# Vignette: BiplotGUI

**Anthony la Grange**

**Abstract**

Biplots simultaneously provide information on both the samples and the variables of a data matrix in two- or three-dimensional representations. The **BiplotGUI** package provides a graphical user interface for the construction of, interaction with, and manipulation of biplots in R. The samples are represented as points, with coordinates determined either by the choice of biplot, principal coordinate analysis or multidimensional scaling. Various transformations and dissimilarity metrics are available. Information on the original variables is incorporated by linear or non-linear calibrated axes. Goodness-of-fit measures are provided. Additional descriptors can be superimposed, including convex hulls, alpha-bags, point densities and classification regions. Amongst the interactive features are dynamic variable value prediction, zooming and point and axis drag-and-drop. Output can easily be exported to the R workspace for further manipulation. Three-dimensional biplots are incorporated via the **rgl** package. The user requires almost no knowledge of R syntax.

*Keywords*: alpha-bag, biplot, circular non-linear, canonical variate analysis, graphical user interface, multidimensional scaling, principal component analysis, principal coordinate analysis, Procrustes, R, Tcl/Tk.

## 1. Introduction

In this first section, a brief overview of biplots, existing biplot software, as well as the statistical programming language and environment R, is given. The main aims of the **BiplotGUI** package are set out in Section 2, while its most important features are showcased in Sections 3 and 4 through the exploration of three data sets. Further features are illustrated in Section 5. In Section 6, some ideas for future releases are listed. The present version of the package is 0.0-6.

The style of the vignette[1] is intentionally non-mathematical. This allows the focus to lie firmly with the package and its features. However, detailed references are provided for those who wish to gain a fuller understanding of the underlying theory. In addition, the main computations performed by the package are set out in Appendix A.

### 1.1. Biplots

Introduced by Gabriel (1971), the biplot is described by Gower and Hand (1996) in their authoritative monograph as the multivariate analogue of the ordinary scatter plot. As such, biplots are representations of multivariate data in which information on both the samples (observations) and the variables of a data matrix is given simultaneously in two or three

[1] This document is an update of La Grange *et al.* (2009); last updated: 14 February 2010.

dimensions: the samples are represented as points, while the variables are represented as labelled, calibrated axes. The axes are either linear and oblique, or non-linear. This new approach to biplots differs from the more traditional approach in which samples and variables are represented as points and/or uncalibrated vectors.

Some dimension-reduction technique is typically used to represent the samples as points, often principal component analysis (PCA) (Pearson 1901; Hotelling 1933) or canonical variate analysis (CVA) (Hotelling 1935, 1936). More generally, scaling techniques such as principal coordinate analysis (PCO) (Torgerson 1952; Gower 1966) or metric or non-metric multidimensional scaling (MDS) (Kruskal 1964a,b; Sammon 1969) are used. Jolliffe (2002) dedicates a monograph to PCA, while Krzanowski (2000) covers general multivariate topics. Cox and Cox (2001) and Borg and Groenen (2005) are standard references for scaling techniques.

The placement of the axes depends partly on the mechanism used in the placement of the points. The PCA biplot provides linear axes for points placed by PCA (Gower and Hand 1996, Chapter 2); similarly the CVA biplot provides linear axes for points placed by CVA (Gabriel 1972; Gower and Hand 1996, Chapter 5). The regression biplot (Gower and Hand 1996, Chapter 3) gives approximate linear axes for any ordination of points. So too does the Procrustes biplot (Gower and Hand 1996, Chapter 3). The regression and Procrustes biplots correspond to the PCA biplot for points determined by PCO based on Pythagorean dissimilarities. The covariance biplot (Greenacre 1984; Underhill 1990) adjusts the points and axes of the PCA biplot so that the cosines of the angles between the axes approximate the correlations between the corresponding variables. The correlation biplot is similar, except that the variables are first scaled to have unit variances.

The placement of the axes may also depend on how they are to be used. *Predictive* axes are positioned and calibrated so that the orthogonal projection of a point onto an axis ‘predicts’ as best as is graphically possible the value of the corresponding sample on the corresponding variable. *Interpolative* axes, on the other hand, are positioned and calibrated so that a new sample may be added to an existing configuration of points at the most appropriate position graphically possible. Interpolation can be either by the centroid or vector sum of the positions on the axes corresponding to the respective variable values of the new sample.

For additive inter-sample dissimilarities (Gower and Hand 1996, p. 105), biplots with non-linear axes (or trajectories) may be constructed for points determined by PCO. The PCO solution itself requires the inter-sample dissimilarities to be Euclidean-embeddable (Gower 1982); dissimilarity measures for which this is the case are discussed by Gower and Legendre (1986). Non-linear predictive axes may make use of circular projection (Gower and Hand 1996, Chapter 6), while non-linear interpolative axes (Gower and Harding 1988; Gower and Hand 1996, Chapter 6) are used in the same way as the linear variety. Non-linear biplots are often most useful to gauge what is otherwise approximated by linear biplots.

While very many examples of biplots of the traditional approach may be found in the literature, there are fewer examples of biplots of the new approach. An important reason has been the lack of software, as is discussed in Sections 1.2 and 1.3. The value of biplots of the new approach, however, has often been demonstrated. In an easy to read introduction, for example, Le Roux and Gardner (2005) cite and showcase many examples of the uses of linear biplots, from such diverse fields as archaeology, agriculture, antiques, education, financial management, mineralogy and process control. Other recent fields of application include cephalometry (Naidoo *et al.* 2006), chemistry (Alves *et al.* 2005) and mineralogy (Jemwa and

Aldrich 2006). Examples of non-linear biplots may be found in Gower and Harding (1988), Gower and Hand (1996), Cox and Cox (2001) and Gower and Ngouenet (2005).

Given the ubiquity of multivariate data and the usefulness of biplots in describing such data, there is still much scope for the further popularisation of the technique.

## 1.2. Existing biplot software

Many statistical packages can be used to produce at least the simplest of biplots of the traditional approach. These include the major statistical packages Minitab (Minitab Inc 2007), SPSS (SPSS Inc 2008), Stata (StataCorp LP 2007) and various products from SAS (SAS Institute Inc 2009). However, functionality is often limited, and the results hard to obtain. Greater functionality is provided by the three dedicated biplot programs **XLS-Biplot** (Udina 2005a,b), **GGEBiplot** (Yan and Kang 2006) and **BiPlot** (Lipkovich and Smith 2002a,b). **XLS-Biplot** is based on XLisp-Stat (Tierney 1990) and has many useful features including a related web-server that can be used to construct biplots online. **GGEbiplot** is aimed mainly at agronomists, crop scientists and geneticists. It supplements the book by Yan and Kang (2003). **BiPlot** is an add-on for **Excel**, and although therefore potentially widely useful, it unfortunately has some minor but serious shortcomings (see Udina 2005b).

The Genstat package (VSN International Ltd 2008) can be used to calculate the coordinates of the elements of a biplot. These can then be drawn using a procedure from an add-on library. Other packages, offering some traditional biplot functionality, include **Manet** (Hofmann 2000), for Macintosh only, and **ViSta** (Young 2001). Some packages are aimed at ecologists—**brodgar** (Highland Statistics Ltd 2008) with R, **Canoco** (Plant Research International 2002) with **CanoDraw** (Smilauer 2003), **MVSP** (Kovach Computing Services 2008) and **PC-ORD** (MjM Software Design 2009)—while the **Excel** add-on **BrandMap** (WRC Research Systems Inc 2007) is aimed at marketers.

**STATISTICA** (StatSoft Inc 2009) is a mainstream statistical package that is capable of producing calibrated new-approach biplots, albeit the PCA biplot only. All the software mentioned are for purchase, except **XLS-Biplot**, **BiPlot**, **Manet** and **ViSta** which are available free of charge. So too is R.

## 1.3. R

R (R Development Core Team 2009) is a free statistical programming language and environment capable of producing high-quality graphics. Initiated by Ihaka and Gentleman (1996), it has become ‘the *de facto* standard for statistical computing’ (Greenacre 2007, p. 213). It is an open-source implementation of the S programming language, available for download for all the major platforms from the R Project home page at `http://www.R-project.org/`. The R core is updated regularly with minor version revisions released roughly every six months. The current version (as of December 2009) is R 2.10.1. Updates are relatively painless. R is easily extensible: a large number of user-written packages is available for download from repositories such as the Comprehensive R Archive Network (CRAN) and BioConductor. These repositories can be accessed via the R Project home page (see also Appendix B). As R has increased in popularity, so too has the number of books devoted to it. Recent general-topic books on R include Braun and Murdoch (2007), Chambers (2007) and Spector (2008). The book by Murrel (2005) deals specifically with graphics in R. Many more resources are freely available from the R Project home page.

As far as biplots are concerned, the `biplot` method in R can be used to produce two variations of Gabriel's (1971) classical biplot. The classical biplot is most similar to the covariance/correlation biplot described earlier. Packages with support for traditional biplots include **ade4** (Dray and Dufour 2007, 2009), **ade4TkGUI** (Thioulouse and Dray 2009, 2007), **bpca** (Faria and Demetrio 2008) and **vegan** (Oksanen *et al.* 2009). In addition, the **calibrate** package (Graffelman 2009) can be used to calibrate both scatter plot and linear biplot axes as described by Graffelman and van Eeuwijk (2005). These calibrations correspond to those of Gower and Hand (1996).[2]

## 2. A new package

The primary aim with the **BiplotGUI** package is to make it easy to construct biplots of the kind advocated by Gower and Hand (1996) – biplots in which samples are represented as points and variables are represented as calibrated axes. The package goes beyond this, however, allowing users to interact with the data through the biplots which are produced. Naturally, the graphical output should be of a high quality, and the graphs easily customisable. Its characteristics make R the ideal environment for the development of such a package.

In the next two sections, the most important features of the **BiplotGUI** package are illustrated. This is done through the exploration of three data sets. Further features are highlighted in Section 5. A systematic account of all features is given in the Features Manual. This manual can be accessed via the `Help` menu from within the package. The package does not currently support biplots of categorical variables. Further resources and tools are available via the package home page at `http://biplotgui.R-Forge.R-project.org/`.

## 3. A first example

In this section a country-comparative data set is introduced. It is used to show how the graphical user interace (GUI) of the package may be initialised, how its features are laid out, and how it may be used to explore multivariate data using, amongst other things, PCA and regression biplots.

### 3.1. The country data

Section 1 shows eight variables measured on the countries with the 15 largest economies (by purchasing price parity (PPP) gross domestic product (GDP)) in 2007. These data have been derived largely from the 2007 CIA World Factbook (Agency 2007) and are for illustrative purposes only. The variables are: PPP GDP per capita in US dollars (GDP); HIV/Aids prevalence as a percentage of the population (HIV.Aids); life expectancy in years (Life exp.); military spending as a percentage of GDP (Mil.); oil consumption in barrels per annum per capita (Oil cons.); population in millions (Pop.); number of fixed line telephones per 1 000 people (Tel.); and percentage unemployed (Unempl.). The aim is to represent these data in

[2] In La Grange *et al.* (2009) and in previous versions of this vignette, it was stated that the biplot calibrations of Graffelman and van Eeuwijk (2005) do not in general correspond to those of Gower and Hand (1996). In the case of predictive linear biplots, the calibrations do in fact correspond. In addition, corresponding interpolative linear biplots can also be constructed. I apologize for this error.

Table 1: The country data. Eight variables measured on the countries with the 15 largest economies (PPP GDP) in 2007; countries listed in alphabetical order.

| Country | GDP | HIV.Aids | Life exp. | Mil. | Oil cons. | Pop. | Tel. | Unempl. |
|---|---|---|---|---|---|---|---|---|
| Brazil | 8 710 | 0.7 | 72.2 | 2.6 | 4.0 | 190 | 204.2 | 9.6 |
| Canada | 35 370 | 0.3 | 80.3 | 1.1 | 25.1 | 33 | 622.3 | 6.4 |
| China | 7 724 | 0.1 | 72.9 | 4.3 | 1.8 | 1 322 | 278.4 | 4.2 |
| France | 29 852 | 0.4 | 80.6 | 2.6 | 11.3 | 64 | 543.5 | 8.7 |
| Germany | 31 941 | 0.1 | 79.0 | 1.5 | 11.7 | 82 | 657.8 | 7.1 |
| India | 3 685 | 0.9 | 68.6 | 2.5 | 0.8 | 1 130 | 44.0 | 7.8 |
| Indonesia | 4 041 | 0.1 | 70.2 | 3.0 | 1.8 | 235 | 63.2 | 12.5 |
| Italy | 30 199 | 0.5 | 79.9 | 1.8 | 11.8 | 58 | 430.8 | 7.0 |
| Japan | 33 100 | 0.1 | 82.0 | 0.8 | 16.0 | 127 | 432.8 | 4.1 |
| Mexico | 10 570 | 0.3 | 75.6 | 0.5 | 6.6 | 109 | 182.7 | 3.2 |
| Russia | 12 350 | 1.1 | 65.9 | 2.7 | 6.5 | 141 | 283.6 | 6.6 |
| S Korea | 24 386 | 0.1 | 77.2 | 2.7 | 16.0 | 49 | 547.8 | 3.3 |
| Spain | 27 418 | 0.7 | 79.8 | 1.2 | 14.2 | 40 | 454.5 | 8.1 |
| UK | 31 723 | 0.2 | 78.7 | 2.4 | 11.0 | 61 | 552.9 | 2.9 |
| USA | 43 369 | 0.6 | 78.0 | 4.1 | 25.1 | 301 | 571.2 | 4.8 |

two or three dimensions so that a single, multivariate visual impression may be obtained, with the calibrated biplot axes incorporating information on the original variables.

