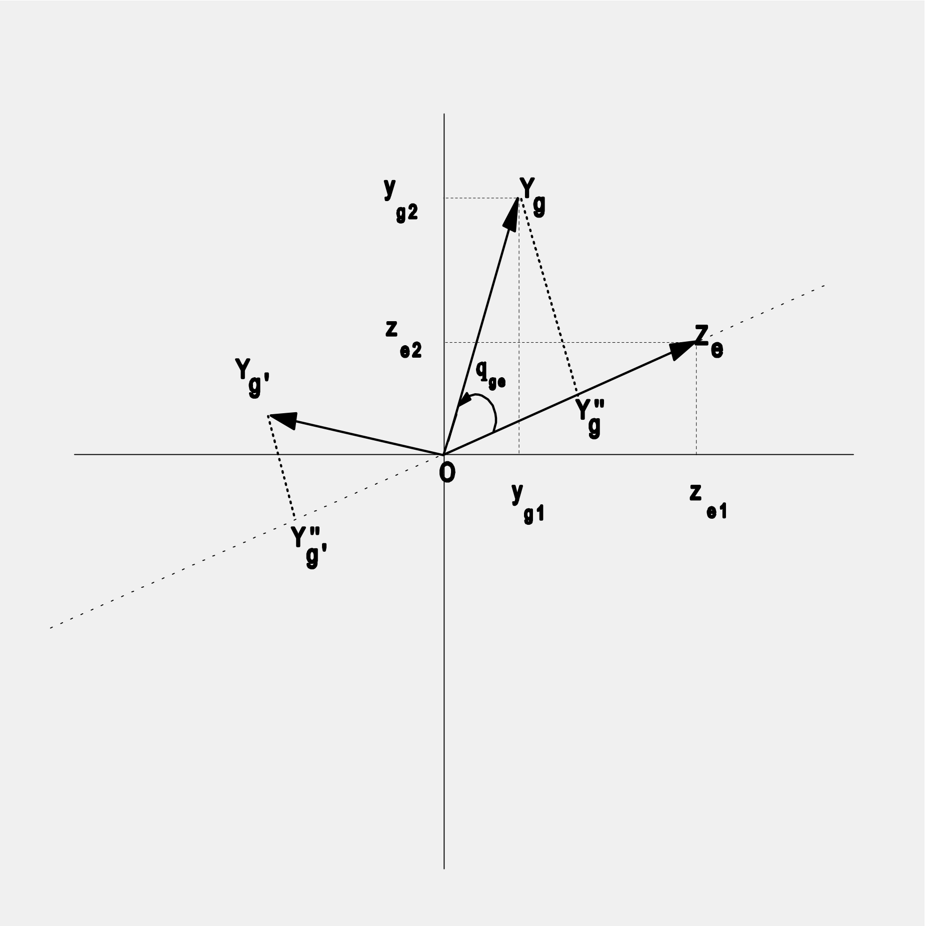
INTRODUCTION TO BIPLOTS FOR $G{\times}E$ TABLES

Pieter M. Kroonenberg
Department of Education

Leiden University



Centre for Statistics The University of Queensland
Research Report #51.
Department of Mathematics, The University of Queensland, Brisbane, Qld 4072 Australia

INTRODUCTION TO BIPLOTS FOR $G{\times}E$ TABLES[1]

Abstract

This report contains an introduction to biplots, a technique to display large tables in a graph. The construction and interpretation is explained at a fairly basic level and is directed at plant breeders. The technique is illustrated with several artificial data sets as well as a real one from maize breeding in drought conditions.

Table of Contents

[1] This report was written while on leave at the Department of Agriculture at the University of Queensland, Brisbane, Australia, and was financially supported by the Netherlands Organisation for Scientific Research (NWO) and by a grant to Dr. K.E. Basford from Rural Industries Research and Development Corporation, Canberra, Australia. Thanks are due to Kaye Basford, Samantha Watson, and Ian Phillips for their careful reading of the manuscript. This manuscript is available from: http://www.uq.edu.au/~agswatso/biplot.zip or http://www.fsw.leidenuniv.nl/~kroonenb/document/biplot.zip
Version 1: July 1995; Version 3: dd. 1-9-97

1. Introduction

Plant breeders typically conduct large-scale trials to investigate the performance of large numbers of genotypes in several environments with the aim of selecting the `best' genotypes for the purpose of further improvements of crops. The data from such trials consist of the scores on one or more attributes for each genotype in each environment, barring missing data. Generally data from several replications are available and the raw results need to be analysed by sophisticated analysis of variance techniques to assess blocking effects, to estimate variance components, etc. (see among others, Searle, Casella, McCulloch, 1992, and the course notes by Cullis and Gilmour, 1995). For the purpose of this report, we assume that such analyses have been carried out, and that for further analysis a Genotype by Environment table with (adjusted) means is available.

This table can be analysed in a straightforward way with a simple two-way analysis of variance procedure, in particular a model with the overall mean, a genotype main effect, an environment main effect, and a genotype-by-environment ($G{\times}E$) interaction may be used. Whether a real residual term is present depends on the usage of raw data with several replications (error term present) or the usage of environment means or equivalently no replications (no error term present). In the latter case, one might consider the two-way (or first-order) interaction as the error term. The latter practice is, however, wasteful and incorrect when there are many levels for either rows or columns or both, because generally there is a (large) amount of structure in the two-way interaction. Moreover, often it is the structure in this interaction which is the focus of the analysis. There are many ways in which a $G{\times}E$ table can be modelled, and Table 1 gives an overview of some of the proposals without claiming completeness or originality.

Table 1
Some models for two-way $G{\times}E$-tables[1]

	Equation	Description
1.	$x_{ij} \approx \mu + g_i + e_j$	main effects
2.	$x_{ij} \approx \mu + g_i + e_j + \lambda g_i e_j$	Tukey (1949) 1-df for interaction model
3.	$x_{ij} \approx \mu + g_i + e_j + \lambda_i e_j$	Finlay-Wilkinson (1963) regression on the environment mean; joint regression analysis
4.	$x_{ij} \approx \mu + g_i + e_j + \lambda_i z_j$	regression on an external variable z_j
5.	$x_{ij} \approx \mu + g_i + e_j + \lambda u_i v_j$	main effects plus 1 multiplicative term
6.	$x_{ij} \approx \mu + g_i + e_j + \Sigma_p \lambda_p u_{ip} v_{jp}$	main effects plus P multiplicative terms (due to Mandel, 1971); also called AMMI-model[2]
7.	$x_{ij} \approx \mu + e_j + \Sigma_p \lambda_p u_{ip} v_{jp}$	genotype main effect is included in the multiplicative model
	$x'_{ij} = x_{ij} - \mu - e_j$ $\approx \Sigma_p \lambda_p u_{ip} v_{jp}$	x'_{ij} is the centred version of x_{ij}
8.	$x_{ij} \approx \mu + e_j + s_j \Sigma_p \lambda_s u_{ip} v_{jp}$	x''_{ij} environment standardised version of x_{ij}
	$x''_{ij} = (x_{ij} - \mu - e_j)/s_j$	s_j is the scaling factor of the j-th environment (usually standard deviation)[3]

[1] ≈: is modelled by; for equality an error term should be added;

[2] AMMI (Additive Main effects and Multiplicative Interaction model) is a name sponsored by Gaugh (see e.g. Gaugh, 1988)

[3] model recommended by Cooper & DeLacy (1994) for use with selection.

When there is a large table with interaction, there is a need for methods to analyse the two-way interaction in such a way that, if there are systematic patterns present, they can be readily assessed and their relevance can be evaluated. Plots which show both the genotypes and the environments simultaneously can be of great assistance in this respect, and these plots, called *biplots* (Gabriel, 1971), are the subject of this report. The prefix *bi* refers to the simultaneous display of *both* rows and columns of the table, and not to the two-dimensionality of the plots.

