence among varieties. This method allows the plant breeder to determine which traits contribute significantly to genetic divergence and, also, to identify the most important among them. An example is presented, related to a trial where 28 varieties of maize were evaluated for two traits.

INTRODUCTION

Genetic divergence is one of the most important parameters evaluated by plant breeders in starting a breeding program. This is a necessary, but not sufficient, condition for the occurrence of heterosis and the generation of a population with broad genetic variability. Subsequently, heterosis is directly proportional to genetic divergence and to dominance squared (Falconer, 1981; Cruz, 1990; Ferreira, 1993) and is also associated with adaptation.

The usual approach to make inferences about genetic divergence is to adopt predictive methodologies. Among them, diallel crosses are the most important. In this case, crosses are made among materials and a great number of hybrids is obtained. These must be evaluated over several years and environmental conditions, which increases the initial cost of the breeding program.

A second approach is to use multivariate methods to estimate genetic divergence and then predict hybrid performance. In this case, it is not necessary to make crosses. Furthermore, a large number of materials may be successfully evaluated (Hallauer and Miranda Filho, 1981).

In the latter approach, a large number of traits must be measured. A canonical variate technique is often used to reduce the number of these traits, through a linear combination of them, without a significant loss of the total variation. Additionally, this technique takes into account the structure of residual covariances. Thus, it allows plant breeders to obtain information about traits that are important for genetic divergence among varieties. This information can be obtained from the correlation between canonical variates and traits. This coefficient of correlation may be subdivided into two parts, the first is variation among varieties (phenotypic) and the other is residual variation (within-group).

Ferreira (1993) presents an approach to estimate the correlation coefficient due to variation among varieties, when the residual covariance matrix is not different from the identity matrix. However, when this assumption is not verified, the approach can lead breeders to discard important traits. This study was initiated to derive the exact correlation coefficients between a canonical variate and the traits used in genetic divergence studies.

METHODOLOGY

Let X_1, X_2, \dots, X_p be the traits (1, 2, ..., p), measured in a replicated variety trial to evaluate genetic divergence. Let Y_1, Y_2, \dots, Y_p be the canonical variates (Fisher's discriminant function) which are linear combinations of the traits, given by:

¹ Departamento de Ciências Exatas, Universidade Federal de Lavras, UFLA, Caixa Postal 37, 37200-000 Lavras, MG, Brasil. Send correspondence to D.F.F

² Aracruz Florestal, Aracruz, Espírito Santos, ES, Brasil.

$$Y_{(i)} = \underline{e}^{(i)'} X = e_1^{(i)} X_1 + e_2^{(i)} X_2 + \dots + e_p^{(i)} X_p \quad (1)$$

$i = 1, 2, \dots, p$

Let T and E be the sum of squares and product matrix ($p \times p$) due to varieties and residual, respectively, and λ_i and $\underline{e}^{(i)}$ be eigenvalues and eigenvectors ($p \times 1$) related to the i th canonical variate. The residual covariance matrix is considered to be the same for every variety generating a pooled matrix. Then, the variance among varieties and the residual variance of the i th canonical variate are:

$$\text{Var}_V(Y_i) = \underline{e}^{(i)'} T \underline{e}^{(i)} = \lambda_i \quad (\text{among varieties}) \quad (2)$$

$$\text{Var}_R(Y_i) = \underline{e}^{(i)'} E \underline{e}^{(i)} = 1 \quad (\text{residual}) \quad (3)$$

The covariance among varieties and the residual covariance between two different canonical variates are:

$$\text{Cov}_V(Y_i, Y_k) = \underline{e}^{(i)'} T \underline{e}^{(k)} = 0 \quad i \neq k = 1, 2, \dots, p \quad (4)$$

$$\text{Cov}_R(Y_i, Y_k) = \underline{e}^{(i)'} E \underline{e}^{(k)} = 0 \quad i \neq k = 1, 2, \dots, p \quad (5)$$

The eigenvalues (λ_i) and eigenvectors ($\underline{e}^{(i)}$) related to the i th canonical variate are obtained from the solution of the homogeneous indeterminate system:

$$(T - \lambda_i E) \underline{e}^{(i)} = 0 \quad (6)$$

A convenient way to solve equation (6) is by transforming it into a problem of determination of principal components (Johnson and Wichern, 1988).