### 3.2. Getting started

After R has been downloaded and installed, it is also necessary to install the **BiplotGUI** package and its dependencies (details are provided in Appendix B). This process needs to be performed only once. To then load the **BiplotGUI** package into R, the following command is entered at the R prompt, followed as usual by the enter key:

```
R> library("BiplotGUI")
```

If the user is acquainted with R, data may be entered at the keyboard or be imported into R and saved as a matrix or a data frame. The country data have already been included in the package as a data frame, and may be viewed from within R by typing the commands

```
R> data("Countries")
R> Countries
```

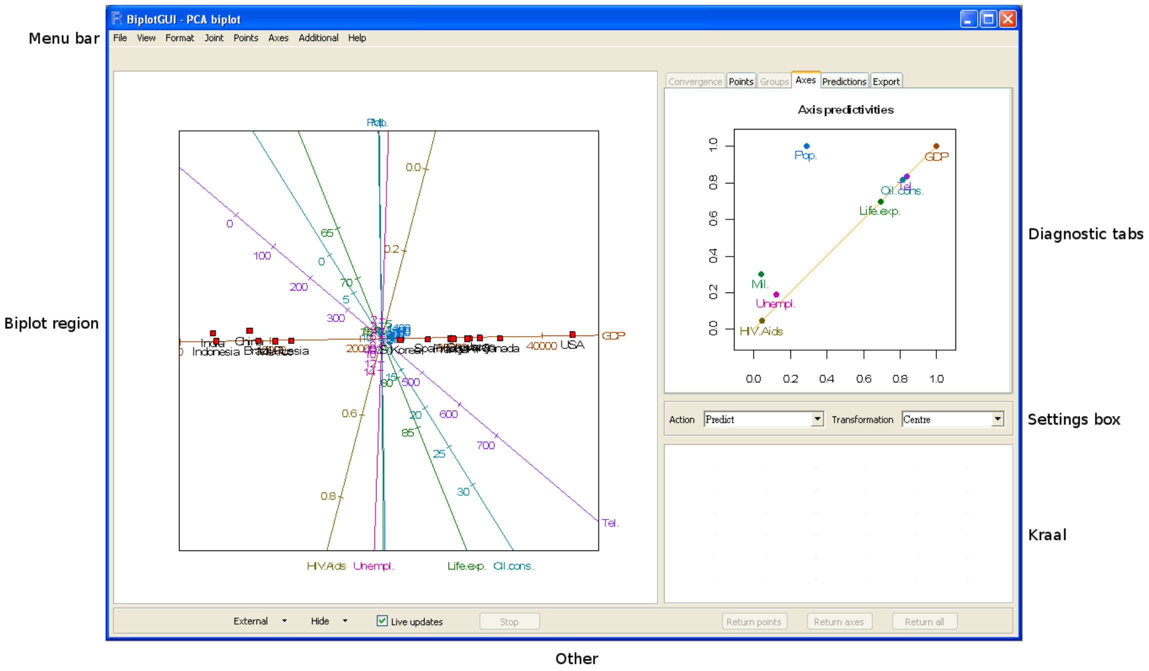
at the R prompt. To initialise the GUI with the country data, the command

```
R> Biplots(Data = Countries)
```

is entered. No further R commands are needed.

### 3.3. The layout

Figure 1 shows the layout of the GUI after it has launched. Six regions are indicated:



Action Predict

Transformation Centre

☑ Live updates

Figure 1: A screenshot of the **BiplotGUI** window as it initially appears. A predictive PCA biplot of the country data is shown towards the left. The axis predictivities are shown top right.

- **The menu bar**, in addition to the settings box, contains the most important options. The menu bar options are laid out in full in Figure 2. The three most important drop-down menus are `Joint`, `Points` and `Axes`. The biplots listed under `Joint` have both their points and axes determined according to a single, joint mechanism. The other biplots have their points determined from the `Points` menu and their axes determined from the `Axes` menu.

- **The biplot region** is where the biplot and optional title and legend are displayed. This space is responsive to mouse clicks and motion.

- **The settings box** may be used to set the action of the biplot axes, either predictive, centroid interpolative or vector sum interpolative. Various data transformations may be effected.

- **The diagnostics tabs** show output related to the currently displayed biplot. The `Convergence` tab shows a graph of convergence; the `Points`, `Groups` and `Axes` tabs show context-specific graphs of goodness-of-fit for points, groups and axes, respectively; the `Predictions` tab shows dynamically predicted variable values; while the `Export` tab allows various objects to be exported to R.

- **The kraal** is where points and axes may be kept, temporarily removing them from consideration.

- **Other**. The options in this section can be used to show the currently displayed biplot in an external window (in two or three dimensions), to control the biplot region or to control the kraal. While the GUI is busy, a progress bar is shown towards the left of this area.

The `Show pop-up help` option in the `Help` menu activates pop-up help messages which appear when the mouse cursor is hovered over the components of the main GUI window.

## 3.4. Exploring

The PCA biplot with predictive axes is shown by default. For the country data, this is the biplot shown towards the left of the screenshot in Figure 1. As should be the case for all biplots, a unit aspect ratio is used to ensure that distances within the biplot are properly represented. In this biplot, the points representing the countries lie ordered along a virtually straight line. In fact, the imagined line corresponds very closely to the biplot axis for GDP, and importantly, the line is almost horizontal. The reason for this becomes clear by looking at the GDP column of the country data set. The values of GDP are orders larger than those of the other variables. Therefore, that linear combination of the variables that has the largest possible variation (the first principal component) is heavily weighted towards GDP. In effect, GDP drowns out the other variables. To avoid this, we choose the `Centre, scale` transformation from the available transformations in the settings box. This transformation independently transforms each variable to unit variance and automatically updates the biplot to the one shown in the top left panel of Figure 3 (see Appendix A.2 for details on the calculations involved). Irrespective of the chosen transformation, however, the axes are always calibrated in terms of the original variable values. The first principal component in this new figure still ranks the countries from least to most wealthy, in some more complicated sense. The developed countries of the West, together with Japan and newly-industrialised South Korea, cluster in the south-east quadrant. Brazil, Russia and Indonesia lie more towards the west, with Mexico straddling the divide. India, and especially China, lie further away.

While the relative positions of the points are interesting, biplots come into their own when the points are related to their original variable values through the axes. By right clicking inside the predictive linear biplot and selecting `Predict cursor positions` from the pop-up menu, an array of orthogonally projecting lines emanates from, and follows, the cursor as it moves over the biplot. If `Predict points closest to cursor positions` is selected instead, the lines project from the point closest to the cursor as it moves, rather than from the cursor itself. So for example, the image in the top right panel of Figure 3 was created by hovering the cursor closer to the point for China than to any other point. These orthogonally projecting lines intersect the axes at the positions at which the optimal approximations to the original variables values are to be read off. It can be seen from the image that China scores relatively low on all the variables except population and military spending. As the cursor moves, these predictions are also given numerically, in real time, in the `Predictions` tab. Dynamic prediction is disabled by right clicking inside the biplot and selecting `Don't predict` from the pop-up menu. (Some numerical predictions are given below.) Notice that although the predictions are optimal, they remain approximations. An unfortunate consequence is that values close to zero on variables measured on the ratio scale can have negative predicted values. For example, in Figure 3 the population of Spain is predicted to be less than zero.

With many variables, a biplot may become crowded. A particular axis can be highlighted by right clicking it and then selecting `Highlight` from the pop-up menu. Doing so greys the other

**File**

- Save as
    - PDF...
    - Postscript...
    - Metafile...
    - Bmp...
    - Png...
    - Jpeg
        - 50% quality...
        - 75% quality...
        - 100% quality...
- Copy
- Print...
- Options. . .
- Exit

**View**

- Show title
- Clip around points
- Clip around points and axes
- Show point labels
- Show point values
- Show group labels in legend
- Don't show axis labels
- Show clinging axis labels
- Show axis labels in legend
- Show Additional labels in legend
- Show next legend entries
- Show previous legend entries
- Calibrate display space axes

**Format**

- Title. . .
- By groups. . .
- Axes. . .
- Interaction. . .
- Diagnostic tabs. . .
- Reset all. . .

**Joint**

- PCA
- Covariance/Correlation
- CVA

**Points**

- Dissimilarity metric
    - Pythagoras
    - Square-root-of-Manhattan
    - Clark
    - Mahalanobis
- PCO
- MDS
    - Run
    - Identity transformation
    - Monotone regression
    - Monotone spline transformation. . .
    - Primary approach to ties
    - Secondary approach to ties
    - Random initial configuration
    - In terms of principal axes

**Axes**

- None
- Regression
- Procrustes
- Circular non-linear
- Default

**Additional**

- Interpolate
    - A new sample. . .
    - Sample group means. . .
- Convex hulls. . .
- Alpha-bags. . .
- Point densities. . .
- Classification regions. . .
- Clear all

**Help**

- Vignette (in PDF)
- Features Manual (in PDF)
- Home page
- Show pop-up help
- About. . .

Figure 2: The menu bar options.

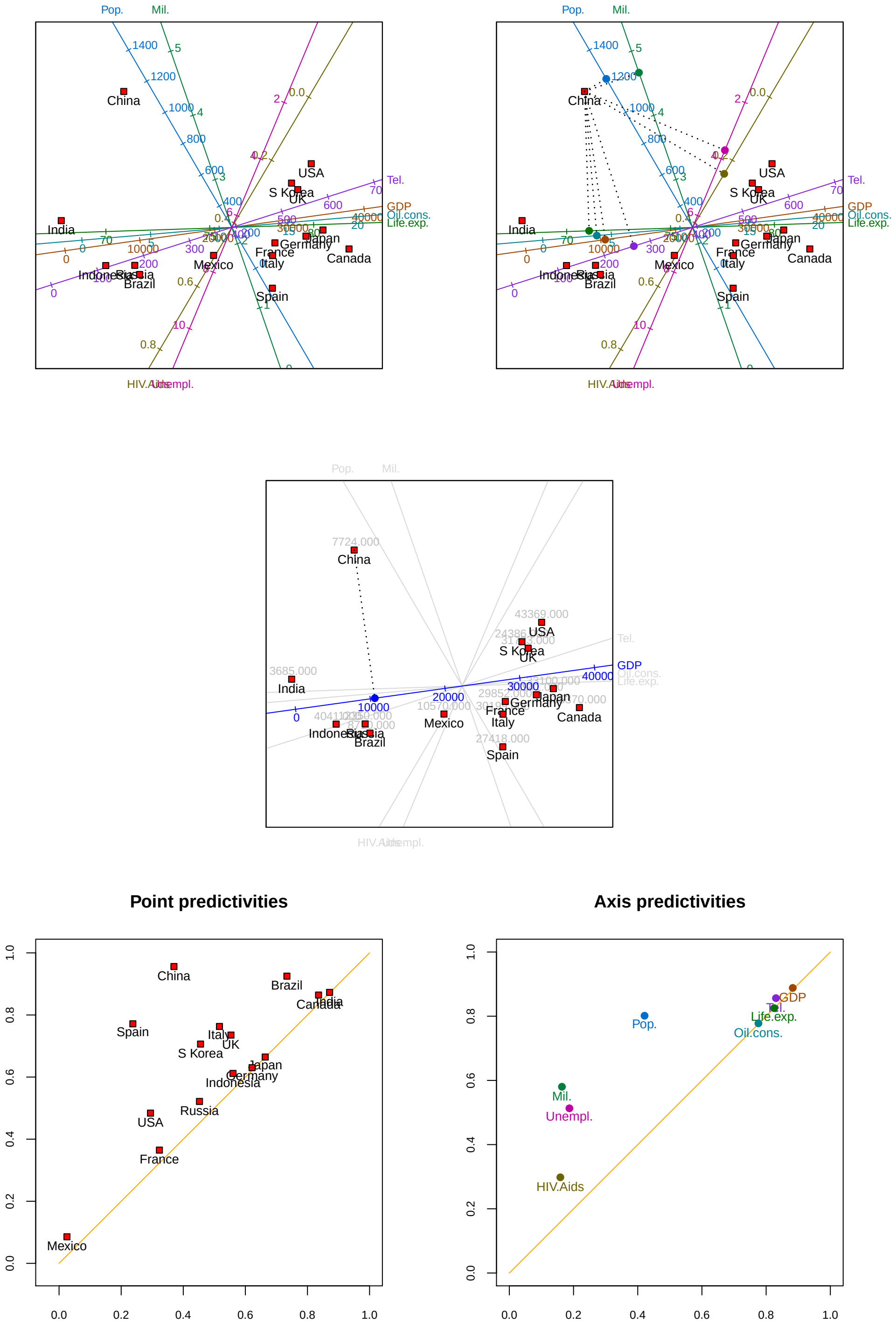


Figure 3: Top left panel: a predictive PCA biplot of the centred, scaled country data. Top right panel: a predictive PCA biplot of the centred, scaled country data with China projected onto all the biplot axes. Centre panel: a predictive PCA biplot of the centred, scaled country data with GDP highlighted and China projected. Bottom left panel: PCA point predictivities of the centred, scaled country data. Bottom right panel: PCA axis predictivities of the centred, scaled country data.

axes, and displays the true variable values of the highlighted axis above the corresponding points. The displays in the diagnostic tabs are shaded accordingly and orthogonal projections are drawn to the highlighted axis only. An example is shown in the centre panel of Figure 3, where GDP is highlighted and China is predicted.

