Data Inspection Using Biplots

Ulrich Kohler

kohler@wz-berlin.de

Wissenschaftszentrum Berlin

Plan of the presentation

- History
- Interpretation
- The math
- Biplot-Types
- Two more options

History

`biplot` has been available on SSC since Stata 5. After arrival of Stata 8 I have revisited `biplot` and made several changes (old version still works under version control).

- Use of the new graph engine
- Allowing for weights for JK-Biplots
- New option `rv` for “compositional data”
- New option `mahalanobis`
- New option `subpop()`
- Change of some default settings

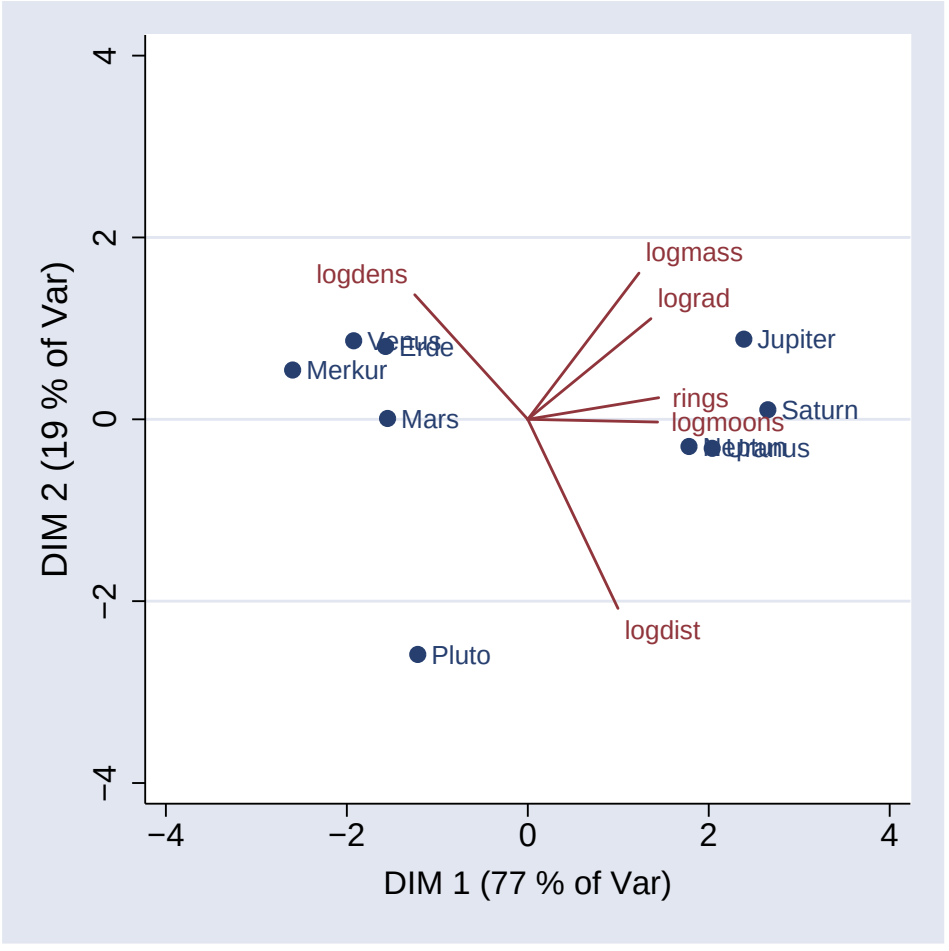
Interpretation

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

- standard deviations of variables
- correlations between variables
- values of observations on variables
- distances between observations in the multidimensional space

