Biplot and Singular Value Decomposition Macros for Excel©

Ilya Lipkovich and Eric P. Smith
Department of Statistics
Virginia Tech
Blacksburg, VA 24061-0439
Email: epsmith@vt.edu

Revised: June 6, 2002

Abstract

The biplot display is a graph of row and column markers obtained from data that forms a two-way table. The markers are calculated from the singular value decomposition of the data matrix. The biplot display may be used with many multivariate methods to display relationships between variables and objects. It is commonly used in ecological applications to plot relationships between species and sites. This paper describes a set of Excel© macros that may be used to draw a biplot display based on results from principal components analysis, correspondence analysis, canonical discriminant analysis, metric multidimensional scaling, redundancy analysis, canonical correlation analysis or canonical correspondence analysis. The macros allow for a variety of transformations of the data prior to the singular value decomposition and scaling of the markers following the decomposition.

1. Introduction

The biplot display is a commonly used multivariate method for graphing row and column elements using a single display (Gabriel, 1971). The method has been used to display objects and variables on the same graph in principal components analysis, row and column factors in correspondence analysis of two-way contingency tables and to detect interaction in two-way analysis of variance tables (Gower and Hand, 1996). Biplot displays are commonly used in the analysis of data from ecological and environmental studies. Data are often collected on the abundance of species at various sites. Interest is in describing the data and the biplot display provides a method for reducing the dimensionality of the data and displaying the species and sites jointly on the same plot. Similarities between species or sites may be gleaned from these types of plots. Also it is common to interpret the axes in the biplot and treat the coordinates as scores on these axes. For example, in an ecological analysis the first axis might represent a moisture gradient while the second a temperature gradient. Species or sites may then be ranked in terms of tolerance to moisture or temperature. Examples of the display and interpretation guidelines are given in Legendre and Legendre (1998) with a focus on ecological data or Gower and Hand (1996) for more general applications. A detailed numerical illustration is presented in Digby and Kempton (1987).

The basis of the display is the singular value decomposition. If we have an n by p matrix $\mathbf{Y}$ of rank r with $r \leq p \leq n$, then the matrix may be decomposed as (Seber, 1984, pg. 504)

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$$

where **U** (n by p) and **V** (p by p) are matrices of singular vectors and $\mathbf{\Lambda}$ (p by p) is a diagonal matrix of singular values. **U** is the matrix with columns corresponding to the p orthogonal eigenvectors of $\mathbf{YY'}$ and **V** is the orthogonal matrix corresponding to the eigenvectors of $\mathbf{Y'Y}$. $\mathbf{\Lambda}$ =diag(λ_1, λ_2, …, λ_p) with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_m = 0$. The singular values are the positive square roots of the eigenvalues of $\mathbf{Y'Y}$.

Typically the matrix is approximated using the first few dominant singular values and vectors. The biplot display is the plot of the row markers **G** and column markers **H** where

$$\mathbf{G} = \mathbf{U}_{(k)}\mathbf{\Lambda}_{(k)}^{\alpha}$$
$$\mathbf{H} = \mathbf{V}_{(k)}\mathbf{\Lambda}_{(k)}^{1\text{-}\alpha}$$

The value of k determines the dimension of the approximation (typically k=2) and the user specified parameter α ($0 \leq \alpha \leq 1$) determines whether emphasis is placed on the rows or columns of **Y**. The matrix **Y** is then approximated by $\hat{\mathbf{Y}}_{(k)} = \mathbf{GH'} = \mathbf{U}_{(k)}\mathbf{\Lambda}_{(k)}^{\alpha}\mathbf{\Lambda}_{(k)}^{1\text{-}\alpha}\mathbf{V}_{(k)}^{'} = \mathbf{U}_{(k)}\mathbf{\Lambda}_{(k)}\mathbf{V}_{(k)}^{'}$. Plots of the coordinates associated with **G** superimposed over the coordinates associated with **H** form the biplot display.

Although many values of α are possible, three are commonly used, 1, ½, and 0. When the value 1 is selected, the result is called a JK or RMP (row metric preserving) biplot. In this display the distances between pairs of rows is preserved (after any centering and scaling is performed) and the display is useful for studying objects (for example interpreting distance matrices). When the value 0 is selected, the result is a GH or CMP (column metric preserving) biplot. This display preserves distances between the columns and is useful for interpreting variance and relationships between variables (for example, interpreting a covariance or correlation matrix). The other value of α, ½, gives equal scaling or weight to the rows and columns. It is useful for interpreting interaction in two factor experiments (Gower and Hand, 1996). In correspondence analysis, the terminology principal coordinates is sometimes used for the coordinates that are weighted by the singular values and standard coordinates are the unweighted coordinates (Greenacre, 1984).

The singular value decomposition may be used as a basis for many multivariate graphical techniques. Standard applications include principal components analysis and correspondence analysis. In addition, biplot displays may be used as summary displays in canonical discriminant analysis, metric multidimensional scaling, redundancy analysis, canonical correlation analysis and canonical correspondence analysis. While there are many statistical packages for running these analyses we have found that students (particularly undergraduate students) have difficulties with the standard statistical packages and understanding of the procedures. In addition, complete graphical packages for biplot displays are not available in many packages or only partially available. To aid in the educational process we have written a user-friendly add-in for Excel© that carries out a singular value decomposition (SVD) for a two-way matrix of data and then plots the results using a biplot display. Our goal is to provide an educational tool that will be useful for an undergraduate multivariate course as well as a graduate level applied multivariate course. In addition, the program is helpful for researchers who desire high quality, editable graphics. Our focus has been toward students and researchers in the fisheries, wildlife and ecological areas although the program is applicable in other areas. These researchers use canonical correspondence analysis to study relationships between the abundance of different organisms and environmental conditions. However, users in other fields may find the program useful for producing biplot displays for methods such as principal components analysis (PCA) and correspondence analysis (CA).