The question of course, is how good the biplot approximation is. This depends on both the points and the axes. As for the points, the 'quality' of the PCA approximation is found by clicking to the `Export` tab, selecting `quality`, and then clicking `Display in console` to display the result in the R console, or otherwise clicking `Save to Workspace` to save the result as an object in the R workspace. In the case of the country data, the quality, 0.693, implies that 69.3% of the variation in the samples is accounted for by the first two principal components. Point and axis predictivities may also be calculated (Gardner-Lubbe *et al.* 2008). Predictivities indicate how well *individual* points or axes are represented in various dimensions of the biplot. Diagrams of point and axis predictivities are available in the `Points` and `Axes` tabs, respectively. Those for the country data are shown in the bottom panels of Figure 3. The points and axes in these figures always appear above the diagonal in the unit square. The further to the right a point or axis appears, the better represented it is in the first (or horizontal) biplot dimension. The closer to the top of the diagram, the better the point or axis is represented overall in the biplot, taking into account the contribution of both the first and the second (vertical) biplot dimension. The marginal contribution of the second biplot dimension is indicated by the vertical distance between the diagonal line and the point or axis. This interpretation suggests that India, Canada and Brazil are relatively well represented in the first biplot dimension. Japan, Germany and Indonesia are represented reasonably in the first dimension, but poorly in the second. France, the United States and Russia are poorly represented overall, and Mexico extremely poorly. China is the best represented country overall. The axes may be similarly interpreted. The two diagrams were saved by right clicking them in the GUI, then making use of the `Save as` options in the pop-up menu. Predictivities are also available numerically from the `Export` tab. The formulae for quality, point predictivities and axis predictivities for PCA biplots are given in Appendix A.3.

Another measure of the goodness of the approximation is its relative absolute error, which may be calculated for any sample on any variable. The relative absolute error is defined to be the absolute difference between the predicted and actual values, expressed as a percentage of the range (max − min) of the actual values of the particular variable. For GDP, for example, the following output is obtained for the country data by selecting `Pred` from the `Export` tab:

```
GDP
          Prediction Actual RelAbsErr%
Brazil        9330.3   8710        1.6
Canada       37282.7  35370        4.8
China        10606.7   7724        7.3
France       27669.1  29852        5.5
Germany      31869.4  31941        0.2
India           40.2   3685        9.2
Indonesia     5054.5   4041        2.6
Italy        27130.3  30199        7.7
Japan        34209.8  33100        2.8
Mexico       19392.0  10570       22.2
```

```
Russia       8865.5  12350        8.8
S Korea     30946.1  24386       16.5
Spain       26507.7  27418        2.3
UK          31644.7  31723        0.2
USA         33889.0  43369       23.9
```

Although the United States, Mexico and South Korea predict poorly on the GDP axis, the overall configuration is optimal. By taking means over the samples, mean relative absolute errors may be obtained for the different variables. From the `Export` tab's `MeanRelAbsErr` entry these are:

```
GDP  HIV.Aids Life.exp.      Mil. Oil.cons.      Pop.      Tel.   Unempl.
7.7      20.1       9.8      14.4      11.2      11.3       8.8      15.1
```

These error rates reinforce what is conveyed by the axis predictivities: that HIV/Aids prevalence, unemployment and military spending are relatively poorly represented, the other variables better. Mean relative absolute errors are useful as a measure of the loss of information in biplots since they can be calculated for any type of biplot. Predictivities are defined only when certain orthogonal decompositions exist (Gardner-Lubbe *et al.* 2008), as they do in the case of PCA, CVA and analysis of distance (AOD) (Krzanowski 2004; Gardner *et al.* 2005) biplots.

For a biplot to be usable in printed form, it must necessarily be two-dimensional. However, assisted by a computer, a user may easily interact with a biplot in three dimensions. Three-dimensional, non-MDS biplots may be obtained in the **BiplotGUI** package by clicking the `External` menu button at the bottom left of the GUI and then selecting the `In 3D` option. Alternatively, the user may simply press the F12 shortcut key shown alongside the option. Doing so renders the three-dimensional version of the currently displayed two-dimensional biplot in an external window. This feature makes use of the **rgl** package (Adler and Murdoch 2009) and allows the biplot to be rotated and enlarged dynamically. Figure 4 shows the three-dimensional predictive PCA biplot of the country data that corresponds to the two-dimensional version at the top left of Figure 3. A further 12.9% of the total variation in the samples is accounted for in the additional dimension – the third value from `eigen` in the `Export` tab divided by the sum of the eigenvalues. An initial 360 degree 'fly-by' of three-dimensional biplots can be enabled via the `File` → `Options` dialogue box.

A PCA approximation results from the projection of samples onto the plane of best fit. In a covariance biplot (`Joint` → `Covariance/Correlation`), these 'scores' are adjusted so that the cosines of the angles between the biplot axes approximate the correlations between the corresponding variables. The correlation biplot is the same as the covariance biplot, but with the variables first scaled to have unit variances (via the settings box, in the usual manner). The correlation biplot of the country data is shown in the top left panel of Figure 5. Three groups of variables are seen to be highly positively correlated (the angles between them are small):

- the number of telephone lines, GDP, oil consumption, life expectancy;
- HIV/Aids prevalence, unemployment;

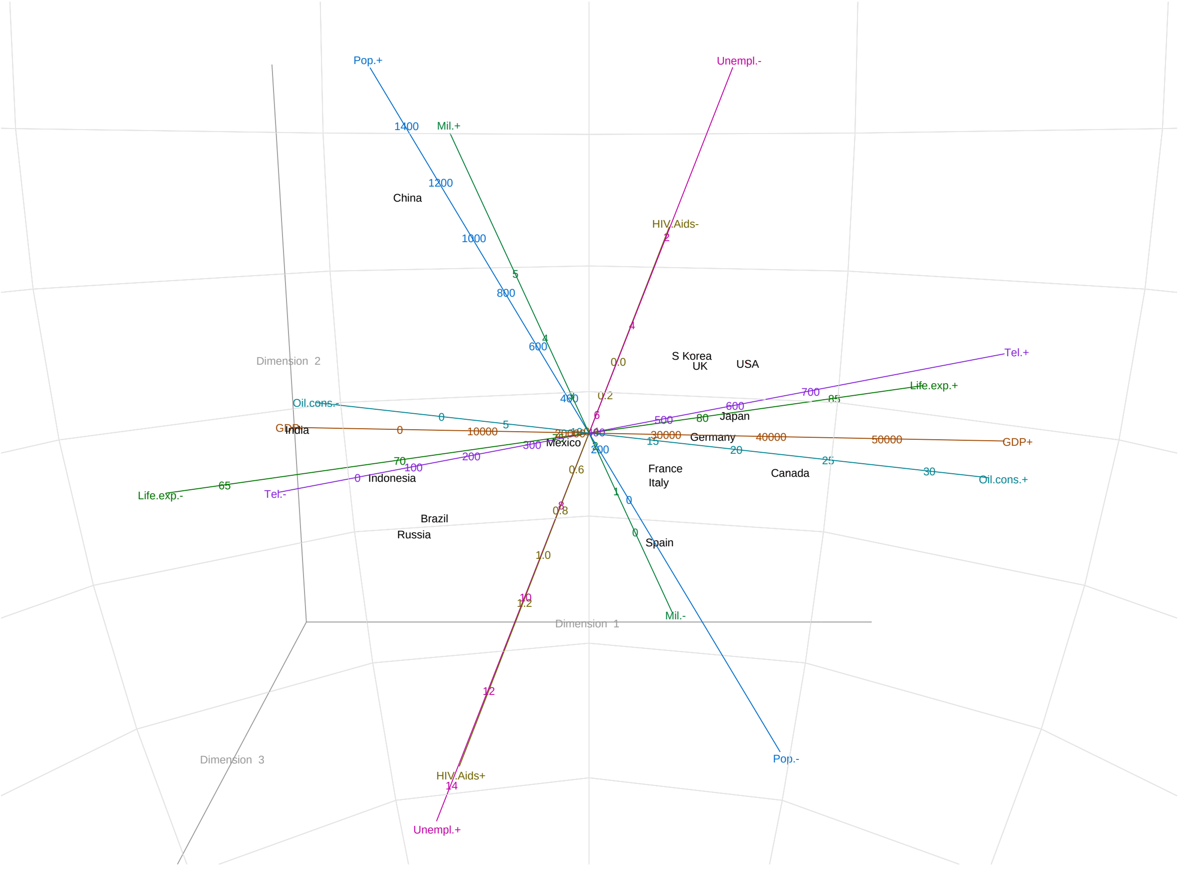


Figure 4: The predictive PCA biplot of the centred, scaled country data, in three dimensions. This figure corresponds to the two-dimensional biplot at the top left of Figure 3.

- population, military spending.

The computations underlying these biplots are set out in Appendix A.4. Notice that the labels of the axes are attached to those ends of the axes that have the higher calibrations. This is the default option for all linear biplots. Alternatively, the axis labels may be given in a legend underneath the biplot, or no axis labels may be given whatsoever. These and other similar options may be set via the `View` menu. Notice also that the option `Joint` → `CVA` is disabled since the samples of the country data have not been grouped in any way, for example by continent. We return to CVA biplots in Section 4.1, where the samples of the antique furniture data set are grouped, and where group differences are investigated.

As opposed to dimension reduction by projection, in MDS the points are chosen so that *stress*, the sum of the squared differences between the inter-sample *disparities* and the inter-point *distances*, is explicitly minimised (details in Appendix A.8). The `Points` → `MDS` menu gives various options. These include taking the inter-sample disparities to be the inter-sample *dissimilarities* themselves (the identity transformation); retaining merely the order of the inter-sample dissimilarities by optimally transforming them into disparities (monotone regression, Kruskal 1964b); or monotonically smoothing the inter-sample dissimilarities into disparities (the monotone spline transformation, Ramsey 1982, 1988). Therefore metric, non-metric and semi-metric MDS representations are available. The inter-sample dissimi-

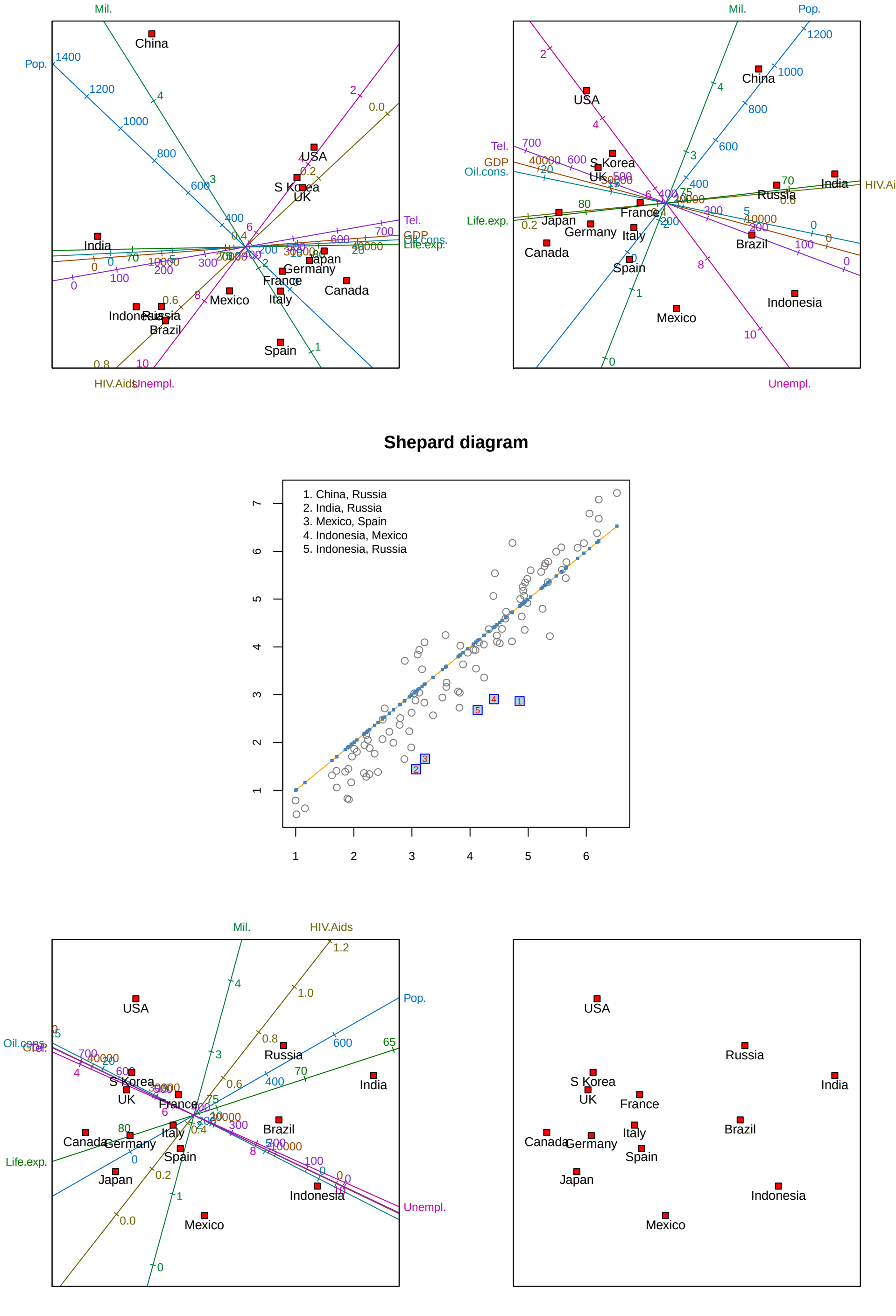


Figure 5: Top left panel: a predictive correlation biplot of the centred, scaled country data. Top right panel: a predictive regression biplot for the metric MDS representation of the centred, scaled country data. The MDS representation is in terms of its principal axes. Centre panel: the Shepard diagram corresponding to the regression biplot of the top right panel. Bottom left panel: a predictive regression biplot for the metric MDS representation of the centred, scaled country data, but with China removed. Bottom right panel: the metric MDS of the centred, scaled data with China removed.

larities are calculated according to the chosen dissimilarity metric (Appendix A.6). Four metrics are currently available from the `Points` → `Dissimilarity metric` menu: `Pythagoras`, `Square-root-of-Manhattan`, `Clark` and `Mahalanobis`. Inter-point distances are always Pythagorean. An iterative majorisation (IM) algorithm (De Leeuw 1977; De Leeuw and Heiser 1980) is used to find the MDS solutions. The IM algorithm converges uniformly, and usually leads to a local minimum, although in theory a saddle-point cannot be ruled out.