Generally, when one has a table of G genotypes and E environments, there are at most *min(G,E)* dimensions possible. For definiteness sake, we will assume in the sequel that there are more genotypes than environments, so that G is greater than E and thus there are atmost E dimensions possible. As displays of more than two dimensions are generally difficult to make and even more difficult to interpret, most biplots show only two dimensions. Obviously, one wants a display in which the interaction between genotypes and environments is presented as well as possible. In other words, one wants to display those dimensions which account for the maximum amount of variation in the table. This implies that we have to find a procedure which provides us with the `best' representation in low-dimensional space. The appropriate tool for this is derived from a theorem presented by Eckart-Young (1936), and the technique is called the *singular value decomposition* (SVD). This technique provides us with coordinates on dimensions (or directions in space); in the mathematical literature these dimensions are called *singular vectors*. The dimensions are arranged in such a way that they are *orthogonal*, i.e. at right angles, and successively represent as much of the variation as possible (see the Appendix for an elementary introduction into vectors, and concepts such as orthogonality). Moreover, the technique provides us with measures (*singular values*) which, if squared, indicate the amount of variability accounted for by each dimension. To display the main variability in the table in a two-dimensional graph, we should use the first two dimensions.

2. Singular value decomposition

2.1 Basic theory

Suppose that we have a two-way data matrix ***X*** with information on a single attribute, say yield, for *G* genotypes in *E* environments, and that there are more genotypes than environments, so that *min(G,E)=E*. The singular value decomposition SVD of the matrix ***X*** is defined as

$$\mathrm{X}=\mathrm{U}\Lambda\mathrm{V}', \tag{1}$$

which may be written in summation notation as

$$x_{g\,e}=\sum_{s=1}^{S}\lambda_s u_{gs} v_{es}\,. \tag{2}$$

where *S* is in most practical cases equal to *E*, i.e. we generally need *E* terms to perfectly reproduce the original matrix ***X***. The scalars λ_s are the singular values arranged in decreasing order of magnitude, ($\boldsymbol{u}_s$) is a set of genotype vectors (the left singular vectors), and ($\boldsymbol{v}_s$) is a set of environment vectors (the right singular vectors). In both sets the vectors are *orthonormal*, i.e. they are pairwise at right angles and have lengths equal to one. ***U*** and ***V*** are matrices which have the vectors $\boldsymbol{u}_s$ and $\boldsymbol{v}_s$ as their columns, respectively. If the entries in the table are the interactions from a two-way analysis of variance on the original table (Model 6 of Table 1), then both $\boldsymbol{u}_s$ and the $\boldsymbol{v}_s$ are *centred*, i.e. each column of ***U*** and ***V*** has a zero mean, because the original table of interaction effects is centred. Moreover, in this case *S* is at most *E-1*, because centring reduces the number of independent dimensions by one.

The $\boldsymbol{u}_s$ and $\boldsymbol{v}_s$ are used to construct the coordinates for graphical representations of the data. In particular, they can be combined with the singular values λ_s in different ways, of which the following two versions are the most common ones:

$$x_{g\,e}=\sum_{s=1}^{S}u_{gs}\big(\lambda_s v_{es}\big)=\sum_{s=1}^{S}y_{gs}\,z_{es} \tag{3}$$

$$x_{g\,e}=\sum_{s=1}^{S}(u_{gs}\lambda_s^{1/2})(v_{es}\lambda_s^{1/2})=\sum_{s=1}^{S}y^*_{gs}\,z^*_{es}\ , \tag{4}$$

where the *y* and the *z* are the genotype and environment coordinates of the first version (*principal component scaled* version), and y^* and the z^* those of the second version (*symmetrically scaled* version), respectively (see section 3.3).

2.2 Low-dimensional approximation

To find a low-dimensional approximation of ***X*** we have to minimise the distance between the original matrix and the approximating matrix, $\hat{X}$. This (Euclidean) distance between two matrices, ***X***=(x_{ge}) and $\hat{X} = (\hat{x}_{ge})$, is defined as

and the Eckart-Young (1936) theorem shows that the best two-dimensional least-squares

$$d(X, \hat{X}) = \sqrt{\sum_{g=1}^{G}\sum_{e=1}^{E}\left(x_{ge} - \hat{x}_{ge}\right)^2}, \qquad (5)$$

approximation of the matrix ***X*** can be obtained from the SVD of ***X*** by summing only the first two terms of equation (2).

2.3 Quality of approximation

To evaluate the quality of the approximation, we have to know how much of the original variability of ***X*** is contained in the approximating matrix $\hat{X}$. The total variability in a matrix, here defined as the uncorrected sum of squares, is equal to the sum of squared entries in the table,

$$\textit{Total variability} = SS_x = \|X\|^2 = \sum_{g=1}^{G}\sum_{e=1}^{E} x_{ge}^2 \,, \qquad (6)$$

where $\_X\_1$ is called the *norm* of ***X***. Because of the least-squares properties of the singular value decomposition, the norm can be split into an explained and a residual part, i.e.

$$\|X\|^2 = \|\hat{X}\|^2 + \|X - \hat{X}\|.^2 \qquad (7)$$

Furthermore, one can use the orthonormality of ***U*** and ***V*** to show that this equation may be expressed in terms of the singular values, i.e.

$$\sum_{s=1}^{S}\lambda_s^2 = \sum_{s=1}^{2}\lambda_s^2 + \sum_{s=3}^{S}\lambda_s^2. \qquad (8)$$

Equation (8) shows that the sum of the first two squared singular values divided by the total sum of the squared singular values will give the proportion of the variability accounted for by the first two singular vectors. Large proportions of explained variability will obviously indicate that the plot based on these two singular vectors will give a good representation of the structure in the table. If only a moderate or low proportion of the variability is accounted for, the main structure of the table will still be represented in the graph, but some parts of the structure may reside in higher dimensions. If the data are environment centred, genotypes located near the origin might either have all their values close to the environment means, or their variability is located in another dimension. Similarly, environments close to the origin may have little variability or may not fit well in two dimensions.

3. Biplots

The most common graph to portray the relationships in a table is the biplot (Gabriel, 1971, 1981, Gabriel & Odoroff, 1990, Kempton, 1984). Fig. 1-5 are the biplots of our examples.

3.1 Standard biplots

A standard biplot is the display of a $G \times E$ (interaction) table $\boldsymbol{X}$ decomposed into a product $\boldsymbol{YZ'}$ of a $G \times S$ matrix $\boldsymbol{Y}=(y_{gs})$ and an $E \times S$ matrix $\boldsymbol{Z}=(z_{es})$. Using this decomposition for $\hat{\mathrm{X}}$, the two-dimensional approximation of $\boldsymbol{X}$ each element $\hat{x}_{g\,e}$ of this matrix can be written as

$$\hat{x}_{g\,e} = y_{g1}\, z_{e1} + y_{g2}\, z_{e2}\,, \tag{9}$$

which is the *inner (or scalar) product* of the row vectors (y_{g1}, y_{g2}) and (z_{e1}, z_{e2}); for further information on inner products see the Appendix. A biplot is obtained by representing each row as a point Y_g with coordinates (y_{g1}, y_{g2}), and each column as point Z_e with coordinates (z_{e1}, z_{e2}) in a two-dimensional graph (with origin O). These points are generally referred to as *row markers* and *column markers*, respectively. Sometimes the word `markers' is also used for the coordinate vectors themselves. Because it is not easy to evaluate markers in a three-dimensional space, the most commonly used biplots are two-dimensional, which thus display the best rank-two approximation of a matrix $\boldsymbol{X}$. With the current state of graphical software, it is likely that three-dimensional biplots will become more common. A straight line through the origin O and a point, say Z_e, is often called a *biplot axis*, and is written as OZ_e, not to be confused with a coordinate axis.

