Slide 1

# Biplots, revisited

**Ulrich Kohler**

kohler@wz-berlin.de

**Wissenschaftszentrum Berlin**

**June 24, 2004**

Slide 2

## Contents





Slide 3

## 1 Interpretation

Biplots show the following quantities of a data matrix in one display:

- the variance-covariance structure of the variables
- the values of observations on variables
- the euclidean distances between observations in the multidimensional space

They are helpful to reveal clustering, multicollinearity and multivariate outliers of a dataset, and they can be also used to guide the interpretation of principal component analyses (PCA).

Slide 4

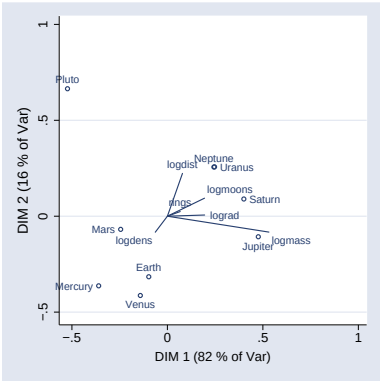


Figure 1: GH-Biplot of planets.dta (Hamilton, 1992, 268)

Slide 5

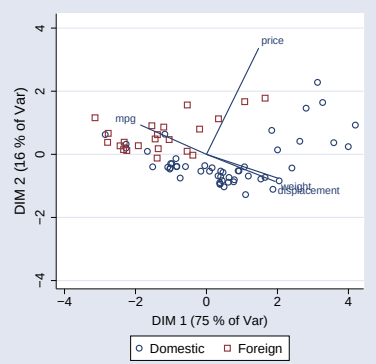


Figure 2: JK-Biplot of auto.dta

Slide 6

# 2 The Math

## Calculation of Coordinates

Let

$$\mathbf{Y} = \mathbf{ULV}' \quad ,$$

be the *singular value decomposition* (SVD) of the matrix $\mathbf{Y}$, which holds the data. From this SVD the coordinates of the observations and variables are calculated by

$$\mathbf{G} = \mathbf{UL}^c \quad \text{and} \tag{1}$$

$$\mathbf{H}' = \mathbf{L}^{1-c}\mathbf{V}' \quad , \tag{2}$$

whereby $c = (0, 1)$.



Slide 7

## Selection of Dimensions

The coordinates in $\mathbf{G}$ and $\mathbf{H}$ have $k$ dimensions. To plot these coordinates in a two dimensional space, one has to select two of them. Usually this is done by choosing the columns of $\mathbf{G}$ and $\mathbf{H}$, which correspondent to the highest Eigenvalues in $\mathbf{L}$.

Using less than $k$ dimensions lead to a loss of information, so that

$$\mathbf{GH}' = \mathbf{UL}^c\mathbf{L}^{1-c}\mathbf{V}' = \mathbf{ULV}' = \mathbf{Y} \tag{3}$$

will only hold approximately. To indicate the quality of the approximation, the default axis-titles of `biplot` mention the amount of explained variances by the selected dimensions. Unless the sum of these explained variances is sufficiently large, "the interpretation of the plot is suspect" (Jackson, 1991, 199)

Slide 8

## Biplot-Types

Choosing a value for $c$ define different types of biplots:

- $c = 0$, the GH-, or column-metric preserving biplot optimally approximates the variance-covariance-structure.
- $c = 1$, the JK-, or row-metric preserving biplot optimally approximates the euclidean distances.
- $c = .5$ the SQ-, or symmetric biplot, optimally approximates the observational values.

# 3 Computational Issues

## Maximum Numbers of Observations

Slide 9

The Stata-command to calculate a singular value decomposition

```
. matrix svd U L V = Y
```

requires that the dataset is stored in a matrix. This restricts the maximum number of observations to be used on 800 in Intercooled Stata and 11000 in Stata/SE.

For the JK-biplot ($c = 1$) the restriction can be circumvented, since $\mathbf{G}$ and $\mathbf{H}$ are equal to the scores and coefficients of a PCA. The JK-biplot is therefore calculated from a PCA, bypassing the SVD. Hence, there is no restriction for the maximum numbers of observations for the JK-biplot.

## Calculation of other Biplots without a SVD?

Slide 10

The JK-biplot can be transformed into the GH-biplot with

$$\mathbf{G}_{\mathrm{GH}} = \mathbf{G}_{\mathrm{JK}}\mathbf{L}^{-1} \quad (4)$$

$$\mathbf{H}'_{\mathrm{GH}} = \mathbf{L}\mathbf{H}'_{\mathrm{JK}} \quad . \quad (5)$$

However, the SVD of $\mathbf{Y}$ is needed to obtain $\mathbf{L}$.