The top right panel of Figure 5 shows a metric MDS of the country data, expressed in terms of its principal axes, with approximate regression biplot axes superimposed (from `Points` → `MDS` → `In terms of principal axes`, thereafter `Points` → `MDS` → `Identity transformation`). The default dissimilarity metric, `Pythagoras`, is retained. In this representation, the relative distances between the points are directly related to the corresponding dissimilarities between the countries. The United Kingdom and South Korea, therefore, are more similar to one another than they are to the other countries with respect to the eight variables. As the algorithm converges, updates of the configuration are shown in the biplot region, together with updates of the graphs in the diagnostic tabs. The `Live updates` option, however, may be disabled to increase the speed at which the algorithm runs (the checkbox in question is amongst the buttons at the bottom of the GUI). A graph of the stress values over iterations is given in the `Convergence` tab; in this instance, from the `Export` tab, convergence is reached after 96 iterations, with a final stress value of 45.8. By default, the IM algorithm is taken to have converged as soon as the relative decrease in stress is lower than $10^{-6}$. The algorithm is also stopped once more than 5 000 iterations have been performed. These options can be adjusted via the `File` → `Options` dialogue box. A Shepard diagram (Borg and Groenen 2005, Section 3.3; Shepard 1962) can be found in the `Points` tab and is shown at the centre of Figure 5. Each circle in the Shepard diagram represents a pair of samples. The horizontal axis indicates the inter-sample dissimilarity; the vertical axis indicates the corresponding inter-point distance. The blue dots on the yellow line (which generalises to a step function or a curve) indicate the disparities. Thus the closer the circles are to the line (or step function or curve), the better the overall fit. The five worst-fitting point pairs are identified in the top left corner of the diagram. The dissimilarity between China and Russia, therefore, is most poorly approximated by the points. The `Points` → `MDS` → `Random initial configuration` option forces the algorithm to start from a random configuration at each run; a new run is initiated by clicking `Points` → `MDS` → `Run` or by re-clicking `Points` → `MDS` → `Identity transformation`. Otherwise, the last PCO or MDS solution is taken to be the new initial configuration, as is the case for Figure 5.

To conclude with the country data, suppose that we feel that China is in many ways atypical, and that we would like to see what the effect would be of removing it from consideration. To do so we need simply 'drag' the point representing China from the biplot into the kraal. We may also right click the point representing China and select `Send to kraal` from the pop-up menu. The biplot region is then automatically updated as if China were never part of the data set. The updated biplot is given in the bottom left panel of Figure 5. Russia's position relative to the other countries seems to have been most greatly affected. There has also been a re-alignment amongst the axes, most notably the axes for HIV/Aids, population and unemployment. Axes may also be removed to the kraal. Points and axes which have been removed to the kraal may be dragged back onto the biplot, or the kraal may be emptied of its points only, its axes only, or of both its points and axes simultaneously by making use of the buttons below it, or by right clicking inside it and selecting the desired option from the pop-up menu. At any stage, the points and/or axes of any representation may be hidden by clicking on the options in the `Hide` menu button at the bottom of the window. The figure at the bottom right of Figure 5 is the same as the one in the bottom left panel, but with the biplot axes hidden as described.

# 4. Two more examples

In this section two more examples are considered. In Section 4.1 focuses on grouped data by investigating antique furniture, while non-linear prediction is illustrated in Section 4.2 at the hand of fighter aircraft data.

## 4.1. Antique furniture

It is often of great interest to collectors, auctioneers and cultural historians to be able to correctly identify the type of wood used to make antique furniture. In the period between 1652 and 1900, wood from both the indigenous *Ocotea bullata* ('Stinkwood') and the imported *Ocotea porosa* ('Imbuia') were used to make Old-Cape furniture in South Africa. Being from the same genus and family (*Lauraceae*), it is often difficult to distinguish between the two types of wood based solely on a traditional analysis of colour, smell, and other observable characteristics. Burden *et al.* (2001) and Le Roux and Gardner (2005) make use of CVA biplots of anatomical measurements to distinguish between the species. A third species, *Ocotea kenyensis*, is also included in the analyses. The microscopically measured variables are: tangential vessel diameter in µm (VesD); vessel element length in µm (VesL); fibre length in µm (FibL); ray height in µm (RayH); ray width in µm (RayW); and the number of vessels per mm$^2$ (NumVes). The 37 observations are the mean values over fifty repeat-measurements made on 20 samples of *Ocotea bullata*, 10 samples of *Ocotea porosa*, and 7 samples of *Ocotea kenyensis*. The data are included in the **BiplotGUI** package as the data frame `AntiqueFurniture`, of which the first column contains the group specifications. The data may be viewed from within R by entering the following instructions at the prompt of the R console:

```
R> data("AntiqueFurniture")
R> AntiqueFurniture
```

To initialise the GUI with the antique furniture data, the following instruction may be entered:

```
R> Biplots(Data = AntiqueFurniture[, -1], groups = AntiqueFurniture[, 1])
```

In other words, the data consist of all the columns of `AntiqueFurniture` except the first, while the group specifications are precisely the contents of the first column.

As was mentioned earlier, upon initialisation of the GUI, the predictive PCA biplot is shown by default. To show the CVA biplot instead, the user simply needs to click the option `Joint` → `CVA`. This option is now available since, in the call to the `Biplots` function, groups were specified. The predictive CVA biplot of the antique furniture data is shown in the top left panel of Figure 6. The positions of the points are determined by the first two canonical variates – those linear combinations of the original variables that maximally separate the group

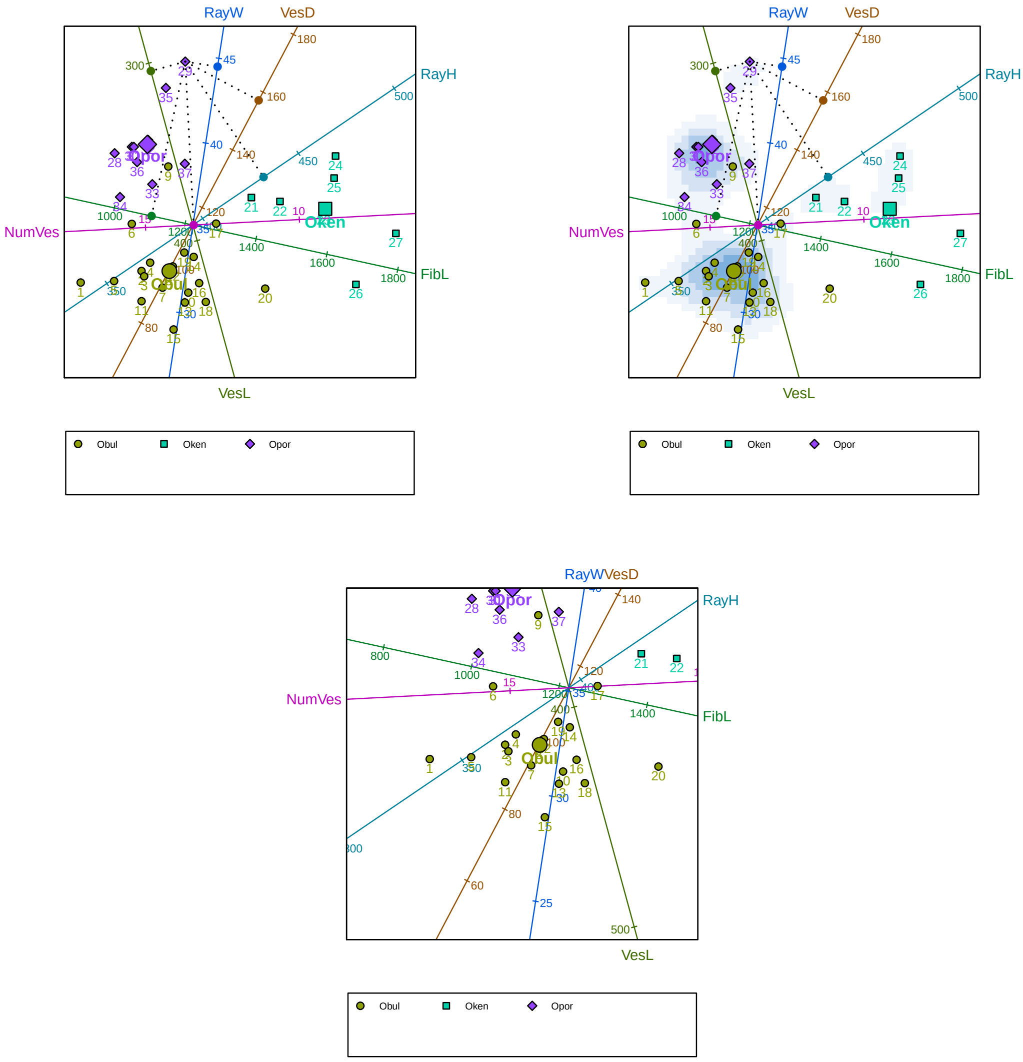


Figure 6: Top left panel: a predictive CVA biplot of the antique furniture data with sample 29 projected onto all the biplot axes. The group means are also shown. Top right panel: as in the top left panel, but with the biplot overlain onto the two-dimensional density estimate of the points. Bottom panel: the same biplot as the other two, but zoomed in around the mean of the species *Ocotea bulluta*.

means, subject to certain restrictions (Krzanowski 2000, Section 11.1). The group means themselves are shown as larger but corresponding symbols (activated by clicking `Additional` → `Interpolate` → `Sample group means`, retaining the default options). Since there is more than one group, an optional legend is included below the biplot by default. The mechanism for the prediction of the variable values is the same as before and is illustrated in the figure in the case of sample 29.

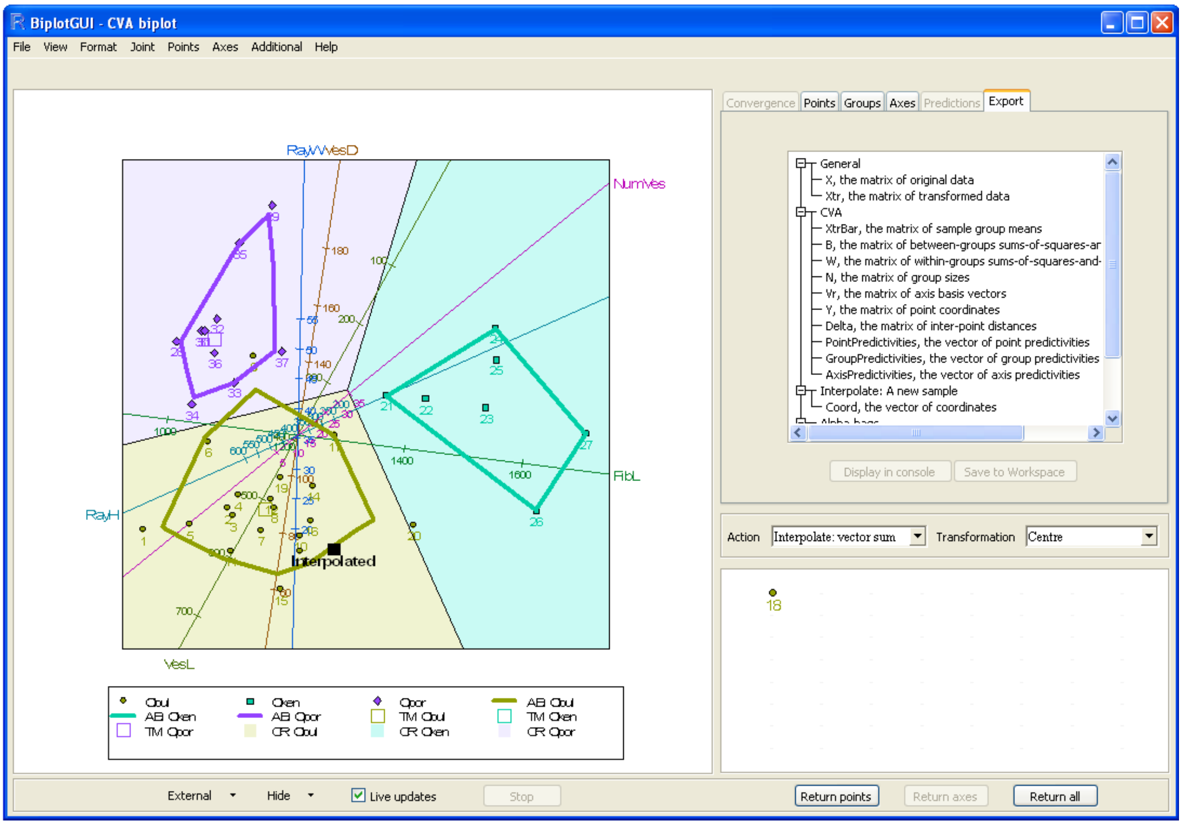
The top right panel of Figure 6 shows the same biplot, now overlain onto a two-dimensional density estimate of the points. The density estimate is obtained by clicking `Additional` → `Point densities` and accepting the default options (amongst other things, for the point densities to be estimated for all points, as opposed to certain groups of points only). The point densities are calculated using the default arguments to the `bkde2D` function of the **KernSmooth** package (Wand 2009). Similar biplots can be found in Blasius *et al.* (2009).

Sometimes it is helpful to zoom into or out of portions of a biplot. This is done by right clicking on a focal point inside the biplot, and selecting the `Zoom in` or `Zoom out` option from the pop-up menu which then appears. The bottom panel of Figure 6 shows the CVA biplot of antique furniture, enlarged around the mean of the species *Ocotea bullata*. The original view can be restored by choosing the `Reset zoom` option from the pop-up menu.

A screenshot of the GUI is shown in Figure 7. To the left, a CVA biplot of the antique furniture data appears. From the settings box, it can be seen that the axes are not predictive; in fact they are vector sum interpolative. Also, the data have not been transformed, except for the obligatory centring of the columns to have zero means. In any case, CVA biplots are unaffected by the scaling of the variables to have unit variance.