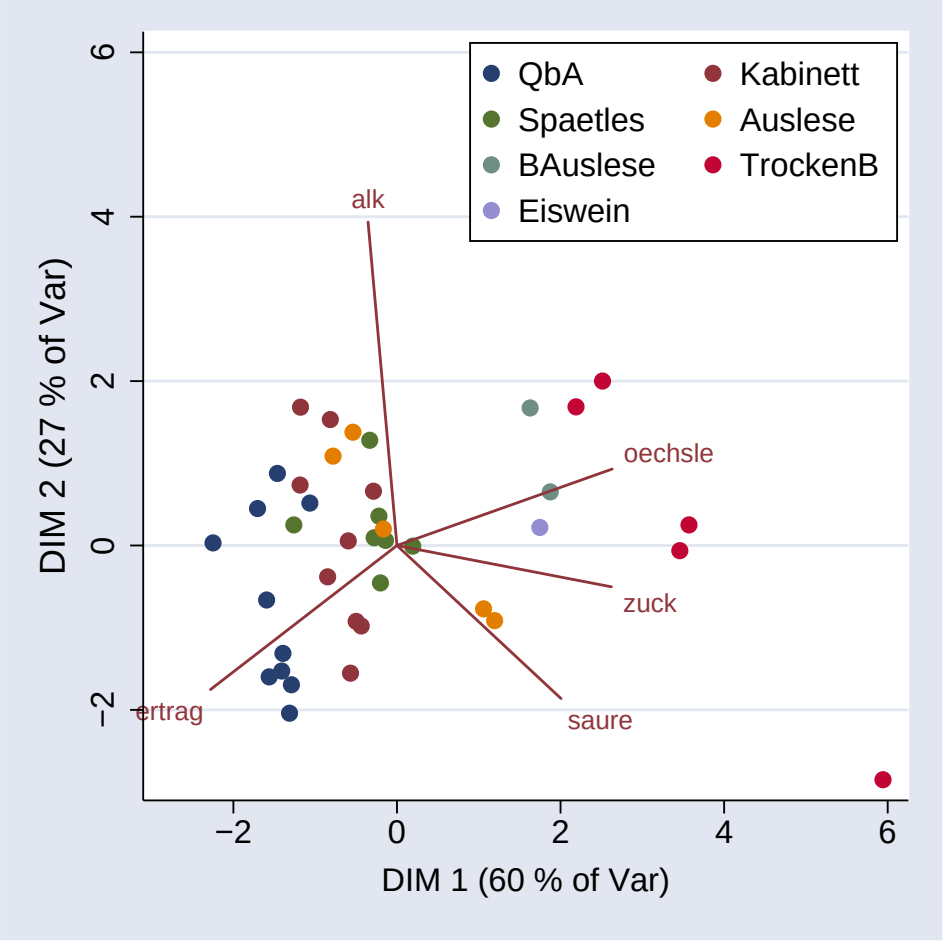
Interpretation

```
. biplot ring-logmoons, mlabel(planet)
```



