The use of biplots in statistical analysis:

with examples in GenStat

Alex Glaser

VSN International, 5 The Waterhouse,
Waterhouse Street Hemel Hempstead, UK

email: `alex@vsni.co.uk`

European GenStat and ASReml Applied Statistics Conference
14 July 2010

Introduction

- Summary of biplots
  - Quick guide to biplots
  - Interpretation/scaling
  - Different visualisation techniques
- Examples of biplots in GenStat and for statistical analysis, e.g.
  - Principal components analysis
  - More specific techniques (GGE Biplot)

Quick guide to biplots

- Introduced by Gabriel (1971) to allow simultaneous display of both samples and variables from a data matrix, **Y**.

- Gower & Hand (1996): biplots can be considered multivariate analogues to a scatterplot.

- Both methods utilize the same technique
  - Samples displayed as points.
  - Variables displayed as vectors or axes (linear or nonlinear).

Factorization

- Any $n \times p$ matrix $\mathbf{Y}$ of rank r can be factorized as

$$\underset{n \times p}{\mathbf{Y}} = \underset{n \times r}{\mathbf{G}} \quad \underset{r \times p}{\mathbf{H}'}$$

- If $r = 2$, then vectors (of order two) $\mathbf{g_1}, \ldots, \mathbf{g_n}$ and $\mathbf{h_1}, \ldots., \mathbf{h_p}$ may be plotted in a standard plane, where:
  - $\mathbf{g_i}$ can be considered 'row' effects
  - $\mathbf{h_j}$ can be considered 'column' effects
- Since 'row' effects and 'column' effects plotted jointly, referred to as a *biplot*
- Both **G** and **H** non-uniquely defined some constraint required, e.g. orthonormality of one matrix.

Singular value decomposition

If **Y** is greater than rank two, cannot fully display all details of the 'row' effects and 'column' effects of the data.
Use singular value decomposition to factorize

$$\underset{n \times p}{\mathbf{Y}} = \underset{n \times r}{\mathbf{U}} \quad \underset{r \times r}{\mathbf{S}} \quad \underset{r \times p}{\mathbf{V}'} \tag{1}$$

where

- **U** and **V** are the orthonormal matrix of the *left* and *right* singular vectors respectively
- **S** is a diagonal matrix of the ordered singular values

Note that the original data matrix **Y** is rarely used in equation (1), usually a transformation is taken

Singular value decomposition

If we construct another matrix using only the first m columns of $\mathbf{U}$ and $\mathbf{V}$, and first m singular values, thus

$$\mathbf{Y}_{(m)} \quad = \quad \mathbf{U}_{(m)} \quad \mathbf{S}_{(m)} \quad \mathbf{V}'_{(m)}$$

then $\mathbf{Y}_{(m)}$ is the least-squares rank m approximation of $\mathbf{Y}$.

Gabriel (1971)

If a matrix $\mathbf{Y}$ can be satisfactorily approximated by a rank two matrix $\mathbf{Y}_{(2)}$, the biplot of $\mathbf{Y}_{(2)}$ may allow useful approximate visual inspection of $\mathbf{Y}$ itself.

Singular value decomposition

Rewriting rank 2 data matrix as

$$\mathbf{y}_{ij} = \sum_{k=1}^{2} u_{ik} s_k v_{kj}$$

$$= \sum_{k=1}^{2} \left(u_{ik} s_k^{\alpha}\right) \sum_{k=1}^{2} \left(s_k^{1-\alpha} v_{kj}\right) \quad 0 \leq \alpha \leq 1$$

Red equation represents 'row' effects (coordinates of samples)
Green equation represents 'column' effects (coordinates of the variables)
Common values of α are 1, $1/2$ and 0.
Different values of α highlight different aspects.