If we write Y_g'' for the orthogonal projection of Y_g on the biplot axis OZ_e, θ_{ge} for the angle between the vectors OY_g and OZ_e, and write $|OZ|$2 for the length of a vector OZ, then we have the geometric equivalent of equation (9) (see also the Appendix)

$$\hat{x}_{g\,e} = |OZ_e||OY_g|\cos(\theta_{g\,e}) = |OZ_e||OY_{g''}|\,. \tag{10}$$

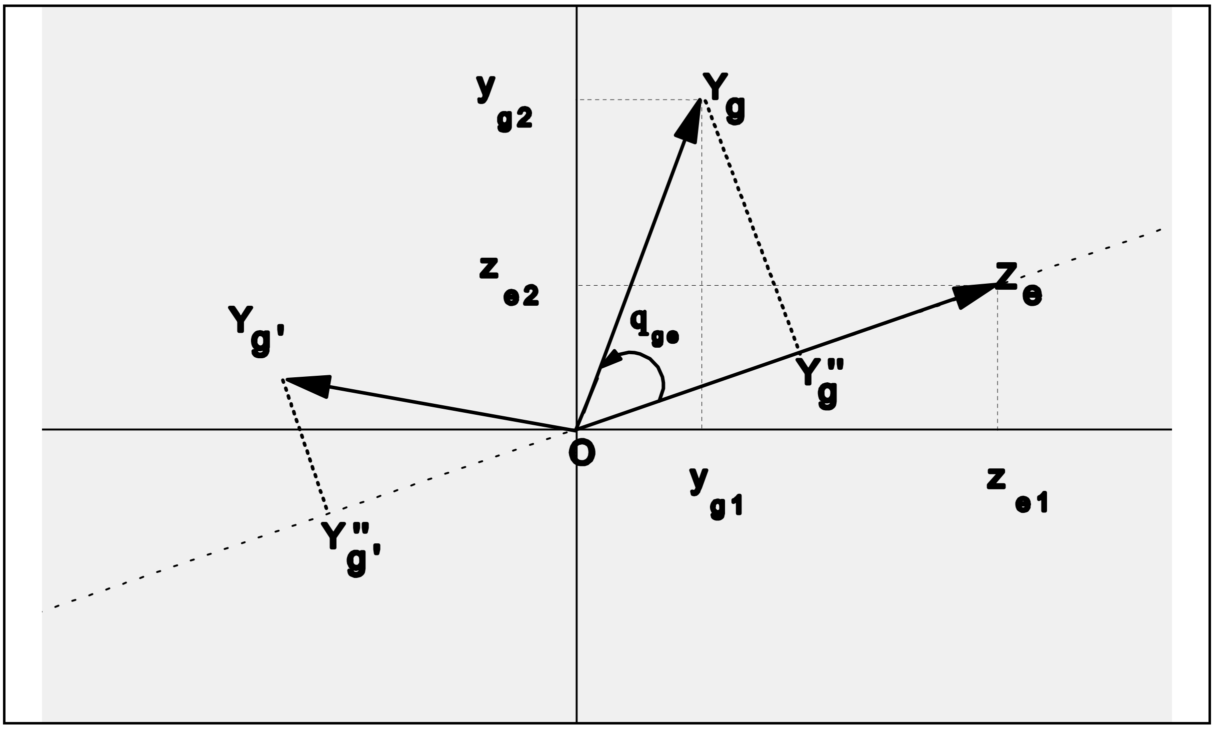


Figure 1 Representation of two genotype markers and one environment marker in a biplot.

Equation (10) shows that $\hat{x}_{ge}$ is proportional to the length of OY_g'', $|OY_{g'}|$. This relationship is of course true for any other genotype g' as well. Thus the relationships or interactions of two genotypes with the same environment can be assessed simply by comparing the lengths of their projections onto that environment. Furthermore, the relationship or interaction between a genotype vector OY_g and an environment vector OZ_e is positive if their angle is acute, and negative in the case of an obtuse angle. When the projection of a marker Y_g onto the environment vector OZ_e coincides with the origin, $\hat{x}_{ge}$ is equal to zero, and the genotype has approximately a mean value for that environment given that the data were environment centred (Models 7 and 8, Table 1). A positive value for $\hat{x}_{ge}$ indicates that genotype *g* has high score in environment *e* relative to the average score in that environment, and a negative value indicates genotype *g* has a relatively low score in environment *e*.

In graphs, the genotype markers Y_g are generally represented by points, and the environment markers Z_e by vectors, so that the two types of markers can be clearly distinguished. This choice is preferred because genotypes are compared with respect to an environment rather than the reverse.

3.2 Calibrated biplots

Because inner products between the coordinates of the genotype markers Y_g and those of a column marker Z_e vary linearly along the biplot axis OZ_e, it is possible to mark (or calibrate) the biplot axis OZ_e linearly in such a way that the $\hat{x}_{ge}$ can be directly read from the graph (Gabriel & Odoroff, 1990; Greenacre, 1993). Note that the approximate value $\hat{x}_{ge}$ does not depend on the position of Y_g, but only on the orthogonal projection $Y_{g'}$ onto the axis OZ_e. When a data matrix is centred as is the case with environment centred data, the approximating matrix is centred as well, and a value of $\hat{x}_{ge}$ equal to zero means that, in the *e*-th uncentred environment, genotype *g* has a value approximately equal to the mean of the *e*-th environment. One could mark the biplot axes according to the (approximations of) the environments according to the centred values. However, sometimes it is also informative to replace the centred values with the `real' values by adding the observed means. After this *decentring*, the origin indicates the true mean values for the environments, rather than zero for all of them.

3.3 Two different versions of the biplot

In section 2.1 the two most common decompositions of $\boldsymbol{X}$ were presented both based on the SVD. These two decompositions lead to different biplots with different properties. Equations (3) and (4) show that the values of the inner products between genotype and environment markers are independent of the version used, so that in this respect the two versions are equivalent. However, when looking at the relationships within each set of markers, the two decompositions lead to different interpretations.

With the principal component scaling (equation (3)) the genotypes are in so-called *standard coordinates*, i.e. they have zero means and unit lengths, and the environments are in *principal coordinates*, i.e. they have unrestricted means and lengths equal to the associated singular values. If in the data matrix $\boldsymbol{X}$ the environments are standardised, then the coordinates of the environments may be interpreted as correlations between the environments and the coordinate axes. Here, all biplots will have this type of scaling.

With the symmetrical scaling (equation (4)) the correlation interpretation cannot be used, because both the genotype components and those of the environments have lengths equal to the square root of the singular values. Therefore, this version should primarily be used when the relations between the genotypes and the environments are the central focus in the analysis, and not the relations among genotypes and/or among environments, or when the row and column variables play a comparable role in the analysis. The advantage of the representation is that lengths of the environment and the genotype vectors in the biplot are approximately equal. With principal component scaling it can easily happen that the genotypes are concentrated around the origin of the plots, while the environments are located on the rim, and vice versa.

3.4 Interpretational rules

An important point in constructing the actual graphs for biplots is that **the physical vertical and horizontal coordinate axes should have the same physical scale**. This will ensure that when one projects genotypes on an environment vector, they will end up in the correct place. Failing to adhere to this scaling will make it impossible to evaluate inner products in the graph.

The most basic property of any kind of biplot of a table at a particular dimensionality, is that the inner product of a row (genotype) vector and a column (environment) vector in the plot is the best approximation to the the corresponding value in the table. If there is a perfect fit in, say two, dimensions, then the inner products are identical to the values in the table. The majority of the rules given below follow from this basic property. Additional interpretations become available if special treatments have been applied to (1) the rows and/or columns, such as centring and standardisation, and (2) to the coordinate axes, such as principal component scaling and symmetric scaling. Below we will only present those interpretational rules which we think are relevant for G×E tables, in particular we will not consider the situation when the original table is analysed without centring.

General (irrespective of scaling coordinate axes)

- genotypes are perferably *displayed* as points and environments as vectors;
- if two genotype vectors have a small angle, they have similar response patterns over environments;
- if two environment vectors have a small angle they are strongly associated.