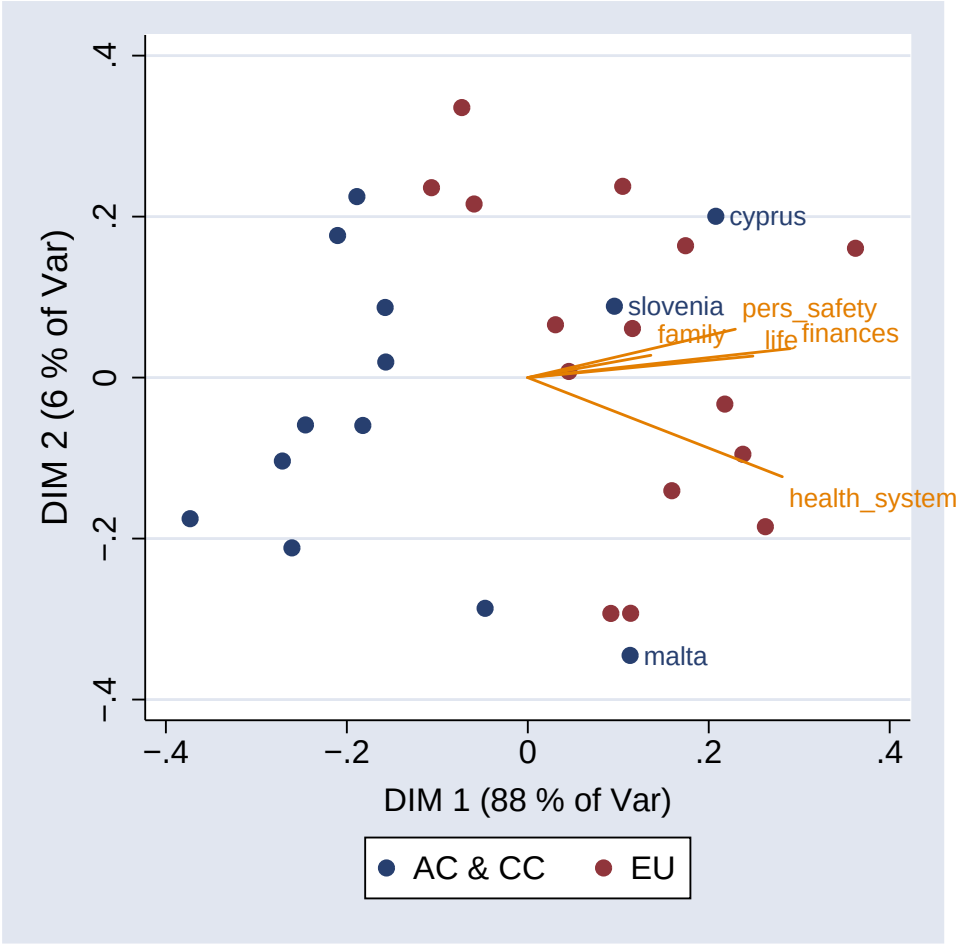
Interpretation

```
. biplot ..., subpop(praed) legend(ring(0) pos(1))
```



Interpretation

```
biplot ...  , gh cov subpop(eu, mlab(label))
```



The Math

Let $\mathbf{Y}$ be a $n \times k$ matrix holding the data. One can decompose $\mathbf{Y}$ with a *singular value decomposition* (SVD) into

$$\underset{n\times k}{\mathbf{Y}} = \underset{n\times k}{\mathbf{U}}\ \underset{k\times k}{\mathbf{L}}\ \underset{k\times k}{\mathbf{V}'} \tag{1}$$

where $\mathbf{L}$ contains the *Eigenvalues*.

From the SVD results the 2×2 matrix $\underline{\mathbf{L}}$ is formed, which contains the two elements of $\mathbf{L}$ with the highest Eigenvalues. The $n \times 2$ matrix $\underline{\mathbf{U}}$ and the $k \times 2$ matrix $\underline{\mathbf{V}}$ are formed by choosing those columns from $\mathbf{V}$ and $\mathbf{U}$ which correspondent to the highest Eigenvalues.

The Math

The coordinates for the observations are given by

$$\underset{n\times 2}{\mathbf{G}} = \underline{\mathbf{U}}\,\underline{\mathbf{L}}^{c} \tag{2}$$

and the coordinates for the variables are given by

$$\underset{2\times k}{\mathbf{H}}' = \underline{\mathbf{L}}^{(1-c)}\underline{\mathbf{V}}' \tag{3}$$

Biplot-Types are defined by choosing the value for c.