Scaling parameter α

Different values of α imply the following:

- $\alpha = 1$ (*row-metric preserving*)
  - Distances between samples approximates their Euclidean distance
  - Projecting a sample at right angles on a variable approximates position of sample on that variable
- $\alpha = 0$ (*column-metric preserving*)
  - Cosine of angle between axes approximates the correlation between variables.
  - Distance of variables from origin approximates variation
- $\alpha = 1/2$ (*symmetric biplots*)
  - Useful for ascertaining the relative magnitude of variation of samples and variables

Correspondence analysis

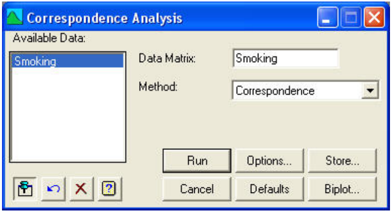
Correspondence analysis is a statistical method for representing categorical data graphically.
Example in Greenacre (1984 & 2006) of smoking habits amongst different staff.

	None	Light	Medium	Heavy
Senior Manager	4	2	3	2
Junior Manager	4	3	7	4
Senior Employee	25	10	12	4
Junior Employee	18	24	33	13
Secretary	10	6	7	2

Would like to see relationship between staff seniority and smoking habits.

Correspondence analysis menu

Extended in the 13th edition.


Correspondence Analysis

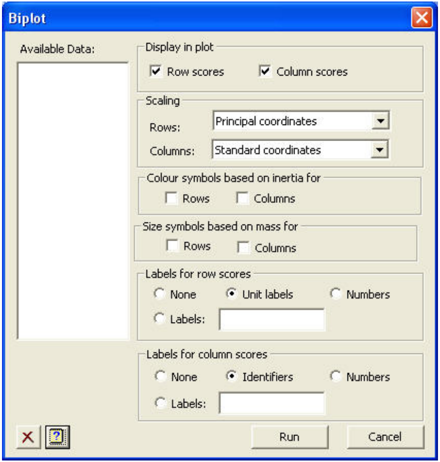
Available Data:

Smoking

Data Matrix: Smoking

Method: Correspondence

- Press the `Biplot` button
- Select scaling and other plotting options
- Press the `Run` button
    - `CABIPLOT`


Biplot

Available Data:

Display in plot

☑ Row scores ☑ Column scores

Scaling

Rows: Principal coordinates

Columns: Standard coordinates

Colour symbols based on inertia for

□ Rows □ Columns

Size symbols based on mass for

□ Rows □ Columns

Labels for row scores

○ None ● Unit labels ○ Numbers

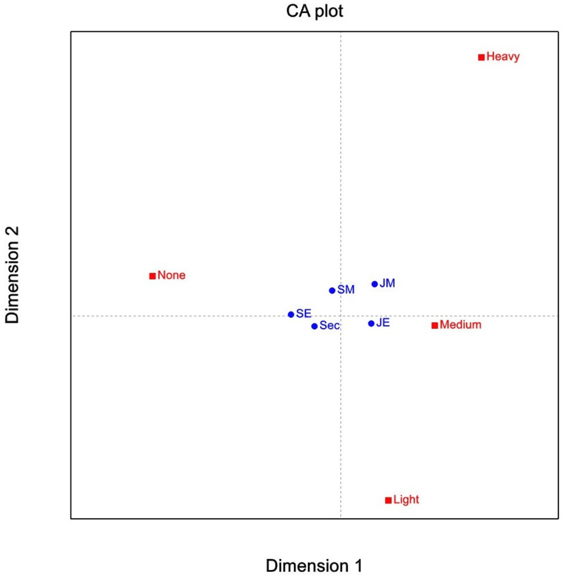
○ Labels:

Labels for column scores

○ None ● Identifiers ○ Numbers

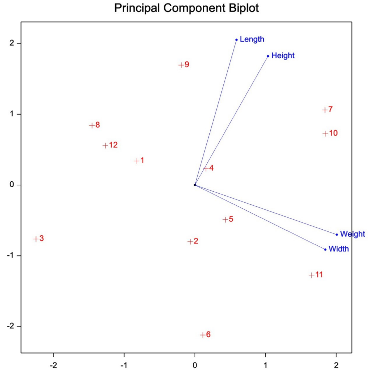
○ Labels:

Asymmetric CA biplot



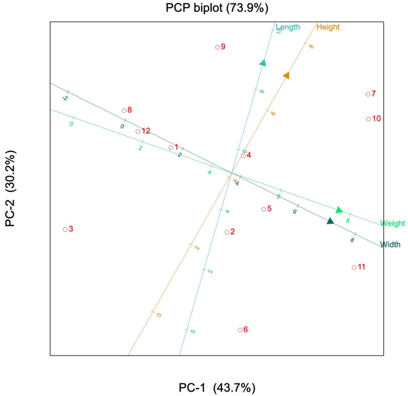
Vectors vs. axes

Vector biplot



- BIPLOT command

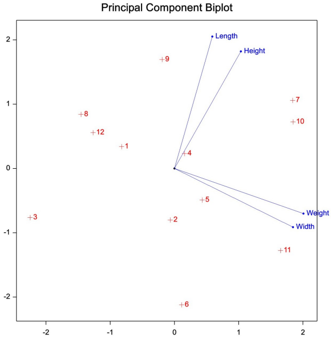
Axes biplot



- DBIPLOT command

Vectors

Vector biplot



Pros

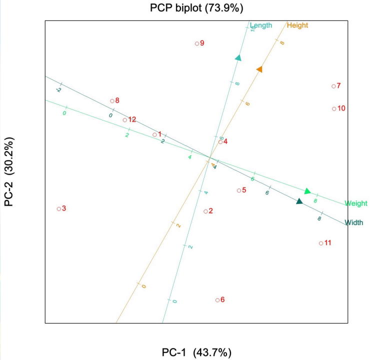
- Simplicity
- Can see contribution of each variable

Cons

- Difficult to know how to relate points to variables

Axes

Axes biplot



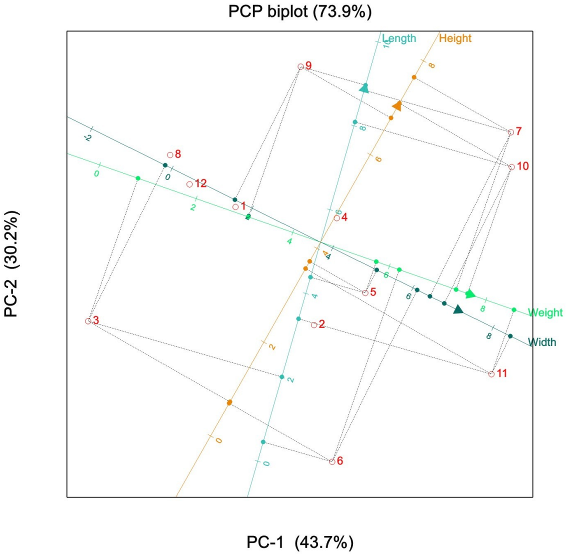
Pros

- Analogous to a scatterplot
- Can relate original data to variables

Cons

- Contribution of each variable to variation not so obvious
- More complicated design

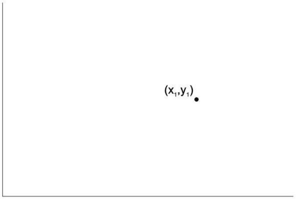
Prediction



PCP biplot (73.9%)
Length
Height
Weight
Width
PC-1 (43.7%)
PC-2 (30.2%)
1
2
3
4
5
6
7
8
9
10
11
12
-2
0
2
4
6
8
10

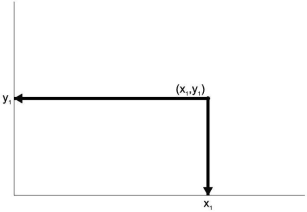
Axes

Single point in a standard plot, at point (x_1, y_1)