The BIPLOT add-in for Excel© is implemented in the Visual Basic for Applications macro language. The add-in requires Excel© 97 or a more recent version of Excel© to function properly. The program will calculate singular value decompositions of the data matrix (or transformed data matrix) and produce a standard biplot display as in principal components analysis or correspondence analysis. In addition, the program also produces displays for the other analyses mentioned above. The steps in adding the macros to Excel© are fairly simple and are described below along with some computational details and descriptions of options.

The macro is stored in an add-in file BIPLOT01.XLA that can be added to the Excel© environment. To add the macro to Excel©, open Excel©, then navigate to TOOLS>ADD-INS. When the add-in window appears, click BROWSE and find the program named BIPLOT01.XLA. It should appear now in the add-in window. Click the box to add it to your set of Excel© macros. If successful, the item **Biplot** will appear on your Excel© menu at the top of the spreadsheet.

Selecting **Biplot** will open a pull down menu that has two main pieces, one to do calculations and the other for plotting. It also has a help menu and an information window. The main calculations are done after selecting options using a *singular value decomposition* (SVD) dialog. A *graphics selection* dialog is then used to choose the type of biplot to display and to specify graphics options.

2. The Singular Value Decomposition Macro (SVD Macro)

Data layout

The data is entered as a two-way array in Excel©. For canonical correlation analysis and canonical correspondence analysis there will be two arrays. It is recommended that data be in the form of a two-way table with row names and column names. In a typical analysis, the rows will correspond to the objects and columns to the variables. Here is a simple example using data collected at 10 sites for four species:

Site	Sp1	Sp2	Sp3	Sp4
S1	1	0	0	0
S2	0	0	0	0
S3	0	1	0	0
S4	11	4	0	0
S5	11	5	17	7
S6	9	6	0	0
S7	9	7	13	10
S8	7	8	0	0
S9	7	9	10	13
S10	5	10	0	0

Species names are in row 1 and site names are in column1 of the data table. Names are not required but are useful in the graphical displays as identifiers. You can have this table anywhere in the spreadsheet. For example, you can have a title above the table and other data below the table. There should be no other columns between the row names and the numerical values that will be analyzed. The data analyzed forms a table and the table is selected so remove any columns or rows that are not to be included in the analysis. There is an option to remove some columns based on certain criteria but there is not option for selecting individual columns or rows to be analyzed.

Sometimes the data will contain the squared symmetric matrix of inner products based on some underlying *n* by *p* matrix. In fact, for MDS (multidimensional scaling) this may be the most typical format since the data often are dissimilarities (or proximities) among stimuli. The data below gives an example of a so-called confusion matrix; here it is percentage of times that the pairs of Morse code signals for two numbers were declared to be the same by 598 subjects (Rothkopf, 1957)

Signal	.----	..---	…--	….-	…..	-….	--…	---..	---.	-----
.----	84	62	16	6	12	12	20	37	57	52
..---	62	89	59	23	8	14	25	25	28	18
…--	16	59	86	38	27	33	17	16	9	9
….-	6	23	38	89	56	34	24	13	7	7
…..	12	8	27	56	90	30	18	10	5	5
-….	12	14	33	34	30	86	65	22	8	18
--…	20	25	17	24	18	65	85	65	31	15
---..	37	25	16	13	10	22	65	88	58	39
---.	57	28	9	7	5	8	31	58	91	79
-----	52	18	9	7	5	18	15	39	79	94

The help file includes a biplot display for this data.

NOTE: At present all entries in the table must be filled. Missing values are not allowed.

The SVD Macro

The computation of singular values and vectors is accomplished through the power method (see for example Heath, 1997). The steps involved are given below.

Algorithm for singular value decomposition $\mathbf{Y} = \mathbf{U\Lambda V'}$:

1. Read the data in matrix **Y** (*note this may have been standardized using the transformation options*)

2. Compute $\mathbf{R}^{(1)} = \mathbf{R} = \mathbf{Y'Y}$ unless *data contains inner product* option is checked; if that is the case, set $\mathbf{R}^{(1)} = \mathbf{R} = \mathbf{Y}$. The latter case can be also used when the user needs to obtain the spectral decomposition of a symmetric data matrix. This becomes relevant for principal coordinates analysis of data on similarities or dissimilarities (see the next section). If this is the case, the primary goal of the analysis are not the singular values but the eigenvalues of **Y** and the reported "singular values" are defined as square roots of the absolute values of eigenvalues of **Y**. They can be interpreted as the singular values associated with the SVD of the underlying **X** such that $\mathbf{X'X} \approx \mathbf{Y}$.