Biplot-Types

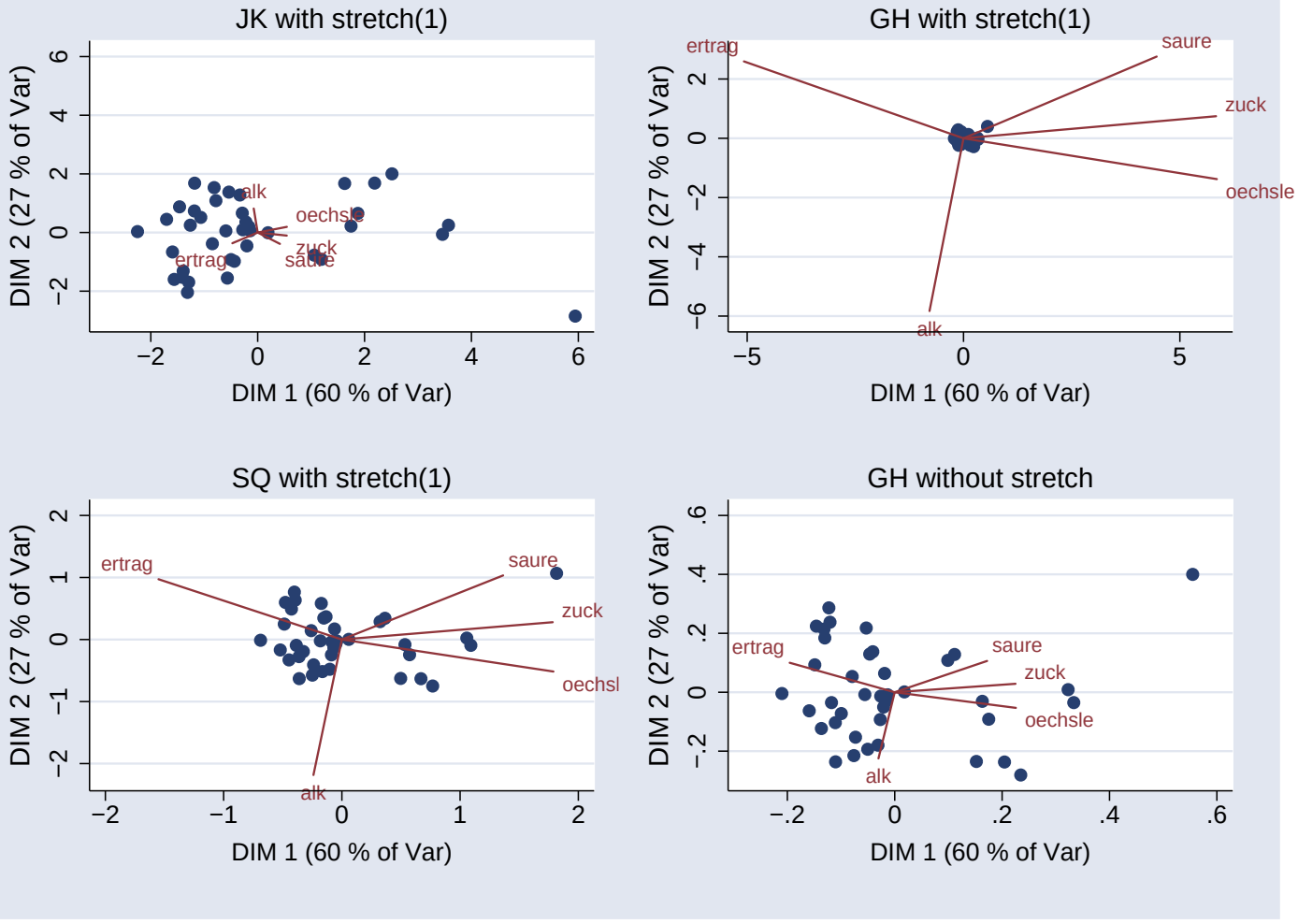
- GH-Biplot: $c = 0$
- JK-Biplot: $c = 1$
- SQ-Biplot: $c = .5$

Note: For $c = 1$ the coordinates for the observations correspond to the first two principal components, and the coordinates for the variables correspond to the first two *Eigenvectors*. Therefore `biplot` calculates a PCA to produce the JK-Biplot.

Biplot-Types

SQ-Biplots are sometimes called *symmetric biplots*. In this type the coordinates of variables and observations tend to be more similar than in the two other types. Regardless of the Biplot-Type, `biplot` automatically chooses a stretch factor for the variable-coordinates making SQ biplots more or less unnecessary.