First, the transformation matrix S^{-1} (Bock, 1975) is obtained by carrying out a Cholesky decomposition of E, such that $E = SS'$, where S is a lower triangular matrix. Then:

$$S^{-1} E (S^{-1})' = I \quad (7)$$

The same transformation may be applied to T:

$$L = S^{-1} T (S^{-1})' \quad (8)$$

The new system to be solved is given as follows:

$$(L - \lambda_i I) \underline{z}^{(i)} = 0 \quad (9)$$

Where $\underline{z}^{(i)}$ is the i th eigenvector of the transformed system (9).

The solution to equation (9) is obtained with the extraction of eigenvalues and eigenvectors from matrix L. The eigenvalues are invariant under nonsingular transformation (Bock, 1975), but the eigenvectors ($\underline{z}^{(i)}$) are modified and must be recovered by:

$$\underline{e}^{(i)} = (S^{-1})' \underline{z}^{(i)} \quad (10)$$

The approximate correlation coefficient due to varieties between the traits and the canonical variates is given by the correlation coefficient ($r_{i,k}^{(V)}$) between the i th principal component of equation (9) and the k th trait (Johnson and Wichern, 1988):

$$r_{i,k}^{(V)} = \frac{z_k^{(i)} \sqrt{\lambda_i}}{\sqrt{S_{k,k}}} \quad (11)$$

where $z_k^{(i)}$ is the k th element of the i th eigenvector ($\underline{z}^{(i)}$), and $S_{k,k}$ is the sum of squares among varieties of the k th trait under the nonsingular transformation, obtained in the diagonal of matrix L, given in equation (8).

To determine the exact correlation due to variation among varieties ($r_{i,k}^{(V)}$) between the i th canonical variate (Y_i) and the k th trait (X_k), let $\underline{1}'_k = [0, \dots, 0, 1, 0, \dots, 0]$, $X_k = \underline{1}'_k X$ and $Y_i = \underline{e}^{(i)'} X$, as presented in equation (1). Thus:

$$\text{Cov}_V(X_k, Y_i) = \text{Cov}_V(\underline{1}'_k X, \underline{e}^{(i)'} X) = \underline{1}'_k T \underline{e}^{(i)} \quad (12)$$

From equation (6), it is clear that:

$$T \underline{e}^{(i)} = \lambda_i E \underline{e}^{(i)} \quad (13)$$

And using (13) in (12):

$$\text{Cov}_{(V)}(X_k, Y_i) = \lambda_i \underline{1}'_k E \underline{e}^{(i)} = \lambda_i \sum_{j=1}^p \sigma_{e(k,j)} e_j^{(i)} \quad (14)$$

where $\sigma_{e(k,j)}$ is the k th row and j th column element of E (symmetric). X_k and Y_i variances are:

$$\text{Var}_V(X_k) = \text{Var}_V(\underline{1}'_k X) = \underline{1}'_k T \underline{1}_k = \sigma_{t(k,k)} \quad (15)$$

where $\sigma_{t(k,k)}$ is the k th row and column element of T.

$$\text{Var}_V(Y_i) = \underline{e}^{(i)'} T \underline{e}^{(i)} = \lambda_i \quad (16)$$

Then:

$$r_{i,k}^{(V)} = \frac{\text{Cov}_V(X_k, Y_i)}{\sqrt{\text{Var}_V(X_k) \text{Var}_V(Y_i)}} = \frac{\sqrt{\lambda_i} \sum_{j=1}^p \sigma_{e(k,j)} e_j^{(i)}}{\sqrt{\sigma_{t(k,k)}}} \quad (17)$$