3. Find the first absolute maximum eigenvalue of $\mathbf{R}^{(j)}$ using the *power method*: start with a random vector $\mathbf{x}_o$, then iteratively compute $\mathbf{w}_i = \mathbf{R}^{(j)} \boldsymbol{x}_{i\text{-}1}$, $\boldsymbol{x}_i = \mathbf{w}_i /\| \mathbf{w}_i \|$ (where $\| \mathbf{w}\|$ is the squared Euclidean norm of vector $\mathbf{w}$) until the relative norm $L = \| \mathbf{x}_i - \mathbf{x}_{i\text{-}1}\|/ \| \mathbf{x}_{i\text{-}1}\| < 10^{-10}$ or abs(L-2)<10^{-10}. The latter condition corresponds to the case of a negative eigenvalue. Then the direction of $\mathbf{x}_i$ switches from iteration to iteration and L tends to 2. The maximum number of iterations is set to 500; if the algorithm does not converge, the program terminates and prints out a report that would contain all elements of matrices **U** and **V** that were determined at the previous steps.

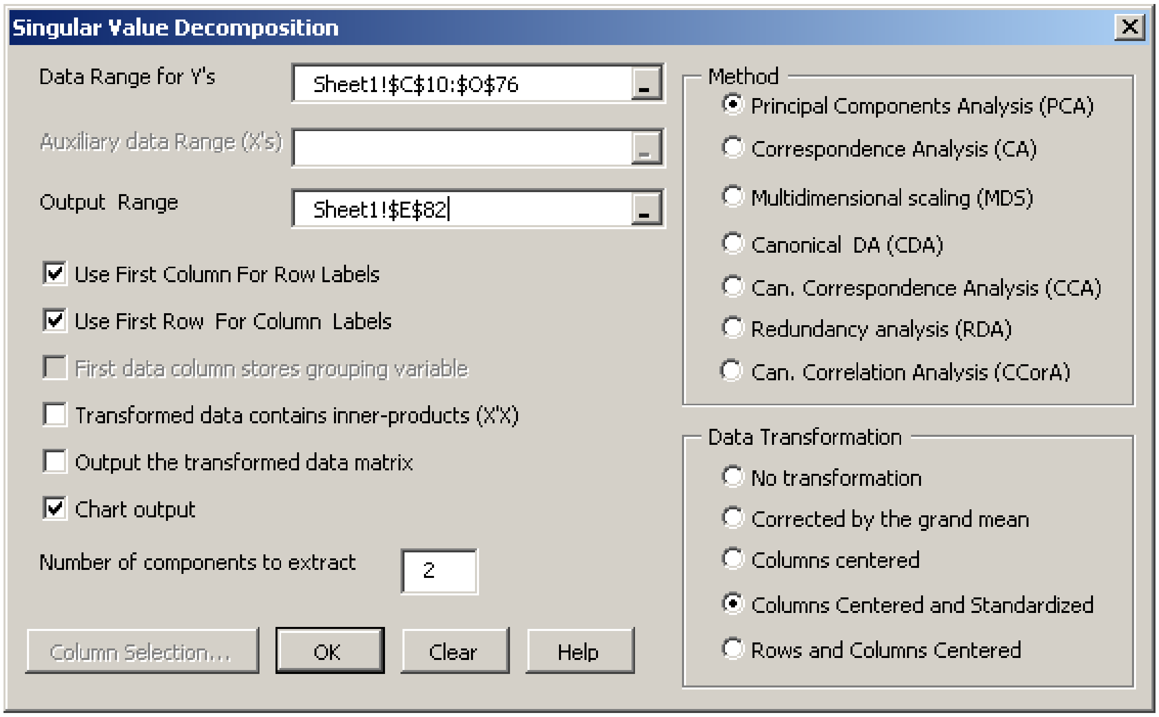
4. Set $\mathbf{v}_j = \mathbf{x}_i$ from the i^{th} iteration of the previous step, form the *j-t*h maximum absolute eigenvalue of **R**, $d_j = \mathbf{v}_j'\mathbf{R}^{(j)} \mathbf{v}_j$, compute the maximum singular value of **Y** as $\lambda_j = \sqrt{abs(d_j)}$, set $\mathbf{u}_j = \mathbf{Y}\mathbf{v}_j / \lambda_j$. Assign the appropriate sign to the singular value. (Actually it can be always taken as positive. Indeed, if it is negative, then the sign of **u** will also change. Therefore there is indeterminacy in the signs of singular values and singular vectors, and the former can always be positive.)

5. Compute the residual matrix $\mathbf{R}^{(j+1)} = \mathbf{R}^{(j)} - d_j\mathbf{u}_j\mathbf{v}_j'$ and repeat steps 2, 3, 4 until all $r = rank(\mathbf{R})$ eigenvalues of matrix **R** are found or until the algorithm fails to converge. The rank of R is the number of nonzero eigenvalues; once the absolute value of an eigenvalues is smaller than the tolerance = 10^{-10}, the algorithm terminates. If the user requests that only a certain number of eigenvalues be extracted, the algorithm terminates after this condition has been met.

6. Print out the report that contains (1) the transformed *n* by *p* data matrix **Y**, if a data transformation was selected, (2) The *p* by *r* matrix **V** for the SVD, (3) the *n* by *r* matrix **U** for the SVD, (4) the singular values and eigenvalues associated with vectors **u** and **v** are printed in descending order, based on the eigenvalues.

The SVD is calculated after options are selected in the SVD dialog box. The box is used to specify locations for data and output, choices for data transformation and method of analysis, and details about the data array. An example of the set of selections chosen for a classical PCA is given below. Here the data set is located in the Excel© Sheet1 with row and column labels as the first row and column, respectively. The data are standardized before the principal components are extracted. The solution will contain results for the first two principal axes. A description of the possible selections is given below.