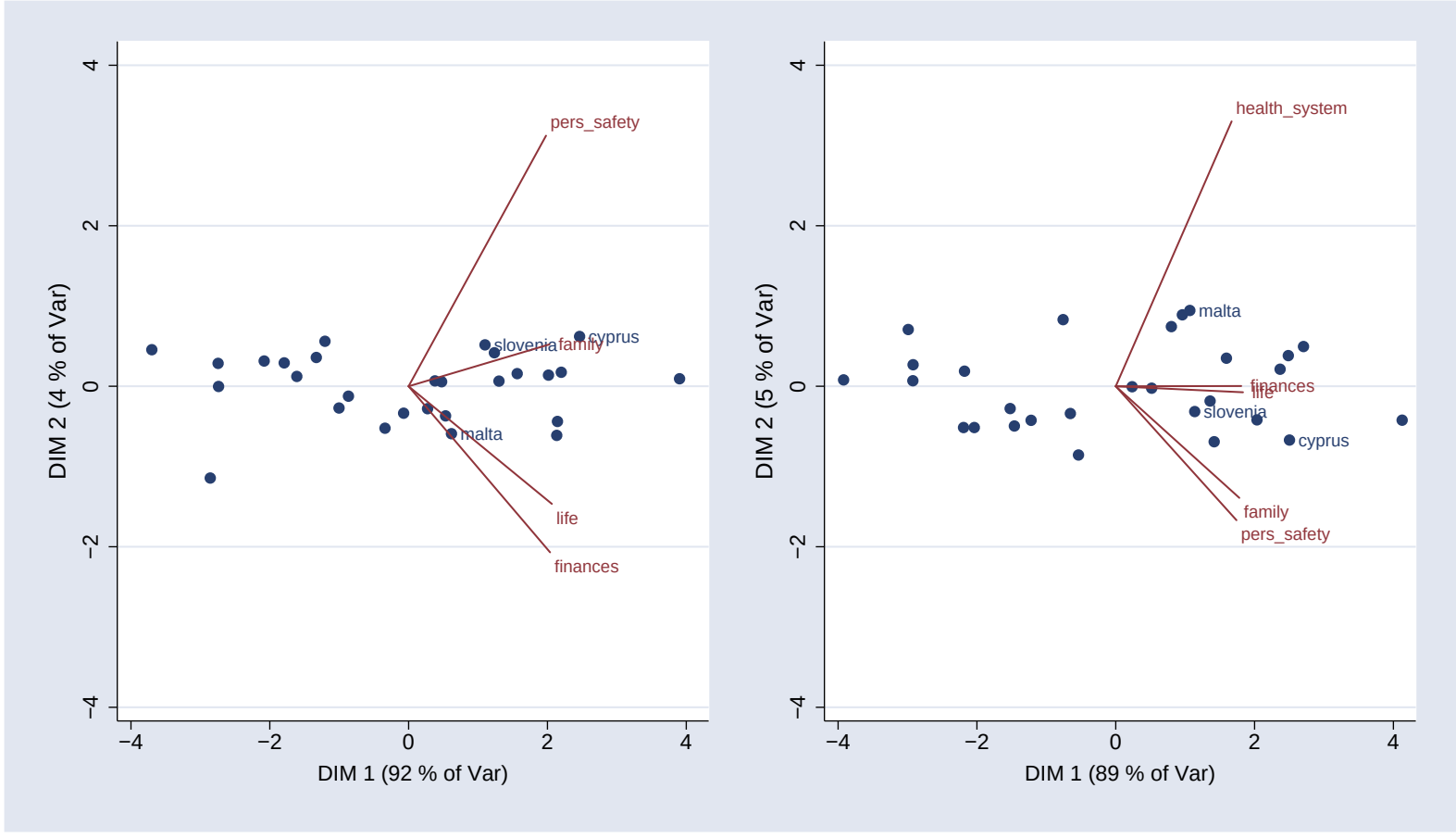
Biplot-Types



Biplot-Types

JK-Biplots are *row metric preserving*, that is, the distances between the objects are more closely approximated in the JK-Biplot than in the other types.