Centred per environment

- the biplot displays the table of genotype main effect plus the two-way interaction (Model 7 in Table 1);
- genotypes are in deviation from the average for each of the environments;
- the origin represents the average value for each environment, i.e. it represents the genotype which has an average value in each environment. This average genotype has a value of zero in the centred data matrix;
- a genotype with a large distance from the origin has a large genotype plus interaction effect;
- the larger the projection of a genotype on an environment vector, the more this genotype deviates from the average in the environment;

Centred per environment and per genotypes

- the biplot displays the two-way interaction table; there are at most min(G,E) dimensions or coordinate axes (Model 6 in Table 1);
- both genotypes and environments are in deviation from their averages;
- the origin represents the average value both for each environment and for each genotype across all environments;
- a genotype (environment) with a large distance from the origin has a large interaction effect with at least one environment (genotype);
- the larger the projection of a genotype on an environment vector, the more this genotype

deviates from the average in the environment, and vice versa.

Principal component scaling: $\boldsymbol{U}$ and $\boldsymbol{V\Lambda}$ (Principal component biplot)

Centred per environment

- the cosine of the angle between any two environments approximate their correlation with equality if the fit is prefect;
- the lengths of the environment vectors are approximately proportional to the standard deviations of the environments with exact proportionality if the fit is perfect;
- the inner product between two environments approximates their covariance with equality if the fit is perfect;
- the euclidean distance between two genotypes does not approximate the distances between their rows in the original matrix but their standardised distance, which is the square root of the so-called Mahalanobis distance (for further details, see Gabriel, 1971, p. 460ff.);
- environments can have much longer vectors than genotypes, making visual inspection awkward; a partial remedy is to multiply all environment coordinates with an *arbitrary* constant, which will make the relative lengths of the environment and genotype vectors comparable. Note, however, that there is no obligation to use such a constant, and that it is an ad-hoc measure.

Standardised per environment

- the lengths of the environment vectors indicate how well the environments are represented by the graph with a perfect fit all vectors have equal lengths;
- the inner product between two environments (and the cosine of the angle between them) approximates their correlation with equality if the fit is prefect;

Symmetric scaling: $\boldsymbol{U\Lambda}^{½}$ *and* $\boldsymbol{V\Lambda}^{½}$

- if two environment vectors have a small angle, they are highly correlated, but their correlation cannot be deduced from the graph; similarly the association between the genotypes cannot be properly read from the graph;
- due to the symmetric scaling of environments and genotypes, both are located in the same part of the space and inner products are easily assessed.

4. Examples with perfectly two-dimensional data

To illustrate some of the properties, biplots will be presented of three variants of a small data set, each of which fits perfectly in two dimensions. The first analysis will be with raw data, in the second one the environment means have been removed (Model 7 of Table 1), and in the third analysis each centred environment has been scaled with its standard deviation (Model 8 of Table 1). The data sets have been derived from each other, but it is impossible to create perfectly two-dimensional centred data by centring a perfectly two-dimensional raw data set, in contrast to creating the standardised data set from the centred one. Because the data sets are fit perfectly in two-dimensions, the biplot will exactly represent the original data. This means, for instance, that the inner products calculated from the biplots are equal to the data themselves. Moreover, the standard deviations of the environments can exactly be gauged from the biplot from the lengths of the environment vectors. If the data had been three-dimensional, these lengths would only have been an approximation. Furthermore in cases of imperfect fit, large data values will be represented by large inner products, small data values by small inner products, but not all values in the original data matrix will be fitted equally well by the inner products.

4.1 Raw data

Raw data have not undergone any preprocessing, i.e. centring and/or scaling, and therefore the dimensions will be strongly influenced by the means. To show this the means of the genotypes and the environments have been included in Table 3. Table 3 shows the (near) perfect rank correlation between the first dimensions and the means for both the environments and the genotypes, indicating that these dimensions represent the differences between the means.

Table 2. Raw Data

		Environments	
Genotypes	A	B	C
G1	.7316	1.4522	.8412
G2	.7665	1.5404	.8812
G3	.6972	1.4712	.8007
G4	.7662	1.5767	.8803
G5	.7358	1.2862	.8481
G6	.6496	1.4294	.7453
G7	.6997	1.1826	.8071
G8	.6330	1.4379	.7257
G9	.7120	0.0251	.8355
Length	2.135	4.037	2.460
σ	0.047	0.481	0.055
s	0.044	0.454	0.051

From Table 3 and Fig. 2 and the size of the variability accounted for (97% and 3%, respectively), the dominance of the first dimension is obvious. The cause is the strong dominance of the means in the analysis. Fig. 2 is rather lopsided, because the representation of the genotypes is in standard coordinates, and that of the environments in principal coordinates with lengths equal to the singular values (5.11 and .81, respectively). Therefore, the vectors for the environments are much longer than those of the genotypes.

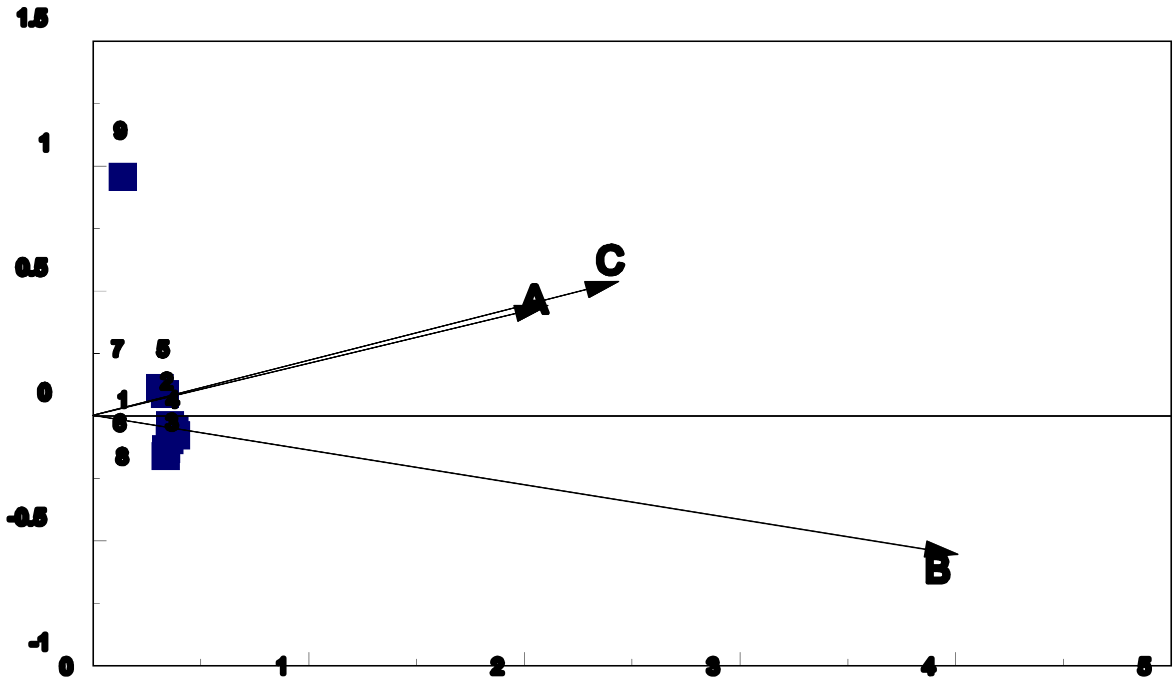


Figure 2 Biplot of the perfectly two-dimensional raw data (Note: The scaling of the horizontal and vertical axes is not equal).

Table 3 Genotype and Environment Coordinates for the Raw Data

		Components$^{\$}$				Components^{ú}	
Genotypes	Mean	1	2	Environments	Mean	1	2
G4	1.07	.384	-.078	B	1.27	3.998	-.549
G2	1.06	.378	-.052	C	0.82	2.403	.528
G1	1.01	.358	-.036	A	0.71	2.086	.444
G3	0.99	.355	-.097				
G6	0.94	.339	-.132				
G8	0.93	.337	-.161	*length*		5.11	0.81
G5	0.96	.334	.088	*proportion*			
G7	0.90	.311	.113	*explained*		.97	.03
G9	0.52	.138	.957				

$^{\$}$ **standard coordinates;** $^{£}$ **principal coordinates**

4.2 Data centred by environments

The raw data have been processed by subtracting the environment means in accordance with Model 7 of Table 1. Subsequently, they have been adjusted to make them perfectly two-dimensional.