Singular Value Decomposition Dialog



Data Range for Y's: Sheet1!$C$10:$O$76

Auxiliary data Range (X's):

Output Range: Sheet1!$E$82

☑ Use First Column For Row Labels

☑ Use First Row For Column Labels

□ First data column stores grouping variable

□ Transformed data contains inner-products (X'X)

□ Output the transformed data matrix

☑ Chart output

Number of components to extract: 2

Method

- ● Principal Components Analysis (PCA)
- ○ Correspondence Analysis (CA)
- ○ Multidimensional scaling (MDS)
- ○ Canonical DA (CDA)
- ○ Can. Correspondence Analysis (CCA)
- ○ Redundancy analysis (RDA)
- ○ Can. Correlation Analysis (CCorA)

Data Transformation

- ○ No transformation
- ○ Corrected by the grand mean
- ○ Columns centered
- ● Columns Centered and Standardized
- ○ Rows and Columns Centered

Dialog Options:

Data Range for Y's: The area on the sheet that contains the data matrix. No missing data are allowed. Simply click on ▭ at the end of the text box to toggle to the data sheet. Select (highlight) the data table to enter elements. Note that the data range includes the row /column labels, if provided.

Auxillary Data range: Use this to set the X variables in canonical correlation/canonical correspondence analysis/redundancy analysis. Note that the first row of the **X** data set should contain labels if the first row of the Y set contains the column labels, however the row labels can be only in the first column of the Y set.

Output Range: Click on the box at the end to toggle to the data file. Then click on a cell in the sheet to define the place of the upper left corner of the output report.

Use first column for row labels: Generally select this – it allows use of row labels (objects or sites) in the graphical displays.

Use first row for column labels: Generally select this – it allows use of column labels (variables or species) in the graphical displays.

Data matrix contains inner-products: When this option is checked, the macro will apply the SVD to the original data matrix. This is useful when the data are already in **X'X** form or when the data can be interpreted as dissimilarities/ similarities, then the SVD can be used to plot the principal coordinates (MDS).

Output the transformed data matrix: Generally this is not used. It allows you to print out the transformed data values.

Chart output: Generally select this as it facilitates the charting of the results; it is needed for automatic selection of columns and rows to plot in the biplot display.

Method

Principal Components Analysis: First the data are transformed using the transformation selected from the data transformation section. Then the SVD algorithm described in the previous section is applied to the transformed data. Depending on the type of transformation the user may obtain a

- PCA of the covariance matrix, when the "columns centered" option is selected
- PCA of the correlation matrix, when the "columns centered and standardized" option is selected
- PCA of two –way interactions, when the "rows and columns centered" option is selected
- Principal Coordinates Analysis of Similarities, when the "rows and columns centered" option is selected along with "transformed data contains cross-products" option.

Correspondence Analysis: First the data are transformed into $z_{ij} = \dfrac{y_{ij} - y_{i\bullet} y_{\bullet j} / y_{\bullet\bullet}}{\sqrt{y_{i\bullet} y_{\bullet j}}}$, where $y_{i\bullet}, y_{\bullet j}, y_{\bullet\bullet}$ are totals for the *i*-th row, *j*-th column and the entire data, respectively. Note that this transformation requires that all column/row totals are greater than zero, or the program will give you an error message. Therefore you should delete all columns and rows with zero totals before invoking the macro, if you have any. This transformation is useful when the data matrix is a contingency table whose entries are frequencies. The transformed z_{ij} is proportional to the square root of the cell's contribution to the Pearson's χ^2 statistic for the independence of row and column classifications. Using the SVD of this data, a classical *Correspondence Analysis* (*CA*) can be performed by plotting $(a_i^{(1)}, a_i^{(2)}) = (u_i^{(1)}\lambda_1 / \sqrt{y_{i\bullet}},\ u_i^{(2)}\lambda_2 / \sqrt{y_{i\bullet}})$ for the row markers and $(b_j^{(1)}, b_j^{(2)}) = (v_j^{(1)}\lambda_1 / \sqrt{y_{\bullet j}},\ \mathrm{v}_j^{(2)}\lambda_2 / \sqrt{y_{\bullet j}})$ for the column markers. This scaling does not preserve inner products (as the singular values appear in both the row and column markers). In this program we use the standard *JK*, *GH*, and *SYM* biplots (preserving the respective inner product approximations; see Gower and Hand, 1996, *p* 180). The classical plot may be obtained by first computing the coordinates in Excel© then using the plotting feature.

Multidimensional Scaling: First, the data, y_{ij}, are transformed to $z_{ij} = -0.5 y_{ij}^2$ then the **double centering** transformation is applied to produce $z_{ij}^* = z_{ij} - \bar{z}_{i\bullet} - \bar{z}_{\bullet j} + \bar{z}_{\bullet\bullet}$. When the original data represent distances d_{ij} for an underlying matrix **X**, the transformed data will be the centered inner product matrix $z_{ij}^* = (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$ and the SVD of this matrix will produce a Principal Coordinate Analysis. The plotted values are the principal coordinates, $(\mathbf{v}_1\lambda_1, \mathbf{v}_2\lambda_2)$, where are the square roots of the eigenvalues of **Z*** and **v**'s are the eigenvectors of **Z***.