Sample 18 has been dragged from the biplot into the kraal. It has therefore not been taken into account in the construction of the biplot. However, using its original variable values—104, 387, 1 290, 381, 22 and 12, respectively—it has subsequently been interpolated onto the biplot towards the bottom of the image (using the `Additional` → `Interpolate` → `A New Sample` option). This is the most appropriate position for the sample in the existing biplot. It is reassuring that the positions assigned to sample 18 in Figures 6 and 7 correspond so closely. This need not have been the case. Also notice that, notwithstanding the removal of sample 18, the calibrations and *directions* of the predictive and interpolative biplot axes differ. This is in general the case for CVA biplots. More details are given in Appendix A.5.

The biplot in Figure 7 also sports colour-coded classification regions. These are the regions in the display space plane closest to the respective group means in a specified number of canonical dimensions, here the default number, two. The classification regions are included by selecting `Classification regions` from the `Additional` menu. They may be used for the classification of interpolated samples. For more on the links between biplots and discrimination, see Gardner and Le Roux (2005). Furthermore, by clicking `Additional` → `Alpha-bags`, alpha-bags (Gardner 2001; Aldrich *et al.* 2004) and Tukey medians have been superimposed for the species *Ocotea bullata* and *Ocotea porosa* (there are too few samples for an alpha-bag for *Ocotea kenyensis* to be constructed; with an appropriate warning, a convex-hull is displayed instead). Alpha-bags are closely related to the bagplots of Rousseeuw *et al.* (1999) and enclose regions that contain approximately the inner $100\alpha\%$ of samples, here 90% of the samples for the two species separately. The alpha-bags and convex hull do not overlap. This emphasises the high degree of separation between the species. For CVA biplots, group predictivities may also be calculated, in addition to the point and axis predictivities



General
- X, the matrix of original data
- Xtr, the matrix of transformed data

CVA
- XtrBar, the matrix of sample group means
- B, the matrix of between-groups sums-of-squares-ar
- W, the matrix of within-groups sums-of-squares-and-
- N, the matrix of group sizes
- Vr, the matrix of axis basis vectors
- Y, the matrix of point coordinates
- Delta, the matrix of inter-point distances
- PointPredictivities, the vector of point predictivities
- GroupPredictivities, the vector of group predictivities
- AxisPredictivities, the vector of axis predictivities

Interpolate: A new sample
- Coord, the vector of coordinates

Action: Interpolate: vector sum

Transformation: Centre

☑ Live updates

Figure 7: A screenshot of the GUI. A vector sum interpolative CVA biplot of the antique furniture data is shown towards the left, with sample 18 removed to the kraal. Sample 18 has then been interpolated to give its implied position. Classification regions are shown, as well as 90% alpha-bags for the species *Ocotea bullata* and *Ocotea porosa*. A convex hull surrounds the points of the species *Ocotea kenyensis*. The `Export` tab is shown top right.

discussed earlier (Gardner-Lubbe *et al.* 2008). A diagram of these is available in the `Groups` tab. Finally, Figure 7 also shows the `Export` tab. As explained previously, various objects are available for export from this tab. The objects may be displayed in the R console or be saved to the current R workspace. The list of available objects depends on what is shown in the biplot.

## 4.2. Fighter aircraft

Measurements of four variables on 22 types of fighter aircraft were extracted by Cook and Weisberg (1982) from a report by Stanley and Miller (1979). Following Gower and Hand (1996), only the first 21 of these aircraft in the biplots below. The four variables are: specific power, proportional to power per unit weight (SPR); flight range factor (RGF); payload as a fraction of gross weight (PLF); and sustained load factor (SLF). These data can be found in the `FighterAircraft` data frame included in the **BiplotGUI** package. The GUI is initialised in the same way as it was for the country data in Section 3.2.

The left panel of Figure 8 shows a regression biplot of the fighter aircraft data with the points determined by PCO and the inter-sample dissimilarities calculated according to the

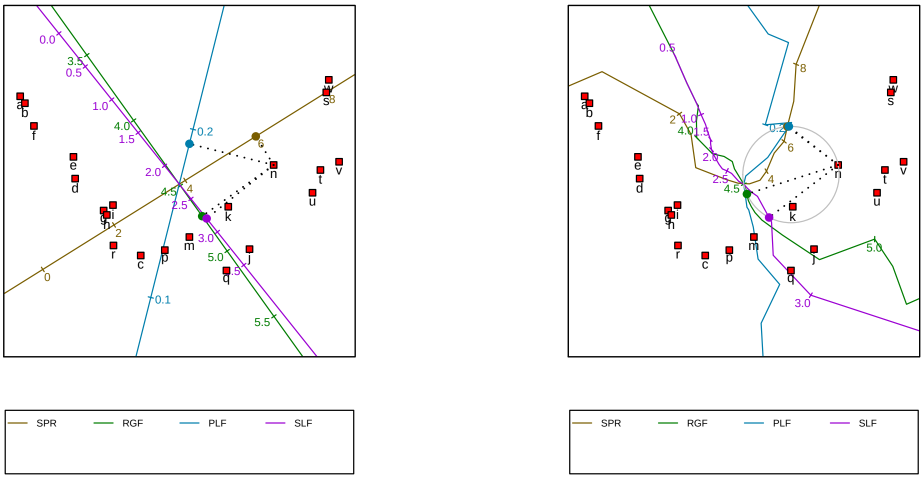


Figure 8: Left panel: a predictive regression biplot of the fighter aircraft data, with points determined by PCO based on the Square-root-of-Manhattan dissimilarity metric. The orthogonal prediction of the variables values of aircraft ‘n’ is shown. Right panel: a predictive circular non-linear biplot of the fighter aircraft data, with points determined by PCO based on the Square-root-of-Manhattan dissimilarity metric. The circular prediction of the variable values of aircraft ‘n’ is shown.

Square-root-of-Manhattan dissimilarity metric. The figure is obtained by clicking `Points` → `Dissimilarity metric` → `Square-root-of-Manhattan` and then `Axes` → `Regression`. Making use of orthogonal projection, the variable values for aircraft ‘n’ are predicted to be 5.980, 4.73, 0.191 and 2.80, respectively. These can be compared to the actual values, 5.855, 4.53, 0.172 and 2.50. The right panel of Figure 8 shows the corresponding circular non-linear biplot (obtained by clicking `Axes` → `Circular non-linear`). Here prediction is performed by completing the circle which has, as diagonal, the line stretching from the origin of the biplot to the point to be predicted. The predicted values are read off at the points at which the circle intersects the axes (Gower and Hand 1996, Section 6.3.2). If a particular axis is intersected at more than one position, the position closest to the point being predicted is used. If an axis isn’t intersected at all, no prediction can be made for the corresponding variable. For aircraft ‘n’, the valid points of intersection are shown in the figure as small, filled circles on the circumference of the larger circle. From the `Predictions` or `Export` tabs, the circular non-linear predictions for aircraft ‘n’ are 6.090, 4.54, 0.174 and 2.90, respectively (these values depend on how finely the non-linear axes are constructed; by default 20 positions are taken into account from each calibrated marker to the next). Except for the fourth variable, the non-linear predictions are very close to the actual values.

Appendix A.7 provides the steps required to perform a PCO; the formulae underlying the regression and circular non-linear biplots are laid out in Sections A.9 and A.11, respectively.

## 5. Further features

This section touches upon the customisation and export of biplots and other graphs produced

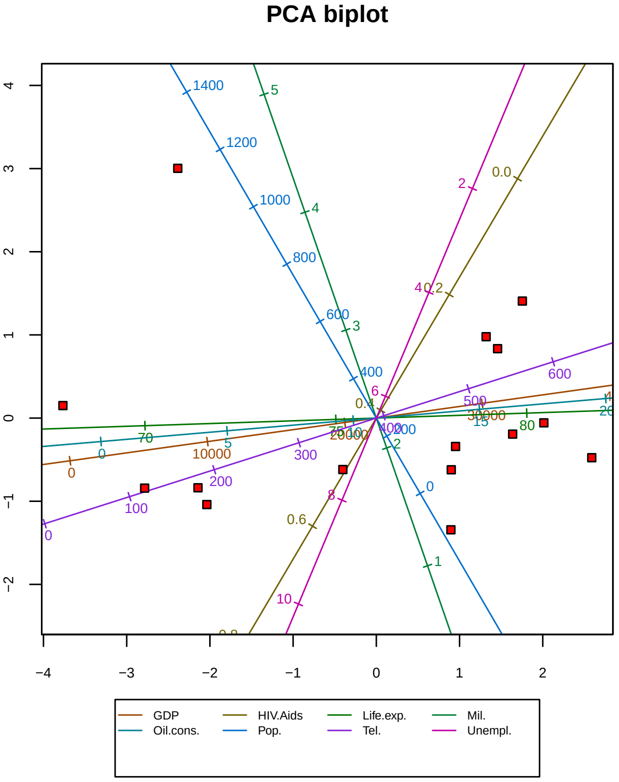


Figure 9: A modified version of the biplot given in the top left panel of Figure 3: a predictive PCA biplot of the country data with a title, hidden point labels, axis labels in a legend, and the axes of the display space calibrated.

using the **BiplotGUI** package.

There are two main ways in which the graphs of the package can be customised. Basic customisation can be performed using the options of the `View` menu, while the `Format` menu options can be used to alter a large number of graphical parameters.

Figure 9 shows the same predictive PCA biplot of the country data that was shown in the top left panel of Figure 3. However, the biplot in Figure 9 has been modified by changing the default selections in the `View` menu. The `Show title` option places a title above the biplot; by default the title reflects the type of biplot, but it may be changed via the `Format` → `Title` option. Furthermore, the point labels have been hidden by deselecting `Show point labels`. Instead of showing the axis labels around the edges of the biplot as in Figure 3, the labels in Figure 9 are shown in a legend (`Show axis labels in legend`). The `Calibrate display space axes` option calibrates the two dimensions of the biplot, but this is generally undesirable in biplots of the new approach (Gower and Hand 1996, Section 2.6).

The `Format` menu allows virtually any of the graphical parameters used internally by the package to be altered. The biplot in Figure 10 serves as an example. This biplot is the same as the biplot that appears in the left panel of Figure 6, but with some of the default graphical parameters changed. The `By group` option allows the graphical parameters that relate to points, sample group means, convex hulls / alpha-bags and classification regions to be set

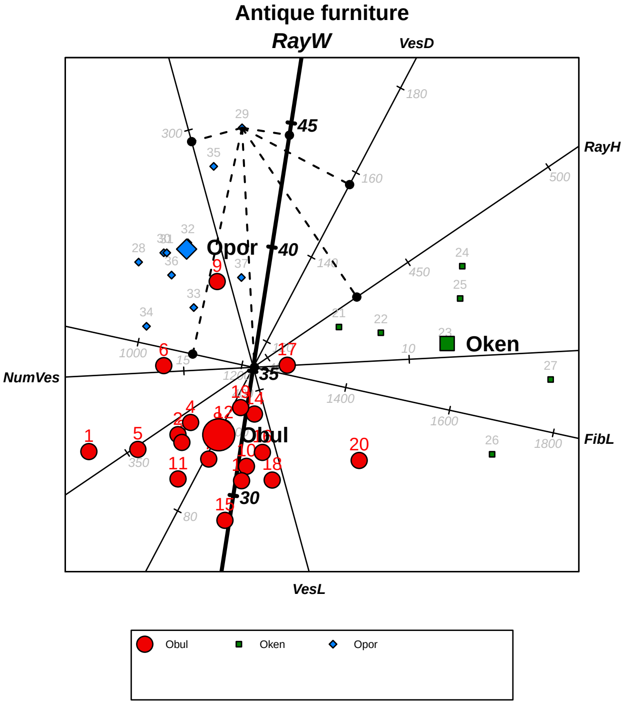
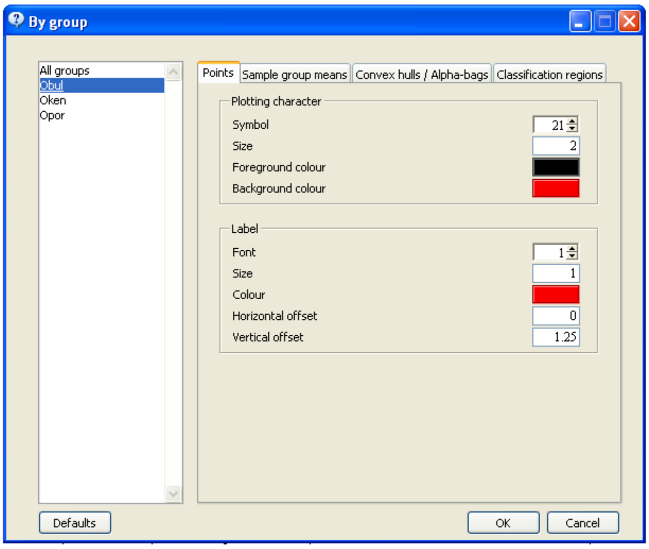


Figure 10: A modified version of the biplot given in the left panel of Figure 6: a predictive CVA biplot of the antique furniture data.

for all groups simultaneously, or for a single group at a time. Figure 11 shows the `By group` dialogue box for the points of the species *Ocotea bullata*. The parameter values shown are as they have been set for Figure 10. The `Axes` option similarly allows the graphical parameters that relate to axes to be set. Figure 12 shows the `Axes` dialogue box for the axis 'RayW', again with the parameters as they have been set for Figure 10. The graphical parameters used for dynamic variable value prediction and in the highlighting of axes can also be modified by clicking `Format` → `Interaction`, while diagnostic tab customisation may be performed via the `Diagnostic tabs` option. The `Reset all` option reverts all the graphical parameters back to their default values. In all, more than 80 different graphical parameters may be set, often-times differently for different groups or axes. All these parameters are documented in detail in the package manual.