Again the representation of the genotypes is in standard coordinates, and that of the environments is in principal coordinates (lengths 4.75 and 2.11, respectively) makes the
vectors for the environments longer than those for the genotypes, but not as much as for the raw data (Table 5). If we choose 4 as an arbitrary appropriate constant to adjust (here divide) all environment coordinates, the plot is more balanced and easier to read (see Fig. 3).

The *length* $|A|$*of Environment A* follows from $|A| = \sqrt{}\ (.940^2 + 1.971^2) = \sqrt{4.77} = 2.18$, which may be found from the Fig. 3 (keeping in mind the adjustment factor of 4). The *length of*

Genotype 6 is $|G6| = \sqrt{(.357^2 + -.319^2)} = .479$. The *inner product* of Genotype 6 and Environment A is $.357 \times .940 + (-.319) \times 1.971 = -.296$, which is equal to the data value for Genotype 6 in Environment A, because of the perfect fit. The *cosine of the angle* between Genotype 6 and Environment A, $\cos\theta_{G6,A}$ is the inner product divided by the lengths of the vectors, or $-.296/(2.18 \times .479) = -.28$ and the *angle* $\theta_{G6,A} = 106°$. The *projection of Genotype 6 onto the Environment A* is the vector $G6''$ and its (signed) length is equal to the length of Genotype 6 times the cosine of $\theta_{G6,A}$ or $|G6|3\cos\theta_{G6,A} = .479 \times -.296 = -.142$, where the minus sign indicates that the projection is on the opposite side from the origin from Environment A. The lengths of the environment vectors are *proportional* to their standard deviations. The background for the calculations is contained in the Appendix.

Table 4 Environment Centred Data

	Environments		
Genotypes	A	B	C
G1	0.6344	0.2027	-0.0968
G2	0.6121	0.9796	0.4942
G3	-1.2009	0.3766	0.7531
G4	0.6725	1.0540	0.5263
G5	-1.1453	1.4442	1.5314
G6	-0.2939	-1.4933	-1.0039
G7	-0.3331	-0.0537	0.0904
G8	0.6709	-2.5429	-2.1688
G9	0.3833	0.0326	-0.1258
Length	2.19	3.61	3.03
σ	0.73	1.20	1.01
s	0.77	1.28	1.07

Table 5 Genotype and Environment Coordinates for the Environment Centred Data

	Components	
Genotypes	1	2
G8	.718	-.003
G6	.357	-.319
G9	.028	.181
G1	.008	.318
G7	-.018	-.160
G2	-.195	.403
G4	-.208	.440
G3	-.210	-.508
G5	-.478	-.354
sum	.000	.000
length	1.000	1.000

	Components		
Environment	1	2	Length
A	0.940	1.971	2.19
C	-3.018	-0.232	3.03
B	-3.536	0.723	3.61
length	4.75	2.11	
proportion explained	.83	.17	
		A	B
correlations	B	-.24	
	C	-.50	.96

In Fig. 3 projections of all genotypes onto Environment A have been drawn, and the relative performance of the genotypes in Environment A can directly be read from the graph. When the data are not perfectly two-dimensional, the inner products and thus the lengths of the projections are only approximations to their real values. Note that the direction of the environment is of vital importance in assessing whether genotypes perform above or below average. Furthermore, note that it is not the closeness of a genotype point to the environment vector, but the size of the projection that determines the relative performance in an environment. For instance, Genotype 9 is much closer to the environment vector than Genotype 8, but the projection of Genotype 8 is larger (.6709) than that of Genotype 9 (.3833), see Table 4. It is thus incorrect to use a Euclidean distance (as one would measure with a ruler) between an environment point and a genotype point to assess their relationship.

Measuring the *angles* between the environments from the graph or calculating them from the coordinates gives $\theta_{A,B}$ = 104°,$\theta_{A,C}$ = 120°, $\theta_{B,C}$ = 16°, corresponding to *correlations* or cosines of $r_{ab} = -.24$, $r_{ac} = -.50$, and $r_{bc} = .96$.

4.3 Data standardised by environments

The environment centred data can be scaled without effecting their perfect two-dimensionality, which makes direct comparison of the results possible. We have used the (population) standard deviation σ, i.e. without degrees of freedom corrections. Alternatively, we could have used *s*.

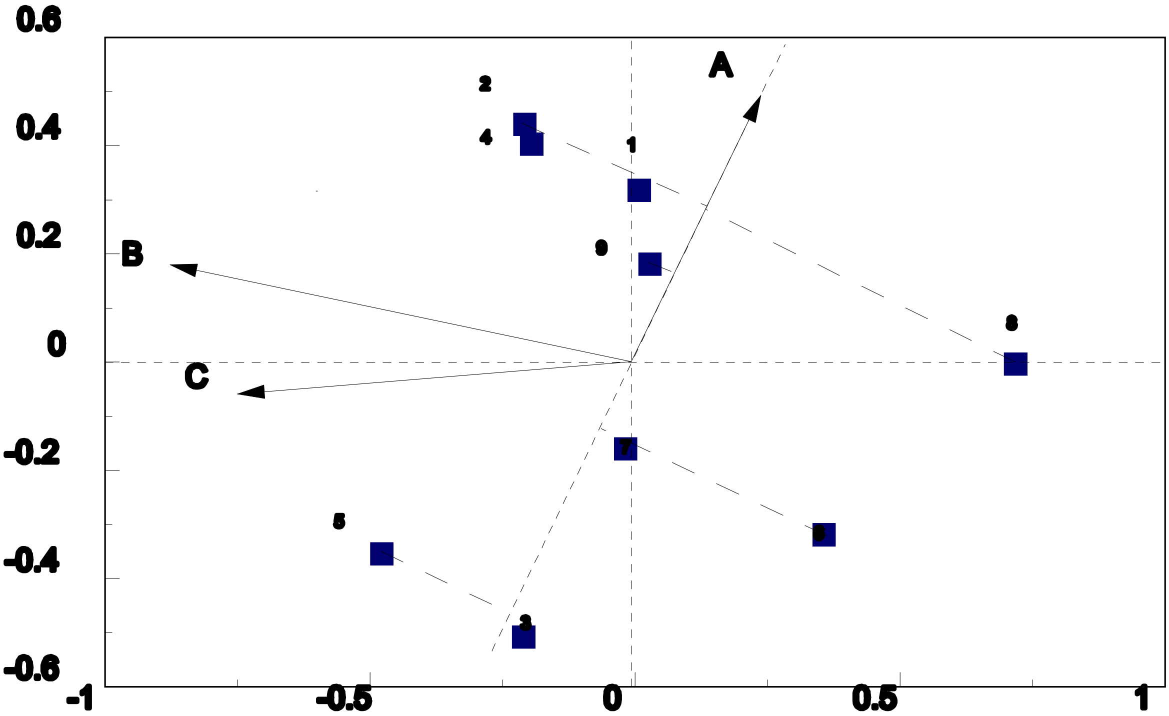


Figure 3 Biplot for the perfectly two-dimensional centred data

Table 6 Environment Standardised Data

	Environments		
Genotypes	A	B	C
G1	0.8707	0.1684	-0.0959
G2	0.8401	0.8139	0.4894
G3	-1.6483	0.3129	0.7457
G4	.9230	0.8757	0.5211
G5	-1.5719	1.1998	1.5164
G6	-0.4034	-1.2406	-0.9941
G7	-0.4572	-0.0446	0.0895
G8	0.9208	-2.1126	-2.1476
G9	0.5261	0.0271	-0.1246
Length	3.00	3.00	3.00
σ	1.00	1.00	1.00
s	1.06	1.06	1.06

From Table 7 we see that all environments have equal length vectors, and in the graph they are necessarily equal as well. When the fit is not perfect, the differences in lengths indicate differences in fit of the environments in the two dimensions shown in the biplot. Fig. 3 and 4 are fairly similar, because the standard deviations of the environments were not very different (see Table 4).