Canonical DA: This option will allow you to do canonical discriminant analysis (CDA) and produce the associated biplot. Note that the grouping variable should be stored in the column after any labels, i.e. in the first data column (that is the second or first column in the data range depending on whether the row labels in the first column are provided or not). The group identifiers may be in either numeric or textual format.

The output will contain the following information:

> *Raw Coefficients of discriminant functions:* If CDA is used, the eigenvectors **V** of the matrix $\mathbf{E}^{-1}\mathbf{H}$ (**E** and **H** are the matrices of within and between sums of squares matrices). The first 2 columns of **V** are used as the coordinates for the column markers of the biplot
>
> *Canonical scores:* $(\mathbf{Y}-\overline{\mathbf{Y}})\mathbf{V}$, can be used as the coordinates for the row markers, however note that for CDA it is more useful to plot the coordinates of the group means, which are produced in a separate table. The user may choose to plot both the canonical scores for the observations and the group means on same chart.

Canonical correspondence analysis:
Allows for CCA, a multivariate technique similar to RDA (redundancy analysis) that is based on the SVD of the fits of the appropriately transformed **Y's** regressed on columns of the standardized **X** matrix (details of algorithm are given in Legendre and Legendre, 1999). A typical application is when **Y** contains so-called abundance data (sites in rows and species in columns), and **X** contains some site level environmental data. The data are first transformed to **Y*** and **X***, where

$$y_{ij}^{*} = \frac{y_{ij} - y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}}{\sqrt{y_{i\bullet}y_{\bullet j}}} \quad \text{and} \quad x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{s_{wj}}$$

where $y_{i\bullet}, y_{\bullet j}, y_{\bullet\bullet}$ are totals for the *i*-th row, *j*-th column and the entire abundance matrix, respectively and columns of $\mathbf{X}^{*}$ are centered and standardized using *column mean* and *standard deviation* (s_{wj}), weighted by the row totals divided by the total of the abundance matrix **Y***, i.e.*

$$\overline{x}_j = \sum_{i=1}^{r} x_{ij} y_{i\bullet} / y_{\bullet\bullet}\ ,\ s_{wj} = \sqrt{\sum_{i=1}^{r} (x_{ij} - \overline{x}_j)^2 y_{i\bullet} / y_{\bullet\bullet}}\ , r \ \text{ is the number of rows.}$$

Note that this transformation requires that all column/row totals be greater than zero, or the program will give you an error message. Therefore you have to delete all columns and rows with zero totals before invoking the macro, if you have any.

Denote $\mathbf{X}_w$=$\mathbf{R}^{1/2}\mathbf{X}^{*}$, (that is rows of $\mathbf{X}^{*}$ are weighted by the row totals of **Y**, the elements of the diagonal matrix **R**). Then we obtain $\hat{\mathbf{Y}} = \mathbf{X}_w\mathbf{B}$, where **B** is the matrix of regression coefficients obtained by applying the multivariate OLS regression to **Y*** and $\mathbf{X}_w$, and

$$\mathbf{B} = (\mathbf{X}^{*\prime}\mathbf{R}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\prime}\mathbf{R}^{1/2}\mathbf{Y}^{*} = (\mathbf{X}_w^{\prime}\mathbf{X}_w)^{-1}\mathbf{X}_w^{\prime}\mathbf{Y}^{*}$$

Notice that the described procedure differs from the traditional weighted least squares in that only the *x*'s are weighted in weighted least squares, and the fitted values are obtained by applying the regression coefficients to the weighted $\mathbf{X}_w$.

Another way of looking at this procedure is by writing $\hat{\mathbf{Y}} = \mathbf{H}_w\mathbf{Y}^*$, where $\mathbf{H}_w = \mathbf{X}_w(\mathbf{X}_w^{'}\mathbf{X}_w)^{-1}\mathbf{X}_w^{'}$ is the idempotent projection matrix based on the standardized and weighted X's. The SVD is applied to the matrix of fitted values $\hat{\mathbf{Y}}$ to produce $\hat{\mathbf{Y}}$ =**UΛV'**.

The output contains the rescaled coordinates for rows of **Y** (sites), columns of **Y** (species), and columns of **X** (environmental variables), calculated as follows:
Site coordinates (in the space of **Y**): $\mathbf{U}^* = \mathbf{Y}^*\mathbf{V}\mathbf{\Lambda}^{-1}$

Fitted site coordinates (in the space of **X**): $\mathbf{U}_f^* = \mathbf{R}^{-\frac{1}{2}}\mathbf{U}$

Species coordinates, $\mathbf{V}^* = \mathbf{C}^{-1/2}\mathbf{V}$, where **C** is the diagonal matrix whose elements are the column totals of **Y**.

Either **U*** or **U**$_f$***** *fitted* can be paired with **V*** to form a biplot using any of the scaling options (JK, GH, or SYMM). The default is the GH biplot which would post multiply **V*** by the diagonal matrix **Λ.**

Note that the **X** variables can be plotted simultaneously on same plot, resulting in a *Triplot*. Coordinates for the **X** variables are formed from the *weighted correlations* with the *fitted site scores.* For additional information see Ter Braak (1986) or Legendre and Legendre (1998).