The *Eigenvalues* ($\mathbf{L}_{\mathrm{JK}}$) of a PCA can be transformed into $\mathbf{L}$ with:

$$\mathbf{L} = \mathbf{U}'\mathbf{Y}_{\mathrm{S}}\mathbf{S}^{-1}\mathbf{U}_{\mathrm{S}}\mathbf{L}_{\mathrm{JK}} \quad , \quad (6)$$

where $\mathbf{S}$ is the covariance-matrix of the data, and $\mathbf{U}_{\mathrm{S}}$ are the PCA-coefficients. However this requires the SVD of $\mathbf{Y}$, to obtain $\mathbf{U}$



## The Practical View

Slide 11

It might be worthwhile for StataCorp to program the calculation of the Eigenvalues from the dataset, without storing the dataset in a matrix beforehand. In this case, at least the GH-biplot could be easily derived from a PCA with (4) and (5).

However, one should keep in mind that the interpretation of the biplot will be suspect, if the variance explained by the dimensions of the biplot are small. Small explained variances are common for datasets with many observations. In so far, the biplot has its strength mainly for datasets with small to moderate number of observations.

# 4 The biplot command

## Syntax

Slide 12

```
biplot varlist [weight] [if exp] [in range] [ ,
   [jk|sq|gh|mixed(jk|sq|gh jk|sq|gh)] covariance
   mahalanobis rv obsonly varonly dimensions(##)
   subpop(varname[, scatter_options ]) stretch(#)
   flip(x|y|xy) scatter_options line_options
   twoway_options ]
```

**Slide 13**

### Default-Setting

Invoking the command `biplot` with a *varlist* and no other options brings up a JK-biplot, which superimposes two of the most often described plots for principal component analysis: the component score plot and the plot of PCA-coefficients (loadings).

```
. biplot sepallen-petalwid
```

**Slide 14**

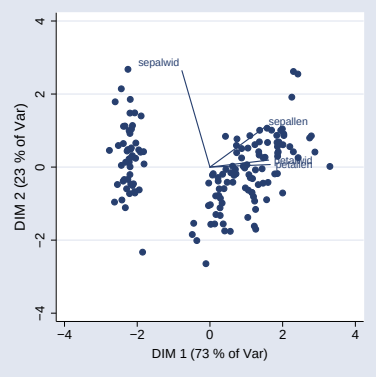


Figure 3: The standard JK-Biplot of iris.dta



**Slide 15**

### Biplots and common Plots for the PCA

It is possible to use `biplot` to produce the common PCA plots.

```
. biplot sepallen-petalwid, stretch(1) varonly

. biplot sepallen-petalwid, obsonly
```

Note: To interpret the square of the plotted PCA-coefficients, it is necessary to "stretch" the variable-lines to their original length.

**Slide 16**

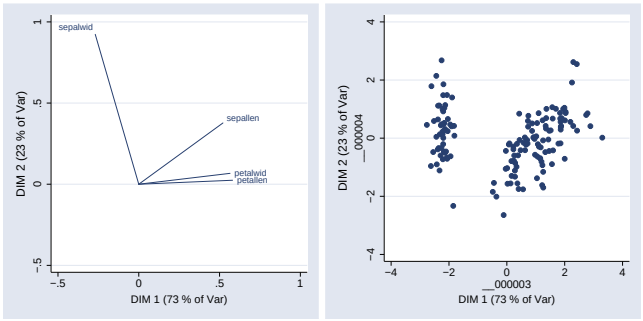


Figure 4: Plot of PCA-coefficients and Component Score Plot

**Slide 17**

## Controlling Dimensions

By default the coordinates which refer to the two highest Eigenvalues are selected for the plot. The option `dimensions(##)` allows to change this. This is useful for JK-biplots, since one might be interested in a display of the PCA-coefficients for arbitrary principal components. Moreover, the component score plot in the space of the two last principal components show a special kind of outlier (Gnanadesikan, 1977, 261).

```
. biplot sepallen-petalwid, dim(3 4)
```

**Slide 18**

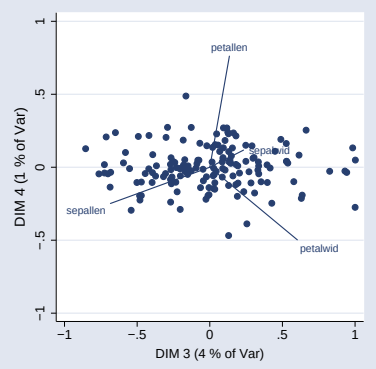


Figure 5: JK-Biplot in the space of the last two principal components



**Slide 19**

## Biplot Types

The JK, GH- and the SQ-biplot can be displayed by using the options `jk`, `gh` or `sq` respectively. It is possible in any case to calculate the coordinates from a standardized or a non-standardized data-matrix. Standardization is the default, which is why the variable-lines tend to have the same length. To get length for the variable-lines according to variances of the variables the option `covariance` has to be used.

```
. biplot sepallen-petalwid, gh cov
```

```
. biplot sepallen-petalwid, gh
```