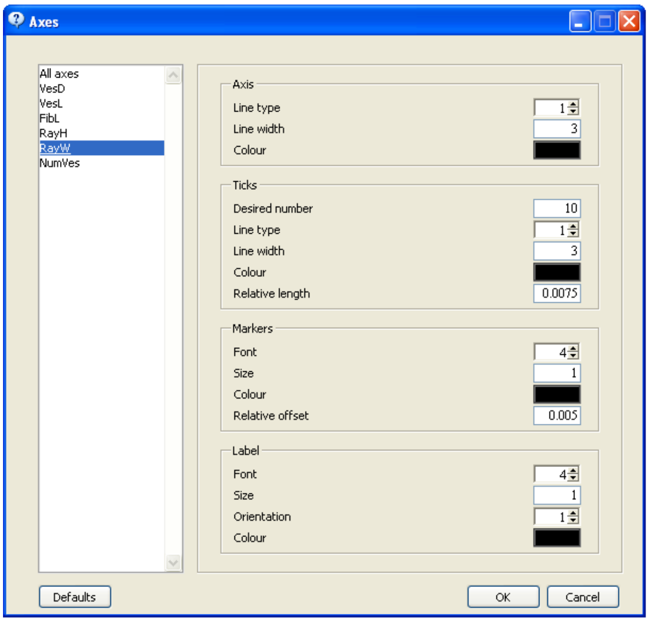
Biplots and diagnostic tab graphs can be saved in various file formats: PDF, Postscript, Metafile, BMP, PNG, JPEG (50%, 75%, 100% quality) and PicTeX. Any graph can be saved by right clicking it and navigating the `Save as` menu. The biplot region may also be saved via the `File` → `Save as` menu. While the images shown onscreen are by necessity Metafile images, the images that appear in this vignette—besides the screenshots—were saved in PDF format. Together with Postscript, such images are of the highest quality. `Copy` and `Print` options are also available.



Plotting character
Symbol 21
Size 2
Foreground colour ⬛
Background colour 🟥

Label
Font 1
Size 1
Colour 🟥
Horizontal offset 0
Vertical offset 1.25

Figure 11: The `Format` → `By group` dialogue box as it appears for the biplot in Figure 10.



Axis
Line type 1
Line width 3
Colour ⬛

Ticks
Desired number 10
Line type 1
Line width 3
Colour ⬛
Relative length 0.0075

Markers
Font 4
Size 1
Colour ⬛
Relative offset 0.005

Label
Font 4
Size 1
Orientation 1
Colour ⬛

Figure 12: The `Format` → `Axes` dialogue box as it appears for the biplot in Figure 10.

# 6. Future work

Being in its first release, there is much that may be improved and expanded upon. Amongst the techniques that might sensibly be incorporated into the package are:

- Special options for CVA biplots in the case of two groups only (Le Roux and Gardner-Lubbe 2008);
- Orthogonal predictive non-linear biplots (Gower and Ngouenet 2005);
- AOD biplots (Krzanowski 2004; Gardner *et al.* 2005);
- The adjustments to the regression and Procrustes biplots suggested by Gower *et al.* (1999) to better suit non-metric MDS representations;
- Sensitivity analysis for PCO-based biplots (Krzanowski 2006);
- Support for categorical variables in the form of generalised biplots (Gower 1992);
- A better approach to the calculation of classification regions (Gower 1993).

Other improvements may also be made. These include:

- Allowing any pair of principal components, canonical variates or principal coordinates to be shown, as opposed to only the first two;
- Supporting a greater number of dissimilarity metrics;
- Allowing interactive orthogonal parallel translation so that axes can be moved towards the edges of biplots (Blasius *et al.* 2009);
- Incorporating a graded legend for point density estimates;
- Improving the three-dimensional biplots (providing support for additional descriptors, allowing dynamic variable value prediction);
- Otherwise improving the GUI and general performance.

As for any such package, suggestions and bug-reports by users are important and greatly encouraged.

# 7. Summary

In this paper, the **BiplotGUI** package for R was introduced. Its features were illustrated using three data sets. Ideas for future releases were briefly explored, and computational details were provided in an appendix.

The package makes it possible to easily construct many types of biplots and to interact with them in various ways. The package is free and its source code shared. Amongst linear biplots, the PCA, covariance/correlation, CVA, regression and Procrustes biplots are supported. Circular non-linear biplots can be created. In addition, PCO and MDS representations can be displayed on their own, without added biplot axes. Additional descriptors can be superimposed, and three-dimensional biplots can be explored using the **rgl** package. Various goodness-of-fit measures are easily accessible.

# References

Adler D, Murdoch D (2010). ***rgl**: 3D Visualization Device System (OpenGL).* R package version 0.90, URL http://CRAN.R-project.org/package=rgl.

Agency C (2007). *The World Factbook: 2007, CIA's 2006.* Potomac Books, Washington, DC, USA.

Aldrich C, Gardner S, Le Roux NJ (2004). "Monitoring of Metallurgical Process Plants by Use of Biplots." *AIChE Journal*, **50**(9), 2167–2186.

Alves MR, Cunha SC, Amaral JS, Pereira JA, Oliveira MB (2005). "Classification of PDO Olive Oils on the Basis of Their Sterol Composition by Multivariate Analysis." *Analytica Chimica Acta*, **549**, 166–178.

Blasius J, Eilers P, Gower JC (2009). "Better Biplots." *Computational Statistics & Data Analysis*, **53**, 3145–3158.

Borg I, Groenen PJF (2005). *Modern Multidimensional Scaling: Theory and Applications.* Springer Series in Statistics. Springer-Verlag, New York, NY, USA, 2nd edition.

Braun WJ, Murdoch DJ (2007). *A First Course in Statistical Programming with R.* Cambridge University Press, Cambridge, UK.

Burden M, Gardner S, Le Roux NJ, Swart JPJ (2001). "Ou-Kaapse Meubels en Stinkhout-Identifikasie: Moontlikhede met Kanoniese Veranderlike-Analise en Bistippings." *South African Journal of Cultural History*, **15**, 50–73.

Chambers JM (2007). *Software for Data Analysis: Programming with R.* Springer-Verlag, New York, NY, USA.

Clark PJ (1952). "An Extension of the Coefficient of Divergence for Use with Multiple Characters." *Copeia*, **2**, 61–64.

Cook RD, Weisberg S (1982). *Residuals and Influence in Regression.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, UK.

Cox TF, Cox MAA (2001). *Multimensional Scaling.* Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL, USA, 2nd edition.

De Leeuw J (1977). "Applications of Convex Analysis to Multidimensional Scaling." In JR Barra, F Brodeau, G Romier, B van Cutsem (eds.), "Recent Developments in Statistics," pp. 133–145. North-Holland Publishing Company, Amsterdam, The Netherlands.

De Leeuw J, Heiser WJ (1980). "Multidimensional Scaling with Restrictions on the Configuration." In PR Krishnaiah (ed.), "Multivariate Analysis," volume V, pp. 501–522. North-Holland Publishing Company, Amsterdam, The Netherlands.

Dray S, Dufour AB (2007). "The **ade4** Package: Implementing the Duality Diagram for Ecologists." *Journal of Statistical Software*, **22**(4), 1–20. URL http://www.jstatsoft.org/v22/i04.

Dray S, Dufour AB (2009). ***ade4**: Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences.* R package version 1.4-14, URL http://CRAN.R-project.org/package=ade4.

Faria JC, Demetrio CGB (2009). ***bpca**: Biplot of Multivariate Data Based on Principal Components Analysis.* UESC and ESALQ, Ilheus, Bahia, Brasil and Piracicaba, Sao Paulo, Brasil. R package version 1.03, URL http://CRAN.R-project.org/package=bpca.

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

Gabriel KR (1972). "Analysis of Meteorological Data by Means of Canonical Decomposition and Biplots." *Journal of Applied Meteorology*, **11**, 1071–1077.

Gardner S (2001). *Extensions of Biplot Methodology to Discriminant Analysis with Applications of Non-Parametric Princial Components.* Unpublished PhD thesis, Stellenbosch University, Stellenbosch, South Africa.

Gardner S, Le Roux NJ (2005). "Extentions of Biplot Methodology to Discriminant Analysis." *Journal of Classification*, **22**, 59–86.

Gardner S, Le Roux NJ, Rypstra T, Swart JPJ (2005). "Extending a Scatterplot for Displaying Group Structure in Multivariate Data: A Case Study." *ORiON*, **21**(2), 111–124.

Gardner-Lubbe S, Le Roux NJ, Gower JC (2008). "Measures of Fit in Principal Component and Canonical Variate Analyses." *Journal of Applied Statistics*, **35**(9), 947–965.

Gower JC (1966). "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." *Biometrika*, **53**(3/4), 325–338.

Gower JC (1982). "Euclidean Distance Geometry." *The Mathematical Scientist*, **7**, 1–14.

Gower JC (1992). "Generalised Biplots." *Biometrika*, **79**, 475–493.

Gower JC (1993). "The Construction of Neighbour-Regions in Two Dimensions for Prediction with Multi-Level Categorical Variables." In O Opitz, B Lausen, R Klar (eds.), "Information and Classification: Concepts-Methods-Applications Proceedings 16th Annual Conference of the Gesellschaft fur Klassifikation," pp. 174–189. Springer-Verlag, Berlin, Germany.

Gower JC, Dijksterhuis GB (2004). *Procrustes Problems.* Oxford Statistical Science Series. Oxford University Press, Oxford, UK.

Gower JC, Hand DJ (1996). *Biplots.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, UK.

Gower JC, Harding SA (1988). "Nonlinear Biplots." *Biometrika*, **75**(3), 445–455.

Gower JC, Legendre P (1986). "Metric and Euclidean Properties of Dissimilarity Coefficients." *Journal of Classification*, **16**, 5–48.

Gower JC, Meulman JJ, Arnold GM (1999). "Nonmetric Linear Biplots." *Journal of Classification*, **16**, 181–196.

Gower JC, Ngouenet RF (2005). "Nonlinearity Effects in Multidimensional Scaling." *Journal of Multivariate Analysis*, **94**, 344–365.

Graffelman J (2010). ***calibrate**: Calibration of Biplot Axes.* R package version 1.7, URL http://CRAN.R-project.org/package=calibrate.

Graffelman J, van Eeuwijk FA (2005). "Calibration of Multivariate Scatter Plots for Exploratory Analysis of Relations Within and Between Sets of Variables in Genomic Research." *Biometrical Journal*, **47**(6), 863–879.

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

Greenacre MJ (2007). *Correspondence Analysis in Practice.* Interdisciplinary Statistics. Chapman & Hall/CRC, London, UK, 2nd edition.

Grosjean P (2010). ***tcltk2**: Additional Tcl/Tk Widgets and Commands for R.* R package version 1.1-2, URL http://CRAN.R-project.org/package=tcltk2.

Highland Statistics Ltd (2008). ***brodgar**, Version 2.5.7.* Highland Statistics Ltd, Newburgh, UK. URL http://www.brodgar.com/.

Hofmann H (2000). ***Manet**, Version 1862.* URL http://stats.math.uni-augsburg.de/manet.

Hotelling H (1933). "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology*, **24**, 417–441.

Hotelling H (1935). "The Most Predictable Criterion." *Journal of Educational Psychology*, **26**, 139–142.

Hotelling H (1936). "Relations Between Two Sets of Variables." *Biometrika*, **28**, 321–377.

Ihaka R, Gentleman R (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

Ihaka R, Murrel P, Hornik K, Zeileis A (2009). ***colorspace**: Colorspace Manipulation.* R package version 1.0-1, URL http://CRAN.R-project.org/package=colorspace.

Jemwa GT, Aldrich C (2006). "Kernel-Based Fault Diagnosis on Mineral Processing Plants." *Minerals Engineering*, **19**, 1149–1162.

Jolliffe IT (2002). *Principal Component Analysis.* Springer Series in Statistics. Springer-Verlag, New York, NY, USA, 2nd edition.

Kovach Computing Services (2008). ***MVSP**, Version 3.1.* Kovach Computing Services, Anglesey, UK. URL http://www.kovcomp.co.uk/mvsp.

Kruskal JB (1964a). "Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis." *Psychometrika*, **29**, 1–27.

Kruskal JB (1964b). "Nonmetric Multidimensional Scaling: A Numerical Method." *Psychometrika*, **29**, 115–129.

Krzanowski WJ (2000). *Principles of Multivariate Analysis: A User's Perspective.* Oxford Statistical Science Series. Oxford University Press, New York, NY, USA, revised edition.

Krzanowski WJ (2004). "Biplots for Multifactorial Analysis of Distance." *Biometrics*, **60**, 517–524.

Krzanowski WJ (2006). "Sensitivity in Metric Scaling and Analysis of Distance." *Biometrics*, **62**, 239–244.

La Grange AM, Le Roux NJ, Gardner-Lubbe S (2009). "**BiplotGUI**: Interactive Biplots in R." *Journal of Statistical Software*, **30**(12), 1–37. URL http://www.jstatsoft.org/v30/i12.

Le Roux NJ, Gardner S (2005). "Analysing Your Multivariate Data as a Pictorial: A Case for Applying Biplot Methodology?" *International Statistical Review*, **73**(3), 365–387.

Le Roux NJ, Gardner-Lubbe S (2008). "Geometrical Considerations and Biplots Associated with Canonical Variate Analysis Involving Two Classes." Conference on High-dimensional Data Modelling. June 2008, Kayseri, Turkey.

Lipkovich I, Smith EP (2002a). ***BiPlot***. Virginia Tech, Blacksburg, VA, USA. URL http://filebox.vt.edu/artsci/stats/vining/keying/biplot_final.zip.

Lipkovich I, Smith EP (2002b). "Biplot and Singular Value Decomposition Macros for Excel." *Journal of Statistical Software*, **7**(5), 1–15. URL http://www.jstatsoft.org/v07/i05.

Mahalanobis PC (1936). "On the Generalised Distance in Statistics." *Proceedings of the National Institute of Science of India*, **12**, 49–55.

Minitab Inc (2007). *Minitab, Version 15.* Minitab Inc, State College, PA, USA. URL http://www.minitab.com/.