Redundancy analysis (RDA) *:*
When this option is selected a redundancy analysis using **Y** and **X** is performed. RDA is based on a SVD of the fits of the centered **Y*** regressed on columns of the centered **X*** matrix. A typical application is when **Y** contains logs of abundance data (sites in rows and species in columns), and **X** contains some site level environmental data. The data are first transformed into **Y*** and **X*** according to one of the available options (centered or centered and standardized). Then the fitted values for multivariate regression $\hat{\mathbf{Y}} = \mathbf{X}^*\mathbf{B}$ are computed, where **B** is the matrix of regression coefficients of **Y*** on **X*** obtained by applying the ordinary least squares regression, i.e. $\mathbf{B} = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$. The SVD is then applied to the matrix of fitted values $\hat{\mathbf{Y}}$ to produce $\hat{\mathbf{Y}} = \mathbf{U\Lambda V}^{'}$**.**

The output contains the rescaled coordinates for rows of **Y** (sites), columns of **Y** (species), and columns of **X** (environmental variables).

Site coordinates (in the space of **Y**) are given by $\mathbf{U}^* = \mathbf{Y}^*\mathbf{V}^{-1}\mathbf{\Lambda}$

Fitted site coordinates (in the space of **X**) are given by $\mathbf{U}$ and species coordinates by **V**. Either **U*** or **U** can be paired with **V** to form a biplot. The default is the JK biplot which would post multiply **U*** or **U** by the diagonal matrix **Λ** of singular values

Note that the **X** variables can be plotted simultaneously on same display thus making it a *Triplot*. Coordinates for **X** variables are formed from their *correlations* with the *fitted site scores* **U.** Additional details are given in Ter Braak (1994).

Canonical correlation analysis:

Selecting this produces a canonical correlation analysis. The data consists of two sets of variables **Y** and **X**. Variables in both sets are centered and standardized. Classical canonical correlation analysis is based on eigenvalues and eigenvectors of the product of the correlation matrices $\mathbf{R}_{\mathbf{YY}}^{-1}\mathbf{R}_{\mathbf{YX}}\mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{R}_{\mathbf{XY}}$. The canonical correlation biplot, as developed by ter Braak (1990) represents the singular value decomposition of the correlation matrix between sets **Y** and **X** as

$$\mathbf{R}_{\mathbf{YX}} = \mathbf{BC}'$$

where **B** is formed of the interset correlations between **Y** and the canonical variates (structural correlations) of the **X** set, and **C** is formed of standardized *canonical coefficients* (canonical weights) of the **X** set. Details concerning the biplot display for this analysis are presented in Ter Braak (1990).

Data Transformation

The data transformation options allow for centering and scaling the data prior to analysis. The options are:

Corrected by the grand mean: $y_{ij}^{*} = y_{ij} - \bar{y}_{\bullet\bullet}$

Columns centered: $y_{ij}^{*} = y_{ij} - \bar{y}_{\bullet j}$ this transformation is used to produce a biplot based on a PCA (principal components analysis) of the covariance matrix. Note that this analysis will differ from a standard PCA as by a factor of *N*-1, as the matrix that is decomposed is the covariance matrix times (*N*-1).

Columns centered and standardized: $y_{ij}^{*} = (y_{ij} - \bar{y}_{\bullet j})/(\sqrt{n-1}s_j)$, this transformation is useful when we want a biplot based on the PCA of the correlation matrix. Here $s_j = \sqrt{\dfrac{\sum_{i=1}^{N}(y_{ij} - \bar{y}_{\bullet j})^2}{n-1}}$ is the standard deviation of the column j.

Rows and Columns centered: $y_{ij}^{*} = y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet}$ (double centering), is useful for biplots that evaluate multiplicative interaction in 2-way tables. Also if the data represent a symmetric matrix of similarities between objects (that should satisfy certain conditions), the transformation will produce the inner product matrix $y_{ij}^{*} = (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$ for some underlying matrix **X** (see for example Mardia et al, 1979) and the SVD of matrix $\{z_{ij}\}$ will allow the user to perform a Principal Coordinate Analysis (a *Multidimensional Scaling* technique).