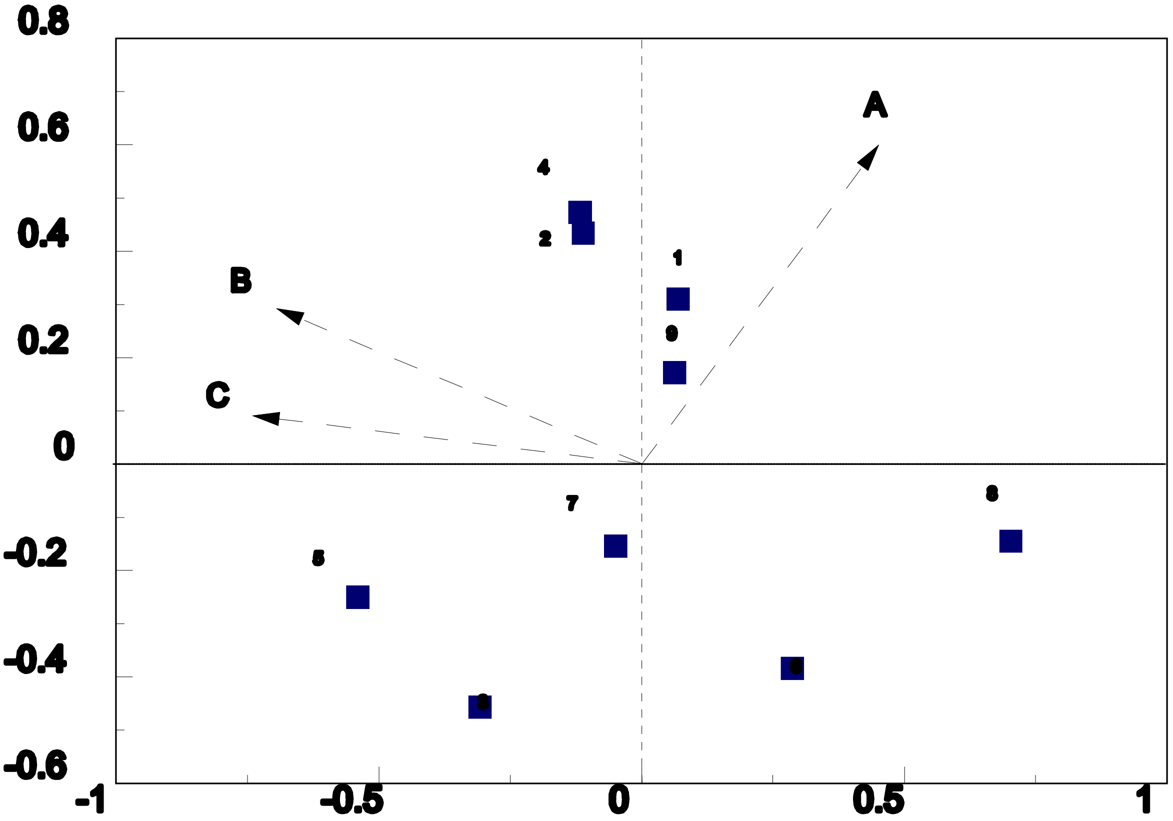


Figure 4 Biplot of the perfectly two-dimensional environment standardised data.

Table 7 Genotype and Environment Co-ordinates for the Environment Standardised Data

Genotypes	1	2
G8	.703	-.145
G6	.287	-.384
G9	.063	.172
G1	.070	.310
G7	-.049	-.154
G2	-.111	.434
G4	-.117	.473
G3	-.307	-.457
G5	-.540	-.250
sum	.000	.000
length	1.000	1.000

Environments	1	2	Length
A	1.804	2.397	3.00
B	-2.764	1.169	3.00
C	-2.977	0.366	3.00
length	4.44	2.69	
proportion explained	.73	.27	

5. Example: Mexican maize data

Ten trials were conducted to evaluate gains with recurrent (S_1 or full-sib) selection in open-pollinated genotypes from three late tropical maize populations (La Posta Sequía, Pool 26 Sequía and Tuxpeño Sequía) that have been especially selected at CIMMYT for tolerance to drought around flowering. The populations have been improved by evaluating and recombining superior families based on their performance under managed drought environments and an irrigated environment. Five of the trials subjected the plants to drought while the other trials were well-watered. The data were analysed to determine gains with selection and to determine how grain yields and other traits had been affected by selection. Included in the trials were three check cultivars which had been improved by convential breeding. Full details about the trials and the analyses as well as all references can be found in Chapman, Edmeades, & Crossa (1996).

Here the yield data will be considered to show the biplot at work with real data in a case where there is no perfect fit. The raw location means were standardised by environments (see Model 8, Table 1). The co-ordinates for the two-dimensional biplot in PCA-scaling are given in Table 8, and the biplot itself in Fig. 5. The two dimensions represent 69% of the variation in the original *G*+*G*×*E* array. A third component accounts for an additional 12%.

Given the environment standardisation of the data the cosines between the angles of two environments represent the best approximation to their correlations in two dimensions. Thus water-stressed environments (including the well-watered, but iron-deficient environment 6) are highly correlated with not much difference between the intermediate (1,5) and severely stressed environments (2,4). There is a clear distinction between stressed and nonstressed environments, apart from environment 3, which takes an intermediate position. The genotypes do not cluster according to population, but there is a clear progression with selection for each population towards increasing yield especially in stressed conditions, as is evident from the negative projections of L1, T1, and P1 on the vectors of the stressed environments, i.e. they had below average yield in those environments. Early selections of Tuxpeño and Pool 26 had also below average yields average in non-stressed environments (7,8,9,10). The latest selections (L4, T3, and P3) all have positive projections on the stressed environments, i.e above average yields. The increase has been most spectacular for Tuxpeño. La Posta yielded above average and continued to do so, although continued selection after L3 led to an increase in yield in stressed environment, but a decrease in the non-stressed environments. From the present data, it is difficult to judge whether this is a systematic or accidental deviation from the pattern. The check cultivars which have gone through convential selection did not improve their drought tolerance, as evident from their below average projections on the environment vectors.

Table 8: Genotype and Environment Coordinates for the Mexican Maize Yield Data

Genotypes		Components		*Environments*			Components[4]	
Variety[1]	Abbr.	1	2	No.	Water Regime	Year[2]	1	2
5 Check La P.	CL	-.48	.17	4	Severe Stress	1992W	.90	.02
1 La Posta C_0	L1	-.27	.18	5	Interm Stress	1992W	.90	-.13
2 La Posta C_1	L2	-.27	.27	2	Severe Stress	1993W	.86	-.15
3 La Posta C_2	L3	.02	.50	1	Interm Stress	1993W	.80	.00
4 La Posta C_3	L4	.11	.34	6	Well-Watered[3]	1992W	.78	.16
				3	Interm Stress	1993S	.86	.45
9 Check Pool	CP	-.11	-.06	8	Well-Watered	1993W	.29	.74
6 Pool 26 C_1	P1	-.07	-.45	9	Well-Watered	1992W	-.14	.80
7 Pool 26 C_2	P2	.16	-.12	10	Well-Watered	1992S	-.22	.63
8 Pool 26 C_3	P3	.26	-.11	7	Well-Watered	1992S	-.42	.51
10 DTP1 C_5	D1	.53	-.00					
11 DTP2 C_2	D2	.22	.07					
15 Check Tuxp.	CT	-.02	-.25					
12 Tuxpe±o C_0	T1	-.33	-.40					
13 Tuxpe±o C_8	T2	.02	-.20					
14 Tuxpe±o C_1	T3	.24	.07					

Proportions Explained variability
(=Squares of singular values) : Comp. 1: .47; Comp. 2: .22; Total:.69

[1] The names of the varieties have been simplified; for a full description see Chapman et al. (1996); The official name for variety 14 is TS6.

[2] W=Winter; S=Summer.