Generally, a normalized solution ($\underline{e}^{*(i)}$) is taken to the eigenvector $\underline{e}^{(i)}$:

$$\underline{e}^{*(i)'} \underline{e}^{*(i)} = 1 \quad (18)$$

In this case:

$$\underline{e}^{*(i)'} E \underline{e}^{*(i)} = c_i \quad (19)$$

where c_i is a scale factor. Then:

$$\lambda_i E \underline{e}^{*(i)} = \sqrt{c_i} \lambda_i E \underline{e}^{(i)} \quad (20)$$

In this specific case, the correlation coefficient given in equation (17) must be divided by $\sqrt{c_i}$. Since equation (20) belongs to the numerator of the expression, it is biased by a factor of $\sqrt{c_i}$, where c_i is obtained from equation (19). Thus:

$$r_{i,k}^{(V)} = \frac{\sqrt{\lambda_i} \sum_{j=1}^p \sigma_{e(k,j)} \underline{e}_j^{*(i)}}{\sqrt{c_i} \sqrt{\sigma_{t(k,k)}}} \quad (21)$$

To determine the correlation coefficient due to residual variation between the i th canonical variate and the k th trait ($r_{i,k}^{(R)}$) the following results are needed:

$$\text{Cov}_R(X_k, Y_i) = \underline{1}'_k E \underline{e}^{(i)} = \sum_{j=1}^p \sigma_{e(k,j)} \underline{e}_j^{(i)} \quad (22)$$

The variances due to X_k and Y_i residual variation are:

$$\text{Var}_R(X_k) = \text{Var}_R(\underline{1}'_k X) = \underline{1}'_k E \underline{1}_k = \sigma_{e(k,k)} \quad (23)$$

$$\text{Var}_R(Y_i) = \underline{e}^{(i)'} E \underline{e}^{(i)} = 1 \quad (24)$$

Then, the exact residual correlation coefficient is:

$$r_{(i,k)}^{(R)} = \frac{\sum_{j=1}^p \sigma_{e(k,j)} \underline{e}_j^{(i)}}{\sqrt{\sigma_{e(k,k)}}} \quad (25)$$

With the normalized solution (18), equation (25) results in:

$$r_{(i,k)}^{(R)} = \frac{\sum_{j=1}^p \sigma_{e(k,j)} \underline{e}_j^{*(i)}}{\sqrt{c_i} \sqrt{\sigma_{e(k,k)}}} \quad (26)$$

EXAMPLE

Part of the data presented by Ferreira (1993) was used in this example. Only two traits were used to illustrate the estimation of the correlation coefficients. The traits were X_1 (stalk diameter - SD) and X_2 (number of leaves - NL) measured in 28 varieties of maize, evaluated in a trial with two replications. The estimated T and E matrices were:

$$T = \begin{bmatrix} 192.8486 & 5.7343 \\ 5.7343 & 126.8434 \end{bmatrix} \quad \text{and} \quad E = \begin{bmatrix} 88.8221 & 25.5061 \\ 25.5061 & 133.2405 \end{bmatrix}$$

The estimated S^{-1} matrix was:

$$S^{-1} = \begin{bmatrix} 0.106106 & 0.000000 \\ -0.025591 & 0.089117 \end{bmatrix}$$

And the estimated L matrix, obtained from equation (8), was:

$$L = \begin{bmatrix} 2.1712 & -0.4694 \\ -0.4694 & 1.1075 \end{bmatrix}$$

The estimated eigenvalues and eigenvectors were:

$$\lambda_1 = 2.3487 \quad \text{and} \quad \lambda_2 = 0.9300;$$

$$\underline{z}^{(1)} = \begin{bmatrix} 0.9353 \\ -0.3537 \end{bmatrix} \quad \text{and} \quad \underline{z}^{(2)} = \begin{bmatrix} 0.3537 \\ 0.9353 \end{bmatrix}$$

The estimates of the approximate correlation coefficient due to variation among varieties, between the canonical variates (Y_1 and Y_2) and the traits (X_1 and X_2), were obtained from equation (11) and are presented in Table I.

Table I - Estimates of approximate correlation coefficients due to variation among varieties, between the canonical variates (Y_1 and Y_2) and the traits (X_1 and X_2), obtained from equation (11).