MjM Software Design (2009). ***PC-ORD**, Version 5.20.* MjM Software Design, Gleneden Beach, OR, USA. URL http://home.centurytel.net/~mjm/pcordwin.htm.

Murrel P (2005). *R Graphics.* Chapman & Hall/CRC, Boca Raton, FL, USA.

Naidoo S, Harris A, Swanevelder S, Lombard C (2006). "Fetal Alcohol Syndrome: A Cephalometric Analysis of Patients and Controls." *European Journal of Orthodontics*, **28**, 254–261.

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Stevens MHH, Wagner H (2010). ***vegan**: Community Ecology Package.* R package version 1.17-0, URL http://CRAN.R-project.org/package=vegan.

Pearson K (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space." *Philosophical Magazine*, **2**, 559–572.

Plant Research International (2002). ***Canoco**, Version 4.5.* Plant Research International, Wageningen, The Netherlands. URL http://www.pri.wur.nl/uk/products/canoco.

Ramsey JO (1982). "Some Statistical Approaches to Multidimensional Scaling Data." *Journal of the Royal Statistical Society. Series A (General)*, **145**(3), 285–312.

Ramsey JO (1988). "Monotone Regression Splines in Action." *Statistical Science*, **3**(4), 425–441.

R Development Core Team (2009). *R: Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rousseeuw PJ, Ruts I, Tukey JW (1999). "The Bagplot: A Bivariate Boxplot." *The American Statistician*, **53**(4), 382–387.

Sammon JW (1969). "A Nonlinear Mapping for Data Structure Analysis." *IEEE Transactions on Computers*, **18**, 401–409.

SAS Institute Inc (2009). Cary, NC, USA. URL http://www.sas.com/.

Shepard RN (1962). "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function." *Psychometrika*, **27**(2), 125–140.

Smilauer P (2003). ***CanoDraw**, Version 4.1*. University of South Bohemia, Ceské Budejovice, Czech Republic. URL http://www.canodraw.com/.

Spector P (2008). *Data Manipulation with R*. Springer-Verlag, New York, NY, USA.

SPSS Inc (2008). *SPSS, Version 17*. Chicago, IL, USA. URL http://www.spss.com/.

Stanley W, Miller M (1979). "Measuring Technological Change in Jet Fighter Aircraft." *Technical Report R-2249-AF*, RAND Corporation, Santa Monica, CA, USA.

StataCorp LP (2007). *Stata, Version 10*. StataCorp LP, College Station, TX, USA. URL http://www.stata.com/.

StatSoft Inc (2009). ***STATISTICA**, Version 9*. Tulsa, OK, USA. URL http://www.statsoft.com/.

Thioulouse J, Dray S (2007). "Interactive Multivariate Data Analysis in R with the **ade4** and **ade4TkGUI** Packages." *Journal of Statistical Software*, **22**(5), 1–14. URL http://www.jstatsoft.org/v22/i05.

Thioulouse J, Dray S (2009). ***ade4TkGUI**: **ade4** Tcl/Tk Graphical User Interface*. R package version 0.2-5, URL http://CRAN.R-project.org/package=ade4TkGUI.

Tierney L (1990). *Lisp-Stat: An Object-Oriented Environment for Statistical and Dynamic Graphics*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, NY, USA.

Tierney L (2008). ***tkrplot**: TK Rplot*. R package version 0.0-18, URL http://CRAN.R-project.org/package=tkrplot.

Torgerson WS (1952). "Multidimensional Scaling: 1. Theory and Method." *Psychometrika*, **17**, 401–419.

Turner R (2010). ***deldir**: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation.* R package version 0.0-12, URL http://CRAN.R-project.org/package=deldir.

Udina F (2005a). "Interactive Biplot Construction." *Journal of Statistical Software*, **13**(5), 1–16. URL http://www.jstatsoft.org/v13/i05.

Udina F (2005b). ***XLS-Biplot**, Version 1.1a.* Universitat Pompeu Fabra, Barcelona, Spain. URL http://tukey.upf.es/xls-biplot/index.html.

Underhill LG (1990). "The Coefficient of Variation Biplot." *Journal of Classification*, **7**, 241–256.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Statistics and Computing. Springer-Verlag, New York, NY, USA, 4th edition. URL http://www.stats.ox.ac.uk/pub/MASS4.

VSN International Ltd (2008). *Genstat, Version 11.1.* VSN International Ltd, Hemel Hempstead, UK. URL http://www.genstat.com/.

Wand M (2009). ***KernSmooth**: Functions for Kernel Smoothing for Wand and Jones (1995).* Ported to R by Ripley BD. R package version 2.23-3, URL http://CRAN.R-project.org/package=KernSmooth.

WRC Research Systems Inc (2007). ***BrandMap**, Version 7.* WRC Research Systems Inc, Downers Grove, IL, USA. URL http://www.wrcresearch.com/brandmap50.htm.

Yan W, Kang MS (2003). *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists.* CRC Press, Boca Raton, FL, USA.

Yan W, Kang MS (2006). ***GGEbiplot**, Version 5.* URL http://www.ggebiplot.com/.

Young FW (2001). ***ViSta**, Version 6.4.* University of North Carolina, Chapel Hill, NC, USA. URL http://visualstats.org/.

# A. Computational details

This section outlines the main formulae used to produce the biplots of the **BiplotGUI** package. Fuller explanations and derivations may be found in the cited works.

## A.1. General

Let $\mathbf{X} : n \times p$ represent a data matrix. The $n$ samples of $\mathbf{X}$ are to be represented as points in a biplot; the $p$ (numerical) variables are to be represented as calibrated biplot axes. Biplots are displayed in $r$ dimensions. Typically $r$ is taken to be 2 or 3. Biplots with higher values of $r$ are abstractions that cannot be drawn.

## A.2. Data transformations

The data matrix $\mathbf{X}$ is first transformed into a matrix $\widetilde{\mathbf{X}} : n \times p$ on which further calculations are performed. By convention, however, biplot axes are always calibrated in terms of the original variable values, those of $\mathbf{X}$. Six transformations are supported: 'centre', 'centre, scale', 'unitise, centre', 'log, centre', 'log, centre, scale' and 'log, unitise, centre'. The transformations are compound functions. The base functions (log, centre, scale, unitise) are performed in the order in which they appear in the option names. When a matrix $\mathbf{A} : n \times p$ is transformed into a matrix $\mathbf{B} : n \times p$ by taking *logarithms*, $[\mathbf{B}]_{ij} = \log_e([\mathbf{A}]_{ij})$. This requires that all $[\mathbf{A}]_{ij} > 0$. When a matrix $\mathbf{A}$ is transformed into a matrix $\mathbf{B}$ by *centring*, $[\mathbf{B}]_{ij} = [\mathbf{A}]_{ij} - \text{mean}(\mathbf{A}_{(j)})$, where $\mathbf{A}_{(j)}$ is the $j$th column of $\mathbf{A}$ and mean($\mathbf{a}$) returns the mean of the elements of $\mathbf{a}$. When a matrix $\mathbf{A}$ is transformed into a matrix $\mathbf{B}$ by *scaling*, $[\mathbf{B}]_{ij} = [\mathbf{A}]_{ij}/\text{sd}(\mathbf{A}_{(j)})$, where sd($\mathbf{a}$) returns the sample standard deviation of the elements of $\mathbf{a}$. When a matrix $\mathbf{A}$ is transformed into a matrix $\mathbf{B}$ by *unitising*, $[\mathbf{B}]_{ij} = ([\mathbf{A}]_{ij} - \min(\mathbf{A}_{(j)}))/(\max(\mathbf{A}_{(j)}) - \min(\mathbf{A}_{(j)}))$, where max($\mathbf{a}$) and min($\mathbf{a}$) return the maximum and minimum, respectively, of the elements of $\mathbf{a}$.

## A.3. The PCA biplot

**Points:** The normalised eigenvectors corresponding to the $r$ largest eigenvalues of $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ form the columns of a *basis matrix* $\mathbf{V}_r : p \times r$. The samples of $\mathbf{X}$ are represented as points in the display space at coordinates $\mathbf{Y} : n \times r = \widetilde{\mathbf{X}}\mathbf{V}_r$. These are simply the 'scores' of the first $r$ principal components.

**Axes:** All biplot axes pass through the origin. Predictive and interpolative biplot axes coincide in direction. The $j$th *predictive* biplot axis is calibrated $\mu$ at coordinates $(\widetilde{\mu}\mathbf{e}_j^\top\mathbf{V}_r)/(\mathbf{e}_j^\top\mathbf{V}_r\mathbf{V}_r^\top\mathbf{e}_j)$, with $\widetilde{\mu}$ the consistently transformed $\mu$, and with $\mathbf{e}_j$ the $j$th column of the identity matrix $\mathbf{I}_p$. Convenient markers, representative of the $j$th column of $\mathbf{X}$, are variously substituted into $\mu$ to calibrate the entire axis. The $j$th *vector sum interpolative* biplot axis is similarly calibrated $\mu$ at coordinates $\widetilde{\mu}\mathbf{e}_j^\top\mathbf{V}_r$, while the $j$th *centroid interpolative* biplot axis is calibrated $\mu$ at coordinates $\widetilde{\mu}p\mathbf{e}_j^\top\mathbf{V}_r$.

**Goodness-of-fit:** With $\lambda_j$ the $j$th largest eigenvalue of $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$, the *quality* of the representation is given by $(\sum_{j=1}^r \lambda_j)/(\sum_{j=1}^p \lambda_j)$. The *adequacy* of the $j$th biplot axis is $\mathbf{e}_j^\top\mathbf{V}_r\mathbf{V}_r^\top\mathbf{e}_j$. The $r$-dimensional *point predictivities* are given by the diagonal elements of

$$\text{diag}(\widehat{\widetilde{\mathbf{X}}}\widehat{\widetilde{\mathbf{X}}}^\top)\text{diag}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1},$$

with $\widehat{\widetilde{\mathbf{X}}}$ the reconstructed matrix $\widehat{\widetilde{\mathbf{X}}} = \widetilde{\mathbf{X}}\mathbf{V}_r\mathbf{V}_r^\top$. The $r$-dimensional *axis predictivities* are given by the diagonal elements of

$$\text{diag}(\widehat{\widetilde{\mathbf{X}}}^\top\widehat{\widetilde{\mathbf{X}}})\text{diag}(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}.$$

Quality, adequacies and predictivities always lie in the interval $[0, 1]$, with higher values better.

**More:** Full details may be found in Gower and Hand (1996, Chapter 2), and, for point and axis predictivities, in Gardner-Lubbe *et al.* (2008).

## A.4. The covariance/correlation biplot

When $\mathbf{X}$ is transformed into $\widetilde{\mathbf{X}}$ by centring only, the biplot is known as a *covariance* biplot. If in addition $\mathbf{X}$ is transformed by scaling, the biplot is known as a *correlation* biplot.

**Points:** The *basis matrix for interpolation* is given by

$$\mathbf{V}_{r,\text{int}} = (n-1)^{1/2}\mathbf{V}_{r,\text{pca}}\text{diag}(\lambda_1, \ldots, \lambda_r)^{-1/2},$$

where $\lambda_1, \ldots, \lambda_r$ are the $r$ largest eigenvalues of $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$ and where the columns of $\mathbf{V}_{r,\mathrm{pca}}$ contain the corresponding normalised eigenvectors. In the case of both interpolative and predictive biplots, the samples are represented as points in the display space at coordinates $\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{V}_{r,\mathrm{int}}$.

**Axes:** All biplot axes pass through the origin. Predictive and interpolative biplot axes do *not* in general coincide in direction. The *basis matrix for prediction* is

$$\mathbf{V}_{r,\mathrm{pr}} = n^{-1/2}\mathbf{V}_{r,\mathrm{pca}}\mathrm{diag}(\lambda_1, \ldots, \lambda_r)^{1/2}.$$

The $j$th *predictive* biplot axis is calibrated $\mu$ at coordinates $(\widetilde{\mu}\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{pr}})/(\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{pr}}\mathbf{V}_{r,\mathrm{pr}}^\top \mathbf{e}_j)$. The $j$th *vector sum interpolative* biplot axis is calibrated $\mu$ at coordinates $\widetilde{\mu}\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{int}}$, while the $j$th *centroid interpolative* biplot axis is calibrated $\mu$ at coordinates $\widetilde{\mu}p\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{int}}$.

**Goodness-of-fit:** The *quality* of the representation is given by $(\sum_{j=1}^{r} \lambda_j)/(\sum_{j=1}^{p} \lambda_j)$.

**More:** Full details may be found in Gardner (2001, Section 2.3.2). See also Greenacre (1984), Underhill (1990) and Gower and Hand (1996, Section 11.5.1).

## A.5. The CVA biplot

**Points:** In CVA biplots, the $n$ samples are grouped. With $g$ groups of size $n_1, \ldots, n_g$, respectively, the matrix of *group sizes* is given by $\mathbf{N} = \mathrm{diag}(n_1, \ldots, n_g)$. The matrix of *sample group means*, $\bar{\widetilde{\mathbf{X}}} : g \times p$, is calculated from the transformed data matrix $\widetilde{\mathbf{X}}$. The *between groups sums-of-squares-and-crossproducts* matrix is given by

$$\mathbf{B} = \bar{\widetilde{\mathbf{X}}}^\top \mathbf{N}\bar{\widetilde{\mathbf{X}}},$$

while the *within groups sums-of-squares-and-crossproducts* matrix is given by

$$\mathbf{W} = \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} - \bar{\widetilde{\mathbf{X}}}^\top \mathbf{N}\bar{\widetilde{\mathbf{X}}}.$$

The normalised eigenvectors corresponding to the non-increasing eigenvalues of the matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ are placed into the columns of a matrix $\mathbf{V}_{\mathrm{temp}}$. Then with $\mathbf{V} = \mathbf{W}^{-1/2}\mathbf{V}_{\mathrm{temp}}$, the *basis matrix for interpolation*, $\mathbf{V}_{r,\mathrm{int}}$, consists of the first $r$ columns of $\mathbf{V}$. In the case of both interpolative and predictive biplots, the samples are represented as points in the display space at coordinates $\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{V}_{r,\mathrm{int}}$. These are simply the 'scores' of the first $r$ canonical variates.