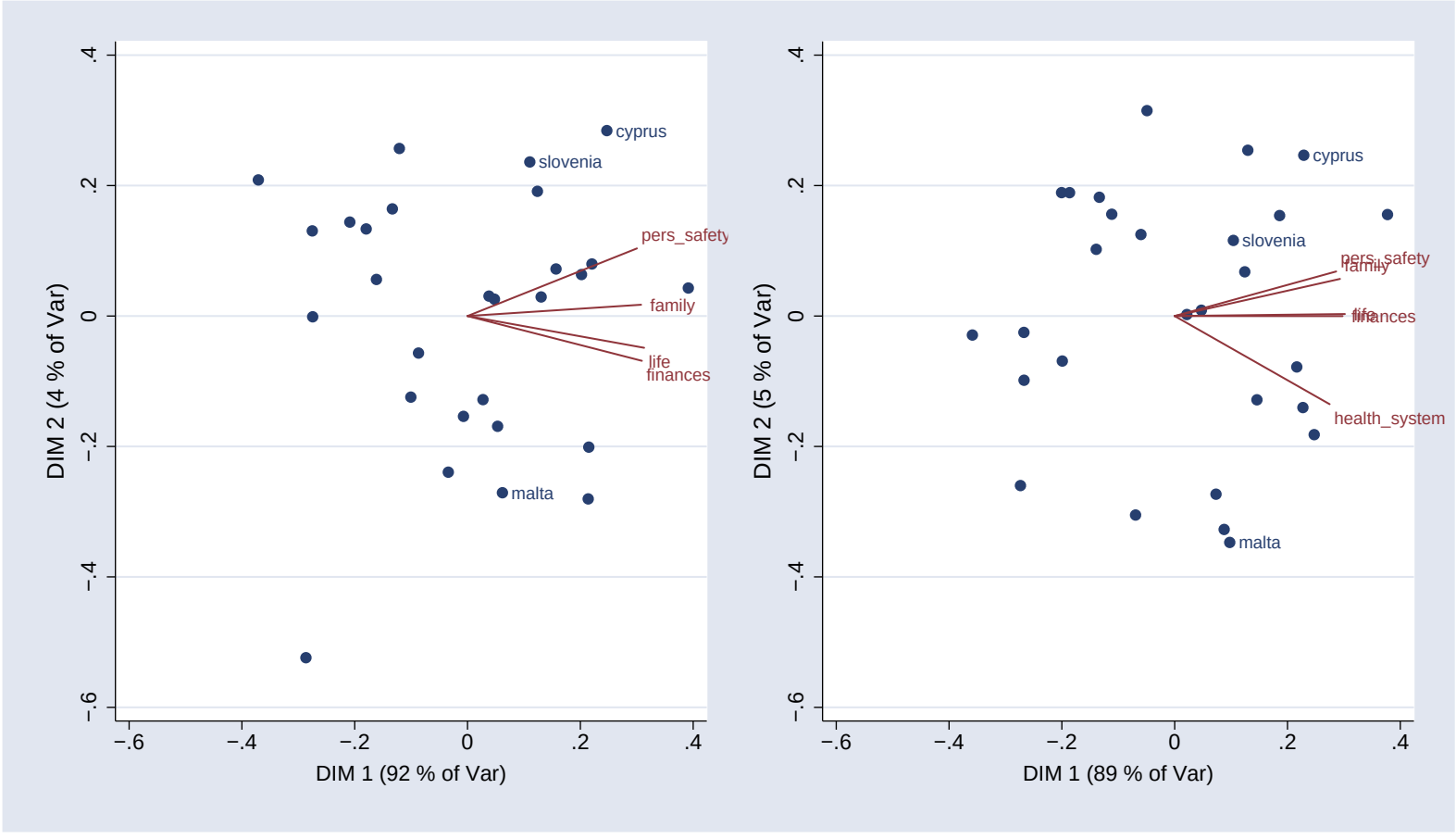
Biplot-Types



Biplot-Types

GH-Biplots are *column metric preserving*, that is, the correlations between the variables are more closely approximated in the GH-Biplot than in the other types.

Biplot-Types



Two new Options

- `rv` is used to produce relative variation diagrams. Relative variation diagrams are Biplots for compositional data and compositional data are data sets with constant row-sums and only positive value (like, for example the row percentages of two-way frequency tables). To get a relative variation diagram the data matrix needs to be transformed before producing the Biplot,. `biplot` does this transformation for you if you specify `rv`.
- `mahalanobis` can be used for GH-Biplots to rescale the graph in a way that the distances between the observations approximates the Mahalnobis distances.