Traits	Canonical variates	
	Y_1	Y_2
X_1 (SD)	0.9728	0.2315
X_2 (NL)	-0.5151	0.8571

SD, Stalk diameter; NL, number of leaves.

The eigenvalues are invariant under nonsingular transformation of variables, but the eigenvectors must be recovered through equation (10). The results are:

$$\underline{e}^{(1)} = \begin{bmatrix} 0.1083 \\ -0.0315 \end{bmatrix} \quad \text{and} \quad \underline{e}^{(2)} = \begin{bmatrix} 0.0136 \\ 0.0834 \end{bmatrix}$$

It can be verified that:

$$\begin{bmatrix} \underline{e}^{(1)'} \\ \underline{e}^{(2)'} \end{bmatrix} E \begin{bmatrix} \underline{e}^{(1)} \\ \underline{e}^{(2)} \end{bmatrix} = I$$

Using this result and equation (17), the exact correlation coefficients among varieties, between the traits and the canonical variates, were obtained and are presented in Table II.

Table II - Estimates of exact correlation coefficients among varieties, between the traits (X_1 and X_2) and the canonical variates (Y_1 and Y_2) obtained from equation (17).

Traits	Canonical variates	
	Y_1	Y_2
X_1 (SD)	0.9728	0.2315
X_2 (NL)	-0.1957	0.9807

SD, Stalk diameter; NL, number of leaves.

The estimates of approximate and exact correlation coefficients among varieties were the same for X_1 and the respective canonical variates, but were different for X_2 and the canonical variate (Tables I and II). Canonical variate Y_1 accounted for 71.64% of total variation. Since X_1 presented a high correlation (0.9728) with it, X_1 was considered the most important trait contributing to variability; therefore, it discriminates among most of the divergent varieties.

Estimates of residual correlation coefficients were obtained from equation (25) and are presented in Table III.

Table III - Residual correlation coefficients between the traits (X_1 and X_2) and the canonical variates (Y_1 and Y_2), obtained from equation (25).

Traits	Canonical variates	
	Y_1	Y_2
X_1 (SD)	0.9354	0.3539
X_2 (NL)	-0.1243	0.9927

SD, Stalk diameter; NL, number of leaves.

X_1 presented a high residual correlation coefficient with the main canonical variate (Y_1). On the other hand, X_2 presented a high estimated residual correlation coefficient with the canonical variate Y_2 , which was less important to genetic divergence among varieties (Table III).

The purpose of this example was to show the utility of the correlation coefficients presented in equations 17, 21, 25 and 26, derived in this paper. This method allows the plant breeder to determine which traits are the most important for genetic divergence among materials under study. This knowledge is useful in choosing parental lines during the first stages of breeding programs and also in the maintenance of germplasm banks.

RESUMO

O coeficiente de correlação exato entre uma variável canônica e a característica mensurada foi derivado com a finalidade de avaliar a divergência genética entre variedades. O presente método permite ao melhorista determinar qual característica tem contribuição significativa para a divergência genética e, também, identificar as mais importantes entre elas. Um exemplo real, relativo a um ensaio com 28 variedades de milho mensuradas para duas características, foi apresentado.

REFERENCES

- Bock, R.D.** (1975). *Multivariate Statistical Methods in Behavioral Research*. MacGraw-Hill, New York, pp. 623.
- Cruz, C.D.** (1990). Aplicações de algumas técnicas multivariadas no melhoramento de plantas. Doctoral thesis, ESALQ, USP, Piracicaba, SP.
- Falconer, D.S.** (1981). *Introduction to Quantitative Genetics*. 2nd edn. Longman, London, pp. 340.
- Ferreira, D.F.** (1993). Métodos de avaliação da divergência genética em milho e suas relações com os cruzamentos dialélicos. Master's thesis, UFLA, Lavras, MG.
- Hallauer, R. and Miranda Filho, J.B.** (1981). *Quantitative Genetics in Maize Breeding*. Iowa State University Press, Ames, pp. 468.
- Johnson, R. and Wichern, D.W.** (1988). *Applied Multivariate Statistical Analysis*. 2nd edn. Prentice Hall, New York, pp. 607.

(Received October 2, 1996)