**Slide 20**

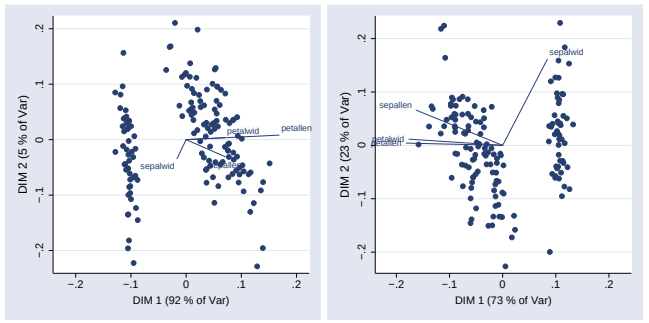


Figure 6: GH-Biplot for Unstandardized and Standardized Data

Slide 21

### The `mixed()`-option

Biplot-types differ in the quality of their approximations of the key-quantities shown in a biplot. It seems therefore straightforward to mix the different biplot-types. Gabriel (2002), for example, proposed a "correspondence analysis" which uses the coordinates of a GH-biplot for the variables and the coordinates of a JK-biplot for the observations. Such mixed biplots can be produced with the option `mixed()`.

```
. biplot sepallen-petalwid, mixed(jk gh)
```

Slide 22

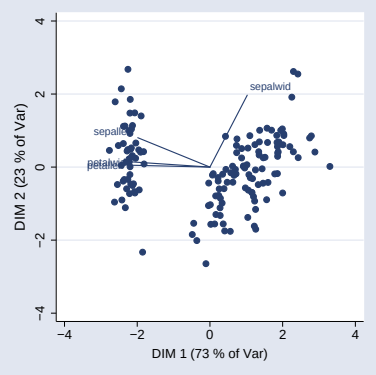


Figure 7: Gabriel's Correspondence Analysis



Slide 23

### Other variants

- Option `rv` for biplots for compositional data (Aitchison, 1990).
- Option `mahalanobis` rescales GH-biplot to reflect mahalanobis distances.

Slide 24

### Options to control the graph appearance

*scatter_options* allow up to two arguments, whereby the first argument refers to the observations (the dots) and the second refers to the points at the end of the variable-lines (which are invisible by default). The *line_options* refer to the variable lines.

The option `subpop()` is used to distinguish observations from different subgroups.

```
. biplot sepallen-petalwid, subpop(species, msymbol(Oh X Th)) legend(ring(0) po
> s(4))
```

Note: The *scatter_options* for the observations are ignored if you specify `subpop()`. However one can use the complete set of *scatter_options* as sub-option within `subpop()` to control the appearance of the observations.

**Slide 25**

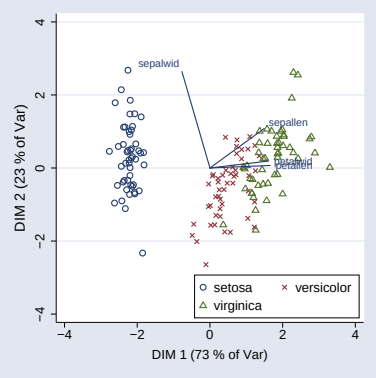


Figure 8: Illustrative example of representation options

## References

Aitchison, J. 1990. Relative Variation Diagrams for Describing Patterns of Compositional Variablity. *Mathematical Geology* 22: 487–512.

Blasius, J. and M. Greenacre. 1998. *Visualization of Categorical Data*. London: Academic Press.

Gabriel, K. 1971. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* 58(3): 453–467.

—. 2002. Goodness of Fit of Biplots and Correspondence Analysis. *Biometrica* 89(2): 423–436.

Gnanadesikan, R. 1977. *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Gower, J. and D. Hand. 1996. *Biplots*. London: Chapman and Hall.

Hamilton, L. C. 1992. *Regression with Graphics. A Second Course in Applied Statistics*. Belmont: Duxburry Press.



Jackson, J. E. 1991. *A User's Guide to Principal Components*. New York: Wiley.

Jolliffe, I. 2002. *Principal Components Analysis. 2nd. Edition*. New York and Heidelberg: Springer.