[3] This environment was well-watered but suffered from iron deficiency, which had an adverse effect on yield.

[4] The environment coordinates have been divided by 4.

6. Relationship with PCA

In principal component analysis we are looking for that linear combination $\boldsymbol{c}=\boldsymbol{Xb}$ which accounts for the largest amount of variation in a set of variables $\boldsymbol{X}$. The standard solution to this problem is constructing the sums-of-squares-and-cross-products matrix (or after centring and scaling the correlation matrix) $\boldsymbol{X'X}$, and decomposing it (via the eigenvectors and eigenvalues) in $\boldsymbol{V}\Lambda^2\boldsymbol{V'}$, furthermore $\boldsymbol{XX'}$ can be decomposed into $\boldsymbol{U}\Lambda^2\boldsymbol{U'}$. It can be shown that $\boldsymbol{U}$, $\boldsymbol{V}$, and Λ are the same as the matrices defined in equation (1). Moreover, $\boldsymbol{c}$ is equal to the first column of $\boldsymbol{U}$ and $\boldsymbol{b}$ is equal to λ_1 times the first column of $\boldsymbol{V}$. In other words, principal component analysis corresponds to the factorisation of equation (9). The parameters for a principal component analysis can thus directly be derived from the singular value decomposition. However, in PCA it is general practice that $X'X$ is a correlation matrix, while this assumption is not made for the singular value decomposition. What this shows is that PCA is really a procedure with two steps, i.e. a centring and scaling followed by a (singular value) decomposition. The separation of these two steps is generally not emphasised in genotype by environment analyses but it becomes essential when analysing three-way data of genotypes by environments by attributes.

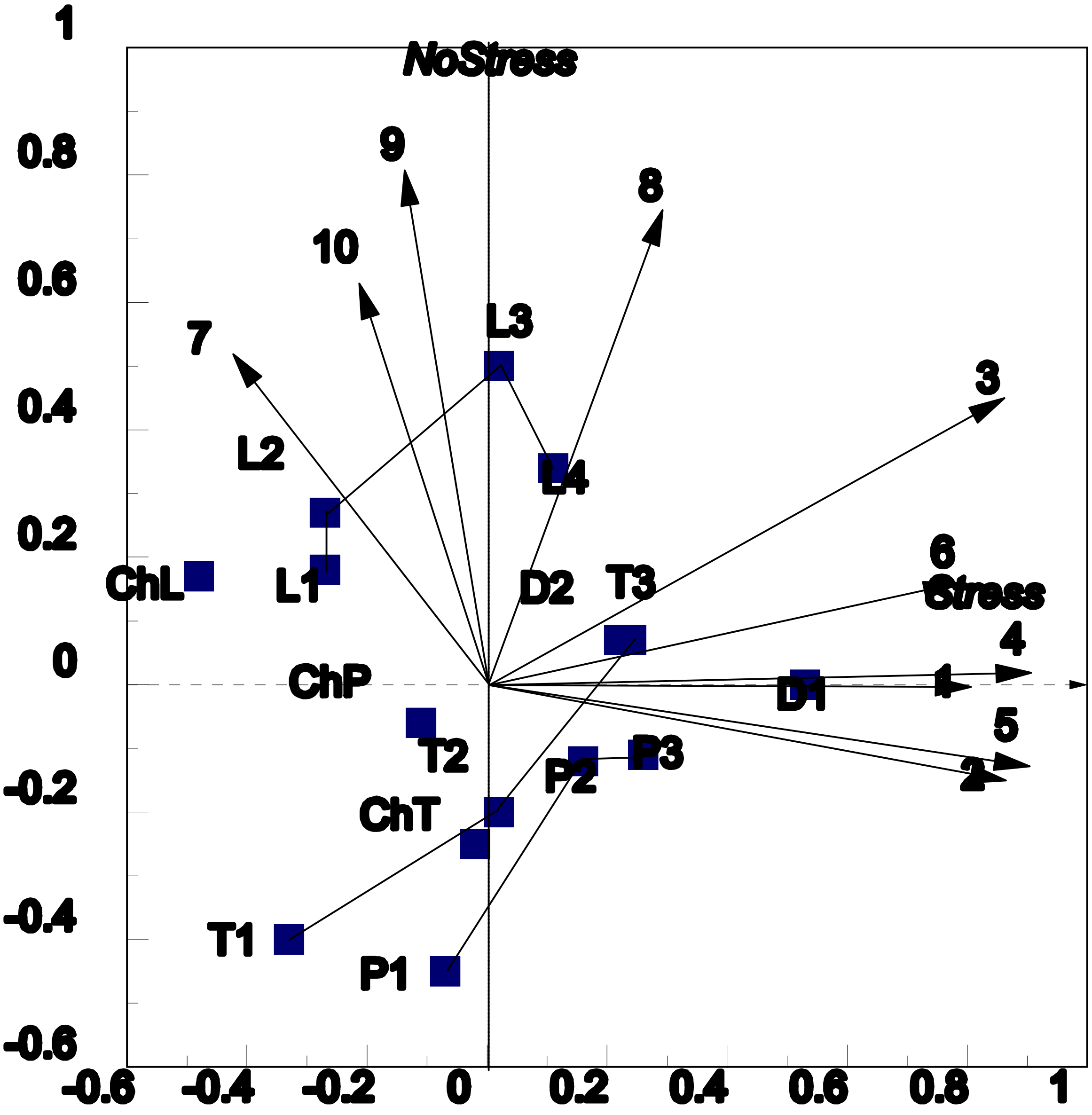


Figure 5: Biplot for Mexican Maize Yield Data

(*Legend*: P = Pool 26; L = La Posta; T = Tuxpeño; Ch* = Check of *; D = Drought Resistant Varieties 1&2)

References

Chapman, S.C., Edmeades, G.O., & Crossa, J. (1996). Pattern analysis of gains with selection for drought tolerance in tropical maize populations. In M. Cooper & G.L. Hammer (Eds.), *Plant adaptation and crop improvement* (pp. 513-527). Wallingford, UK: CAB International.

Cooper, M., & DeLacy, I.H. (1994). Relationships among analytic methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics, 88*, 561-572.

Cullis, B.R. & Gilmour, A.R. (1995). *Statistical methods for small plot field experiments in a variety evaluation programme.* Course notes "Statistical methods for small plot field experiments", Yanchep, WA, 11-17 February 1995.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*, 211-218.

Finlay, K.W., & Wilkinson, G.N. (1963). The analysis of adaptation in plant breeding. *Australian Journal of Agricultural Research, 14,* 742-754.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*, 453-467.

Gabriel, K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In V. Barnett (Ed.), *Interpreting multivariate data* (pp. 147-173). Chicester, UK: Wiley.

Gabriel, K.G., & Odoroff, C.L. (1990). Biplots in biomedical research. *Statistics in Medicine, 9*, 469-485.

Gaugh, H.G. (1988). Model selection and validation for yield trials with interaction. *Biometrics, 44*, 705-715.

Greenacre, M.J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics, 20,* 251-269.

Kempton, R.A. (1984). The use of biplots in interpreting variety by environment interactions. *Journal of Agricultural Science, Cambridge, 122*, 335-342.

Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics, 13,* 1-18.

Searle, S.R., Casella, G., & McCulloch, C.E. (1992). *Variance components*. New York: Wiley.

Tukey, J.W. (1949). One degree of freedom for additivity. *Biometrics, 5*, 232-242.

Wickens, T. (1995). *The geometry of multivariate analysis*. Hillsdale, NJ: Erlbaum.

APPENDIX

Some basic vector geometry relevant to biplots[2]

The interpretation of biplots depends heavily on properties of vectors in the plane or three-dimensional space. This appendix provides a minimal introduction into the most basic properties of vectors leading up to the ideas of inner products and projections.