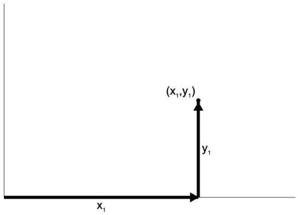
Prediction

Orthogonally project from (x_1, y_1) onto axes

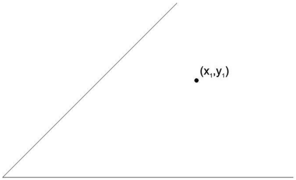


Interpolation

Alternatively, travel x_1 along x-axis and y_1 along y-axis

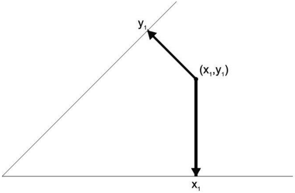


Same point in a biplot



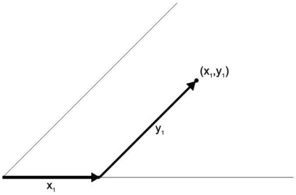
Prediction

Orthogonally project from (x_1, y_1) onto axes

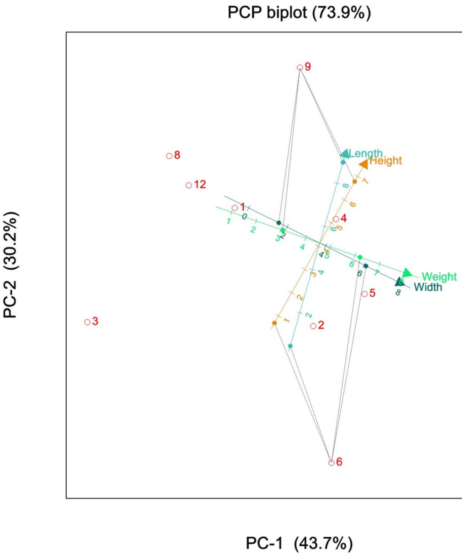


Interpolation

Alternatively, travel x_1 along x-axis and y_1 along y-axis



Interpolative axes



GGE Biplot

- Yan & Kang (2003): Observed phenotypic variation (P) of genotypes across environments is made up of environment variations (E), genotype variations (G) and genotype-by-environment interaction (GE).
- This can be written as

$$P - E = G + GE$$

- Usually E is the dominant source of variation, so environmental means removed and analysis concentrates on the genotype variation and genotype-by-environment interaction.

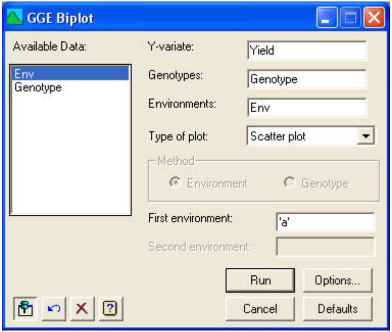
Example

Tolerance to infection by pink stem borer of seven winter wheat genotypes (*A*...*G*) in seven environments (*a*...*g*).

	a	*b*	*c*	*d*	*e*	*f*	*g*
A	27.5	35.7	46.4	53.7	33.3	64.9	43.3
B	35.7	37.5	46.2	40.8	51.9	45.6	57.5
C	46.4	46.2	38.7	49.1	50.4	55.6	69.4
D	53.7	40.8	49.1	51.2	49.4	48.1	57.5
E	33.3	51.9	50.4	49.4	42.5	63.1	68.9
F	64.9	45.6	55.6	48.1	63.1	60.0	63.1
G	43.3	57.5	69.4	57.5	68.9	63.1	43.7

GGE Biplot menu

Use GGE Biplot menu from Meta Analysis section of Stats



Available Data:
Env
Genotype

Y-variate: Yield

Genotypes: Genotype

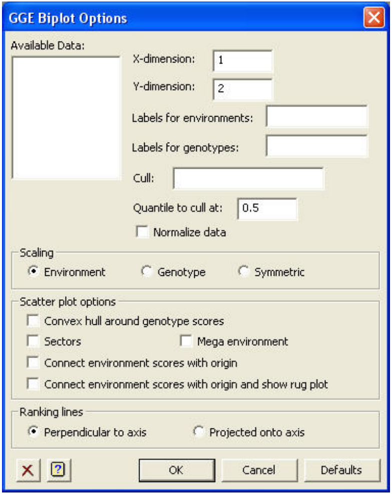
Environments: Env

Type of plot: Scatter plot

Method
● Environment ○ Genotype

First environment: 'a'

Second environment:



Available Data:

X-dimension: 1

Y-dimension: 2

Labels for environments:

Labels for genotypes:

Cull:

Quantile to cull at: 0.5

□ Normalize data

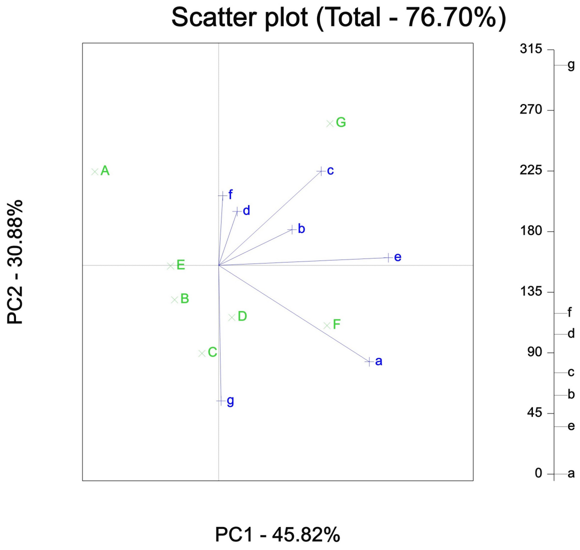
Scaling
● Environment ○ Genotype ○ Symmetric

Scatter plot options
□ Convex hull around genotype scores
□ Sectors □ Mega environment
□ Connect environment scores with origin
□ Connect environment scores with origin and show rug plot

Ranking lines
● Perpendicular to axis ○ Projected onto axis

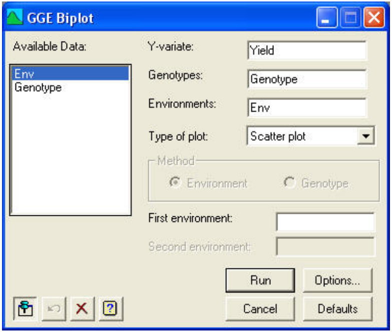
- Click on Options
- Select Connect environment scores with origin and show rug plot

GGE biplot


Scatter plot (Total - 76.70%)
PC2 - 30.88%
PC1 - 45.82%
A
G
c
f
d
b
e
E
B
D
F
C
a
g
315
270
225
180
135
90
45
0
g
f
d
c
b
e
a

GGE biplot

Staying with scatter plot



GGE Biplot

Available Data:
- Env
- Genotype

Y-variate: Yield

Genotypes: Genotype

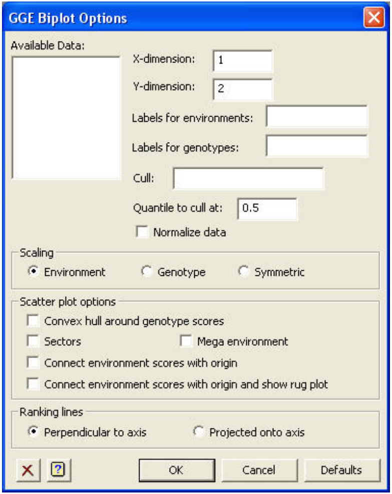
Environments: Env

Type of plot: Scatter plot

Method: ● Environment ○ Genotype

First environment:

Second environment:



GGE Biplot Options

Available Data:

X-dimension: 1

Y-dimension: 2

Labels for environments:

Labels for genotypes:

Cull:

Quantile to cull at: 0.5

□ Normalize data

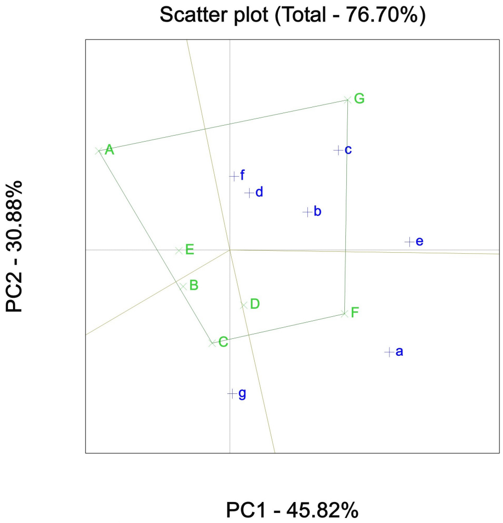
Scaling: ● Environment ○ Genotype ○ Symmetric

Scatter plot options
- □ Convex hull around genotype scores
- □ Sectors
- □ Mega environment
- □ Connect environment scores with origin
- □ Connect environment scores with origin and show rug plot