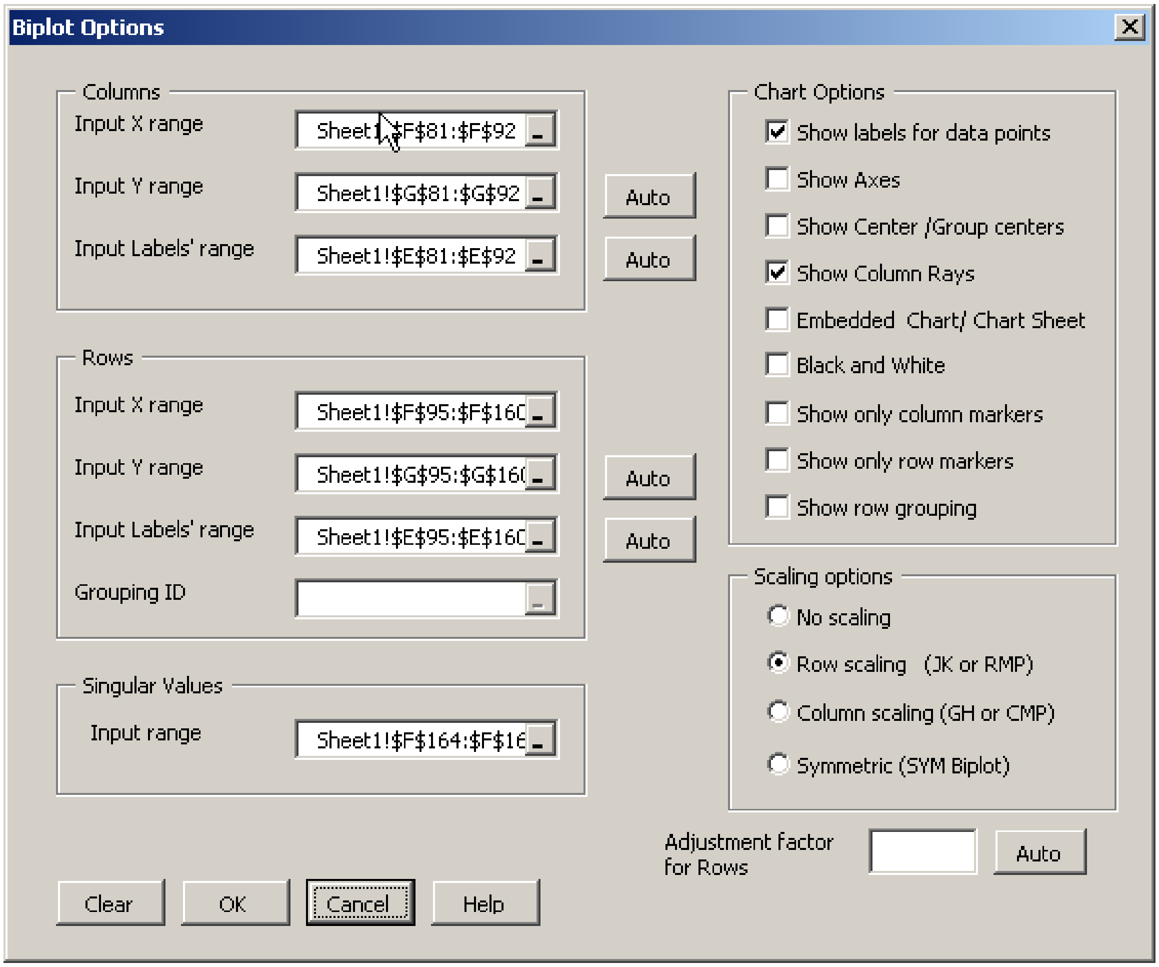
Column Selection Criteria:

In analysis of biological data, it is common to drop some of the rare species, as they may not be important in the analysis. The user can use any combination of the following criteria to select columns of the abundance data when performing CA or CCA.

- Drop all columns that have a fraction of zero counts larger than a pre-specified value. For example you may want to remove all columns (species) where more than 10% of cells are zero.
- Drop all columns whose totals are less than a pre-specified amount. For example you may want to discard all columns (species) whose totals are below some absolute value
- Drop all columns whose total is less than a pre-specified percent of the grand total. This is the same as the previous rule, but now you set the cut-off value as a percentage to the total.
- Drop all columns that are in a certain lower percentile of the column totals. For example you may want to analyze only species with the largest counts and therefore you want to drop the most rare species whose cumulative counts do not exceed say 20% of the grand total. To do this, check the associated check box and enter 20.

Following computation of the singular value decomposition, the resulting values are used to obtain the biplot scores for plotting. The BIPLOT MACRO display box is selected and used to choose the type of scaling of the singular vectors, plot options and location of the results of the SVD. In a typical analysis, the chart output box is selected on the singular value decomposition box and the location of the results of the SVD are automatically entered in the box. The user then need not worry about the location of the data to be plotted. This macro uses the built-in Excel© scatter plot chart facilities to produce a biplot with various options. The options are described below.

3. The Biplot Macro



Biplot Options

Columns

Input X range: Sheet1!$F$81:$F$92

Input Y range: Sheet1!$G$81:$G$92

Input Labels' range: Sheet1!$E$81:$E$92

Rows

Input X range: Sheet1!$F$95:$F$16

Input Y range: Sheet1!$G$95:$G$16

Input Labels' range: Sheet1!$E$95:$E$16

Grouping ID:

Singular Values

Input range: Sheet1!$F$164:$F$16

Chart Options

☑ Show labels for data points

□ Show Axes

□ Show Center /Group centers

☑ Show Column Rays

□ Embedded Chart/ Chart Sheet

□ Black and White

□ Show only column markers

□ Show only row markers

□ Show row grouping

Scaling options

○ No scaling

● Row scaling (JK or RMP)

○ Column scaling (GH or CMP)

○ Symmetric (SYM Biplot)

Adjustment factor for Rows:

Location Options

The Biplot Macro plots the row and column markers in a 2-dimensional display. The markers associated with the largest singular value are plotted on the horizontal axis and are denoted as the *X* values. If the chart option is selected, values associated with the second largest singular value are plotted on the vertical axis and denoted as the *Y* values. If chart output is selected, the *X*, *Y* and singular value information is automatically given. To plot other axes, the user must select the information to be plotted. Note that the macro may be used independently of the singular value macro for example to plot data from other packages by importing markers into Excel©. Then select the first column to plot for rows. You can select the range for the second column and labels or you can click on the Auto box to fill in the details automatically, if needed. The Auto-fill assumes that the *Y*-range is in the next rightmost column to the *X* range, and that the label range is in the next leftmost column. Also select the input range for the singular values if chart output is not selected.

Columns: Input X range Specify the location with the horizontal coordinates for the column markers (usually the variables or species).

Columns: Input Y range Specify the location with the vertical coordinates for the column markers.

Columns: Input labels range Specify the location with the labels for the column markers.

Rows: Input X range Specify the location with the horizontal coordinates for the row markers (sites).