**Axes:** All biplot axes pass through the origin. Predictive and interpolative biplot axes do *not* in general coincide in direction. The *basis matrix for prediction*, $\mathbf{V}_{r,\mathrm{pr}}$, consists of the first $r$ columns of $(\mathbf{V}^{-1})^\top$. The $j$th *predictive* biplot axis is calibrated $\mu$ at coordinates $(\widetilde{\mu}\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{pr}})/(\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{pr}}\mathbf{V}_{r,\mathrm{pr}}^\top \mathbf{e}_j)$. The $j$th *vector sum interpolative* biplot axis is calibrated $\mu$ at coordinates $\widetilde{\mu}\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{int}}$, while the $j$th *centroid interpolative* biplot axis is calibrated $\mu$ at coordinates $\widetilde{\mu}p\mathbf{e}_j^\top \mathbf{V}_{r,\mathrm{int}}$.

**Goodness-of-fit:** The $r$-dimensional *point predictivities* are given by the diagonal elements of

$$\mathrm{diag}(\widehat{\mathbf{Y}}_g^\top \widehat{\mathbf{Y}}_g)\{\mathrm{diag}(\mathbf{Y}_g^\top \mathbf{Y}_g)\}^{-1},$$

where $\mathbf{Y}_g = (\mathbf{I} - \mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-1}\mathbf{G}^\top)\widetilde{\mathbf{X}}$ and $\mathbf{G}$ is the sample-group indicator matrix which has $i$th row $\mathbf{e}_k$ if sample $i$ belongs to group $k$, $i = 1, \ldots, n$, $k = 1, \ldots, g$. The matrix $\widehat{\mathbf{Y}}_g$ is given by $\widehat{\mathbf{Y}}_g = \mathbf{Y}_g \mathbf{V}_{r,\mathrm{int}}(\mathbf{V}_{r,\mathrm{pr}})^\top$. The $r$-dimensional *group predictivities* are given by the diagonal elements of

$$\mathrm{diag}(\widehat{\widetilde{\mathbf{X}}}^\top \mathbf{W}^{-1}\widehat{\widetilde{\mathbf{X}}})\{\mathrm{diag}(\widetilde{\mathbf{X}}^\top \mathbf{W}^{-1}\widetilde{\mathbf{X}})\}^{-1},$$

with $\widehat{\widetilde{\mathbf{X}}}$ the reconstructed matrix $\widehat{\widetilde{\mathbf{X}}} = \widetilde{\mathbf{X}}\mathbf{V}_{r,\mathrm{int}}(\mathbf{V}_{r,\mathrm{pr}})^\top$. The $r$-dimensional *axis predictivities* are given by the diagonal elements of

$$\mathrm{diag}(\hat{\widetilde{\mathbf{X}}}^\top \mathbf{N}\hat{\widetilde{\mathbf{X}}})\{\mathrm{diag}(\widetilde{\mathbf{X}}^\top \mathbf{N}\widetilde{\mathbf{X}})\}^{-1}.$$

**More:** Full details may be found in Gower and Hand (1996, Chapter 5), and, in the case of point, group and axis predictivities, in Gardner-Lubbe *et al.* (2008).

## A.6. Dissimilarity metrics

One of four dissimilarity metrics can be used to calculate an inter-sample dissimilarity matrix $\mathbf{D} : n \times n$ from the transformed data matrix $\widetilde{\mathbf{X}}$. Under the *Pythagoras* dissimilarity metric, $[\mathbf{D}]^2_{ii'} = d^2_{ii'} = (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_{i'})^\top(\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_{i'})$, with $\widetilde{\mathbf{x}}_i$ the $i$th row of $\widetilde{\mathbf{X}}$. Under the *Square-root-of-Manhattan* dissimilarity metric, $d^2_{ii'} = \sum_{j=1}^p |\widetilde{x}_{ij} - \widetilde{x}_{i'j}|$. According to *Clark*'s (1952) dissimilarity metric, $d^2_{ii'} = \sum_{j=1}^p \left((\widetilde{x}_{ij} - \widetilde{x}_{i'j})/(\widetilde{x}_{ij} + \widetilde{x}_{i'j})\right)^2$. Finally, under the *Mahalanobis* (1936) dissimilarity metric, $d^2_{ii'} = (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_{i'})^\top \mathbf{S}^{-1}_{\widetilde{\mathbf{X}}}(\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_{i'})$, with $\mathbf{S}_{\widetilde{\mathbf{X}}}$ the sample covariance matrix of $\widetilde{\mathbf{X}}$. Note that the term ‘Pythagoras’ is used instead of ‘Euclidean’ in order to avoid confusion with the term ‘Euclidean-embeddable’ of the next section.

**More:** See also Cox and Cox (2001, Section 1.3) and Borg and Groenen (2005, Section 6.3).

## A.7. PCO

**Points:** An inter-sample dissimilarity matrix $\mathbf{D}$ is calculated from $\widetilde{\mathbf{X}}$ according to one of the four dissimilarity metrics of Appendix A.6. All four dissimilarity metrics are Euclidean-embeddable (Gower 1982), as required by PCO. The *sums-of-squares-and-crossproducts* matrix is given by

$$\mathbf{B} = (\mathbf{I} - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{D}(\mathbf{I} - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top).$$

The $r$ largest eigenvalues of $\mathbf{B}$, $\lambda_1, \ldots, \lambda_r$, are taken to form the diagonal elements of matrix $\mathbf{\Lambda} : r \times r$. The corresponding normalised eigenvectors are placed into the columns of matrix $\mathbf{V}_r : n \times r$. The samples are then represented as points in the display space at principal coordinates $\mathbf{Y} = \mathbf{V}_r\mathbf{\Lambda}^{1/2}$. PCO is also known as ‘classical scaling’.

**Goodness-of-fit:** The *quality* of the representation is given by $(\sum_{j=1}^r \lambda j)/(\sum_{j=1}^p \lambda_j)$. A Shepard diagram may also be drawn.

**More:** Full details may be found in Gower and Hand (1996, Section A.5.2) and Cox and Cox (2001, Section 2.2). The Shepard diagram is described by Borg and Groenen (2005, Section 3.3).

## A.8. MDS

**Points:** An inter-sample dissimilarity matrix $\mathbf{D}$ is calculated from $\widetilde{\mathbf{X}}$ according to one of the four dissimilarity metrics of Appendix A.6. The samples are represented in the display space at coordinates $\mathbf{Y} : n \times r$, with $\mathbf{Y}$ found algorithmically to minimise the *stress* criterion

$$S(\hat{\mathbf{d}}, \mathbf{Y}) = \sum_{i<i'} (\hat{d}_{ii'} - \delta_{ii'}(\mathbf{Y}))^2.$$

In this expression, $\hat{d}_{ii'}$ represents the inter-sample *disparity* between sample $i$ and $i'$. Inter-sample disparities are derived from the inter-sample dissimilarities $\mathbf{D}$ by one of three transformations: the identity transformation, monotone regression, or a monotone spline transformation. The quantity $\delta_{ii'}(\mathbf{Y})$, in turn, represents the Pythagorean distance between rows $i$ and $i'$ of $\mathbf{Y}$, called the inter-point *distance*. An IM algorithm is used to find the minimising matrix $\mathbf{Y}$. The algorithm converges uniformly to a local minimum, although in theory a saddle-point cannot be ruled out. In practice, the algorithm is taken to have converged as soon as the relative decrease in stress becomes smaller than some pre-set value. The algorithm is also stopped as soon as a certain maximum number of iterations has been performed.

**Goodness-of-fit:** A Shepard diagram may be drawn.

**More:** For more details and a full description of the implementation of the IM algorithm, see Borg and Groenen (2005, Chapters 8, 9). The Shepard diagram is described by Borg and Groenen (2005, Section 3.3).

## A.9. The regression biplot

**Points:** The samples are represented as points with coordinates $\mathbf{Y} : n \times r$. The coordinates are typically determined by PCO or MDS, but any scaling method can be used.

**Axes:** The *basis matrix* is given by $\mathbf{V}_r = (\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\widetilde{\mathbf{X}}$. All other details are the same as for the axes of the PCA biplot.

**More:** Full details may be found in Gower and Hand (1996, Sections 3.3.2, 3.4.3).

## A.10. The Procrustes biplot

**Points:** The samples are represented as points with coordinates $\mathbf{Y} : n \times r$. The coordinates are typically determined by PCO or MDS, but any scaling method can be used. When the points are determined by PCO based on Pythagorean dissimilarities, the PCA, regression and Procrustes biplots coincide.

**Axes:** All biplot axes pass through the origin. Predictive and interpolative biplot axes do *not* in general coincide in direction. By the method of orthogonal Procrustes analysis, the *basis matrix for prediction*, $\mathbf{V}_{r,\mathrm{pr}}$, is given by the first $r$ columns of $\mathbf{B}\mathbf{A}^\top$, where $\widetilde{\mathbf{X}}^\top\mathbf{Y} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top$ is the singular value decomposition of $\widetilde{\mathbf{X}}^\top\mathbf{Y}$. The $j$th *predictive* biplot axis is calibrated $\mu$ at coordinates $(\widetilde{\mu}\mathbf{e}_j^\top\mathbf{V}_{r,\mathrm{pr}})/(\mathbf{e}_j^\top\mathbf{V}_{r,\mathrm{pr}}\mathbf{V}_{r,\mathrm{pr}}^\top\mathbf{e}_j)$. The *basis matrix for interpolation*, $\mathbf{V}_{r,\mathrm{int}}$, minimises $\|\mathbf{Y}-\widetilde{\mathbf{X}}\mathbf{V}\|$ out of all projection matrices $\mathbf{V}$ with orthonormal columns. The solution is found by minimal error projection Procrustes. The $j$th *vector sum interpolative* biplot axis is calibrated $\mu$ at coordinates $\widetilde{\mu}\mathbf{e}_j^\top\mathbf{V}_{r,\mathrm{int}}$, while the $j$th *centroid interpolative* biplot axis is calibrated $\mu$ at coordinates $\widetilde{\mu}p\mathbf{e}_j^\top\mathbf{V}_{r,\mathrm{int}}$.

**More:** Full details may be found in Gower and Hand (1996, Sections 3.3.1, 3.4.2). For more on orthogonal Procrustes analysis, see Gower and Hand (1996, Appendix A.10.1). The algorithm for minimal error projection Procrustes is described in Gower and Dijksterhuis

(2004, p. 57). See also Gower and Hand (1996, Appendix A.10.2) and Cox and Cox (2001, Section 5.2.2).

### A.11. The circular non-linear biplot

**Points:** The samples are represented as points with coordinates $\mathbf{Y} : n \times r$ determined by PCO. The dissimilarity metric is assumed to be additive (Gower and Hand 1996, p. 105), as all the dissimilarity metrics of Appendix A.6 are besides Mahalanobis. Matrices $\mathbf{B}$ and $\mathbf{\Lambda}$ are carried forward from PCO. Define $[\mathbf{E}]_{ii'} = -\frac{1}{2}[\mathbf{D}]^2_{ii'}$, with $\mathbf{D}$ the dissimilarity matrix.

**Axes:** The biplot axes can be chosen to pass through a common point. Predictive and interpolative biplot axes do *not* in general coincide. To find the position of marker $\mu$ on the $j$th *vector sum interpolative* biplot axis, the vector $\mathbf{d}_{n+1} : n \times 1$ is taken to contain the dissimilarities, divided by $-2$, between $\widetilde{\mu}\mathbf{e}_j$ and the $n$ samples of $\widetilde{\mathbf{X}}$. The axis is calibrated $\mu$ at the first $r$ coordinates of $\mathbf{y} = \mathbf{\Lambda}^{-1}\mathbf{Y}^\top(\mathbf{d}_{n+1} - \frac{1}{n}\mathbf{E1})$. The $j$th *centroid interpolative* biplot axis is calibrated $\mu$ at the first $r$ coordinates of $\mathbf{y} = p\mathbf{\Lambda}^{-1}\mathbf{Y}^\top(\mathbf{d}_{n+1} - \frac{1}{n}\mathbf{E1})$. The construction of predictive non-linear biplot axes is more complicated and is described in Gower and Hand (1996, Section 6.3.2).

**More:** Full details may be found in Gower and Hand (1996, Chapter 6). See also Gower and Harding (1988) and Gower and Ngouenet (2005).

## B. Setup

R can be downloaded from CRAN at `http://CRAN.R-project.org/` for any of the three major platforms: Linux, MacOS X and Windows. At the time of writing, the latest version of R is 2.10.1. At present, **BiplotGUI** is intended to be run under Windows. Future releases will be more general.

The **BiplotGUI** package and its dependencies can most easily be downloaded and installed into R by issuing the following command in the R console:

```
R> install.packages("BiplotGUI")
```

**BiplotGUI** depends on the following R packages: **colorspace** (Ihaka *et al.* 2009), **deldir** (Turner 2009), **KernSmooth** (Wand 2009), **MASS** (Venables and Ripley 2002), **rgl** (Adler and Murdoch 2009), **tcltk** (R Development Core Team 2009), **tcltk2** (Grosjean 2009) and **tkrplot** (Tierney 2008).

**BiplotGUI** runs slightly better when the R console is in SDI mode, rather than MDI mode.

## C. Known issues

- The graphs in the diagnostic tabs occassionally do not show when the tabs are opened. Click to another tab and back.
- Interpolative circular non-linear biplots do not always initialise correctly in 3D. If this happens, first display a 3D predictive circular non-linear biplot.
- Extensive use of the GUI leads to memory leaks. If the system becomes noticably slower, close the R console after saving, and re-open.

**Affiliation:**

Department of Genetics
Stellenbosch University
South Africa
amlg@sun.ac.za