Ranking lines: ● Perpendicular to axis ○ Projected onto axis

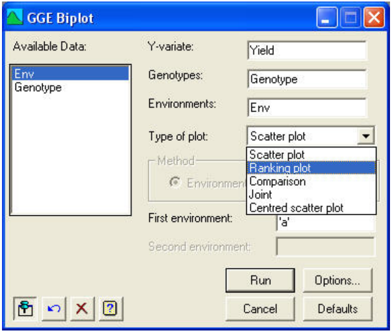
- Click on Options
- Select Convex Hull and Sectors

GGE biplot



GGE biplot - Ranking plot

Click on `Type of plot` drop-down menu on `GGE Biplot`



Y-variate: Yield

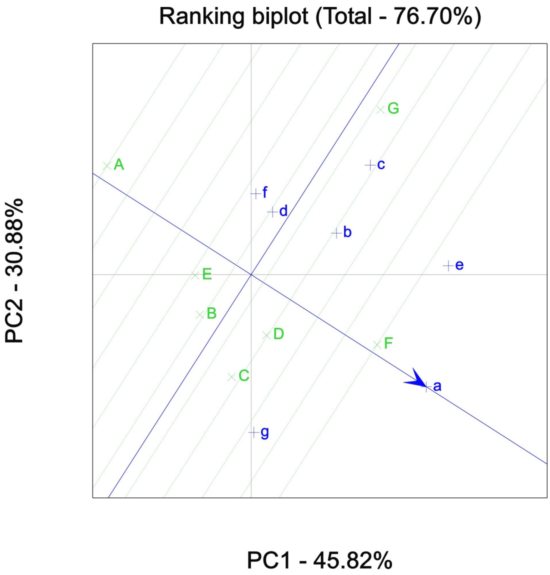
Genotypes: Genotype

Environments: Env

Type of plot: Scatter plot

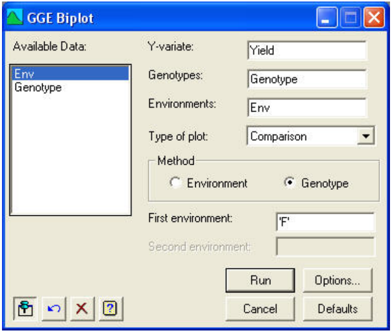
First environment: 'a'

GGE biplot - Ranking plot



GGE biplot - Comparison plot

Choose Comparison from Type of plot



GGE Biplot

Available Data:

Env
Genotype

Y-variate: Yield

Genotypes: Genotype

Environments: Env

Type of plot: Comparison

Method

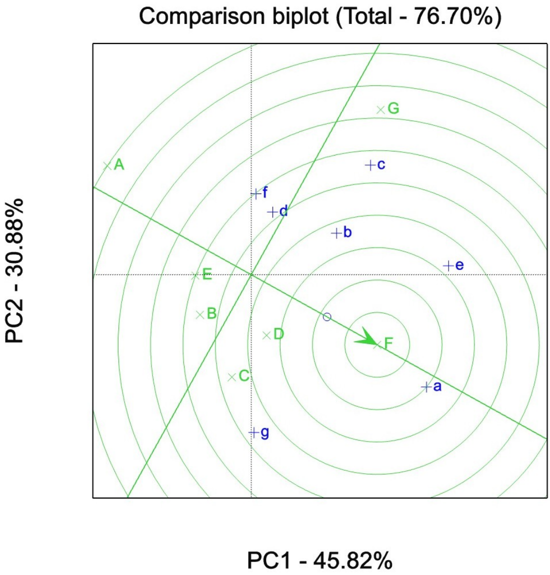
○ Environment ● Genotype

First environment: 'F'

Second environment:

- Select Genotype radio button in Method box
- Select Genotype to be used as base Genotype in First Environment

GGE biplot - Comparison plot



References

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

Gower, J.C. & Hand, D.J. (1996). *Biplots. Monographs on Statistics and Applied Probability 54*. Chapman & Hall, London.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Greenacre, M.J. (2007). *Correspondence Analysis in Practice, second edition*. Chapman & Hall, London.

Legendre, P. & Legendre, L. (1998). *Numerical Ecology, Second English Edition*. Elsevier, Amsterdam.

Yan, W. & Kang, M.S. (2003). *GGE Biplot Analysis: a Graphical Tool for Breeders, Geneticists and Agronomists*. CRC Press, Boca Raton.