Vector: Symbol: ***x*** or $\bar{x}$

(Fig. 6A) A *vector* is a directed line segment; it has a *length* and a *direction*. Mostly vectors in biplots start at the origin, the point (0,0) in a two-dimensional biplot. The coordinates of $\bar{x}$ in the two-dimensional case are (x_1,x_2), where x_1 is the value on the horizontal coordinate axis and x_2 the value on the vertical coordinate axis. Therefore, a vector $\bar{x}$ runs from (0,0) to (x_1,x_2).

Length: The length of a vector is indicated by $|\bar{x}|$, and it is found via the

(Fig. 6A) Pythagorean theorem ($a^2=b^2+c^2$): $|\bar{x}| = \sqrt{}(x_1^2+x_2^2) = \sqrt{}\,(\Sigma_i\, x_i^2)$.

Scalar multiplication:

(Fig. 6B) $\bar{y} = a\bar{x}$. The vector $\bar{x}$ is multiplied by a scalar *a*, and the resulting vector $\bar{y}$ has the same direction as $\bar{x}$, but is *a* times as long. Thus $|\bar{y}| = a|\bar{x}|$, and $y_1 = ax_1 + ax_2$.

Addition: $\bar{z} = \bar{x} + \bar{y}$, with coordinates $z_1 = x_1 + y_1$ and $z_2 = x_2 + y_2$.

(Fig. 6C)

Subtraction: $\bar{z} = \bar{x} - \bar{y}$ or $\bar{z} = \bar{x} + (-\bar{y})$ with coordinates $z_1 = x_1 - y_1$ and $z_2 = x_2 - y_2$.

(Fig. 6D)

Linear combination:

(Fig. 7A) $\bar{z} = b_x\bar{x} + b_y\bar{y}$, which is a combination of vector addition and scalar multiplication.

Angle: The angle between two vectors can be directly read or measured from a

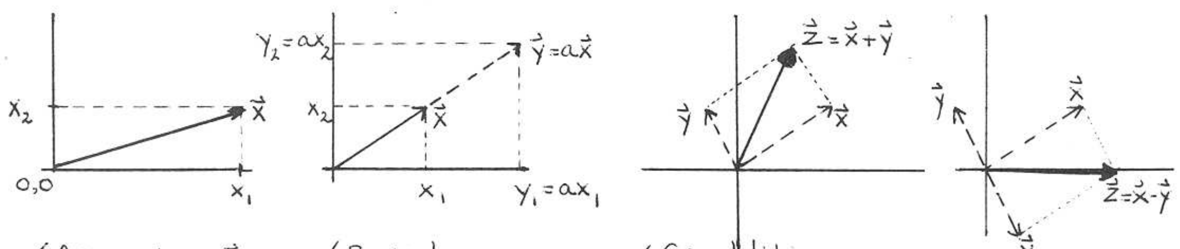
(Fig. 7B) graph, and we will indicate an angle between $\bar{x}$ and $\bar{y}$ as θ_{xy}. The angle can be computed algebraically via the inner product or dot product.

Inner product/Dot product:

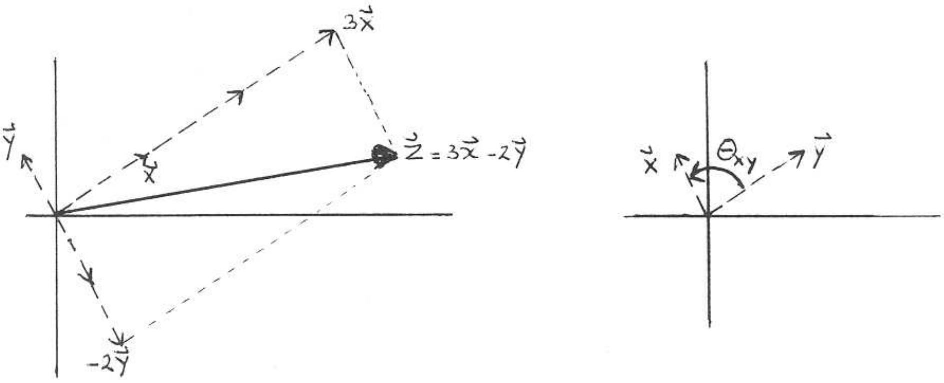
The dot product between two vectors is indicated by $\bar{x} \bullet \bar{x}$ when using vector geometry, and by $\bar{x}'\bar{y}$ when $\bar{x}$ and $\bar{y}$ are considered vectors. In the latter case the product is referred to as the inner product or scalar product of $\bar{x}$ and $\bar{y}$.

The dot product is defined as $\bar{x} \bullet \bar{y} = x_1y_1 + x_2y_2$ or in more geometric terms: $\bar{x} \bullet \bar{y} = |\bar{x}|\,|\bar{y}|\cos\theta_{xy}$, which is the length of $\bar{x}$ times the length of $\bar{y}$ times the cosine of the angle between them.

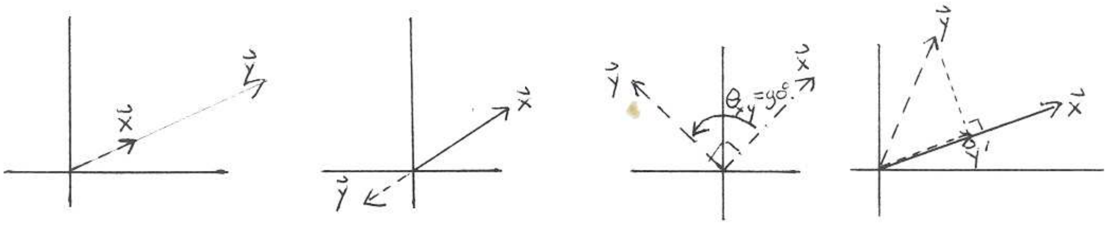
[2] Abstracted from Thomas. D. Wickens (1995). *The geometry of multivariate statistics*. Hillsdale, NJ: Erlbaum.



6A: vector $\vec{x}$ 6B: scalar multiplication 6C: addition 6D: subtraction



7A: linear combination 7B: angle



8A: collinear – b>0 8B: collinear - b<0 8C: orthogonal 8D. projection of $\vec{y}$ on $\vec{x}$

Calculation of angle:

First calculate the cosine of the angle: cosθ_{xy} = ($\vec{x}$4•$\vec{y}$5)/│$\vec{x}$6││$\vec{y}$7│, then convert the cosine to an angle via the "inverse cosine" button on your pocket calculator or look it up in a table.

Special angles:

(Fig. 8A) θ_{xy} = 0° → cos θ_{xy} = 1: $\vec{x}$8 and $\vec{y}$9 are *collinear*, i.e. they lie on the same line in the same direction; $\vec{y}$10 = *b*$\vec{x}$11 with *b*>0; $\vec{x}$12 is collinear with itself θ_{xx}=0;

(Fig. 8B) θ_{xy} = 180° → cos θ_{xy} = -1: $\vec{x}$13 and $\vec{y}$14 are *collinear*, i.e. they lie on the same line but in opposite directions; $\vec{y}$15 = *b*$\vec{x}$16 with *b*<0;

(Fig. 8C) θ_{xy} = 90° → cos θ_{xy} = 0: $\vec{x}$17 and $\vec{y}$18 are *orthogonal* (perpendicular); $\vec{x}$19•$\vec{y}$20 = 0.

Projection:

(Fig. 8D) The projection $\vec{y}$21′ of $\vec{y}$22 on $\vec{x}$23 is a vector collinear with $\vec{x}$24 which can be found by dropping a perpendicular line from $\vec{y}$25 onto $\vec{x}$26 (see figure). Thus $\vec{y}$27′ = *d*$\vec{x}$28. The length of $\vec{y}$29′ is │$\vec{y}$30│cosθ_{xy}, and *d* = ($\vec{x}$31•$\vec{y}$32)/│$\vec{x}$33│^{2}

Equality between cosines and correlations:

If the environments are centred, then the cosine of θ_{xy},the angle between two environments $\vec{x}$34 and $\vec{y}$35 is equal to their correlation r_{xy},

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}} = \frac{x \bullet y}{|x||y|} = \cos \Theta_{xy}$$

where we have used the fact that the means are zero.