Rows: Input Y range Specify the location with the vertical coordinates for the row markers.

Rows: Input labels range Specify the location with the labels for the row markers.

Rows: Grouping ID Specify the location with the grouping variable for the row markers. The grouping variable represents the categories' identifiers and may be either numbers or names.

Singular values: Input range Specify the location with the first two singular values, those that will be used when some of the non-trivial *Scaling options* are selected (JK, GH, or SYM biplots). Note that only the default (no scaling option) should be used with CDA and Canonical Correlation Analysis biplots.

Note: to plot the column rays select the "show column rays" check box. Click the box for "show labels for data points" to have the row and column names displayed on the chart.

Note: All ranges specified in the above options must be vertically oriented in the Sheet.

Chart Options

Show labels for data points: Check if you want to have labels for each row/column marker on the chart. By default the labels will be named as *row*1, *row*2,.., *col*1, *col*2,.... To override the default naming, you can store your labels in the sheet and specify them in the *Input labels range* options. If labels are included with the data and chart output selected, labels are automatically stored and identified.

Show Axes: Check if you want axes with axes markers displayed on the chart.

Show Center/Group Centers: Check if you want a marker(s) for the center or group centers (if there is a grouping variable). This is useful for Canonicial Discriminant Analysis.

Show Column Rays: Check if you want the column rays to be displayed on the chart.

Embedded chart / Chart Sheet: Check if you want the chart embedded in the current sheet rather than occupying a new Chart Sheet.

Black and White: Check if you want the chart to be black and white (no other colors).

Show only column markers: Check if you want to display only column markers on the chart.

Show only row markers: Check if you want to display only row markers on the chart

Scaling options: checking the associated box will produce the following displays

No scaling: The data displayed will be as specified in *Data Location* controls.

Row scaling: Corresponds to the JK (RMP) biplot that is aimed a representing the row distances (sites or row objects). The row markers' *x* and *y* coordinates are multiplied by their respective singular values, specified in the *Singular values: Input range* text box. Column markers are not adjusted.

Column scaling: Corresponds to the GH (CMP) biplot that is aimed at representing the column (species or variables) distances. The column markers' *x* and *y* coordinates are multiplied by their respective singular values, specified in the *Singular values*: *Input range* text box. Row markers are not adjusted.

Symmetric scaling Corresponds to the SYM biplot that gives equal weight to both column and row markers. Both column and row markers' coordinates are multiplied by the square roots of their respective singular values, specified in the *Singular values: Input range* text box.

Adjustment factor for rows Sometimes the plots do not look good because either columns or rows are emphasized. For example, the row markers plot nicely but the column markers are all close to the origin. This option will automatically scale the row markers to make the plot look better. Place the cursor in the box next to the **auto** button then click on the **auto** button to let the program find the best factor, which will be the ratio of *max*(Column Coordinates)/*max*(Row coordinates), or enter your own value. The scaling factor will be used to adjust the row coordinates by the same amount for **both** axes.

Note: if a nontrivial scaling option or adjustment factor is selected, the macro will compute the adjusted scores and place them in a separate sheet named Tmp<number> where the number is adjusted based on the number of temporary sheets. Therefore the chart will have references to that sheet. After the chart is drawn you can of course make any adjustments necessary using standard Excel© ***options.***

4. Examples

A variety of examples are given in the help files. Users are encouraged to copy data from the examples in help and paste it into their spreadsheet to be able to reproduce charts themselves.

5. Disclaimer

This software has been extensively validated with many data sets and all the problems that have been known to us as of May 1, 2002 were fixed. However, the user should understand that there may be other undetected bugs and problems and we will be grateful for any feedback with relevant comments and suggestions for improvements.

6. Acknowledgements

This research has been supported through grants from the U.S. Environmental Protection Agency's Science to Achieve Results (STAR) (Grant No. R82795301) and National Center for Environmental Assessment (NCEA) (Grant No. CR827820-01-0) programs. Although the research described in the article has been funded wholly or in part by the U.S. Environmental Protection Agency's STAR and NCEA programs, it has not been subjected to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. Comments from an anonymous referee greatly enhanced the presentation of the paper and program. Ilya Lipkovich programmed the biplot macro. The macros are provided at no charge. Please cite the paper if you use the macro in publications.

7. References

Digby, P.G.N. and Kempton, R.A. (1987). *Multivariate Analysis of Ecological Communities.* Chapman and Hall, London.

Gabriel, K.R. (1971). The biplot-graphic display of matrices with application to principal component analysis. *Biometrika* **58**, 453-467.

Gower, C. and Hand D.J. (1996). Biplots. Chapman & Hill, London.

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

Heath, M.T. (1997). Scientific Computing. McGraw-Hill, New York.

Legendre, P., and Legendre, L. (1998). Numerical Ecology, Elsevier, Amsterdam.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). Multivariate Analysis, London: Academic Press.

Rothkopf, E.Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *J. Exp. Psychol*., **53**, 94-101

Seber, G.A.F. (1984). Multivariate Observations, John Wiley and Sons, New York.

Ter Braak, C.J.F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis, Ecology 67:5, 1167-1179.

Ter Braak, C.J.F. (1990). Interpreting canonical correlation analysis through biplots of structural correlations and weights. Psychometrika 55, 519-531.

Ter Braak, C.J.F. and Looman, C.W. (1994). Biplots in reduced-rank regression. Biometrical Journal. 36:8, 983-1003.