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
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
. biplot ring-logmoons, mlabel(planet)
Interpretation

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

![Biplot diagram](image)

QbA, Kabinett, Spaetles, Auslese, BAuslese, TrockenB, Eiswein

DIM 2 (27% of Var) vs. DIM 1 (60% of Var)
Interpretation

biplot ... , gh cov subpop(eu, mlab(label))
Let $Y$ be a $n \times k$ matrix holding the data. One can decompose $Y$ with a *singular value decomposition* (SVD) into

$$
Y = U L V' 
\quad (1)
$$

where $L$ contains the *Eigenvalues*. From the SVD results the $2 \times 2$ matrix $L$ is formed, which contains the two elements of $L$ with the highest Eigenvalues. The $n \times 2$ matrix $U$ and the $k \times 2$ matrix $V$ are formed by choosing those columns from $V$ and $U$ which correspond to the highest Eigenvalues.
The Math

The coordinates for the observations are given by

\[ G_{n \times 2} = U L^c \]  \hspace{1cm} (2)

and the coordinates for the variables are given by

\[ H'_{2 \times k} = L^{(1-c)} V' \]  \hspace{1cm} (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 \texttt{biplot} calculates a PCA to produce the JK-Biplot.
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, \texttt{biplot} automatically chooses a stretch factor for the variable-coordinates making SQ biplots more or less unnecessary.
Biplot-Types

JK with stretch(1)

GH with stretch(1)

SQ with stretch(1)

GH without stretch
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
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.
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 Mahalanobis distances.