Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Appendix C Offprint

Glossary of Terms

First published: September 2010
ISBN: 978-84-923846-8-6

Supporting websites:
http://www.fbbva.es
http://www.multivariatestatistics.org

Appendix C

Glossary of Terms

In this appendix an alphabetical list of the most common terms used in this book is given, along with a short definition of each. Words in italics refer to terms which are contained in the glossary.

— *adjusted principal inertias*: a modification of the results of a *multiple correspondence analysis* that gives a more accurate and realistic estimate of the inertia accounted for in the solution.

— *aspect ratio:* the ratio between a unit length on the horizontal axis and a unit length on the vertical axis in a graphical representation; should be equal to 1 for any *biplot* or *map* that has a spatial interpretation.

— *asymmetric biplot/map:* a joint display of the rows and columns where the two clouds of points have different scalings (also called *normalizations*), usually one in *principal coordinates* and the other in *standard coordinates*.

— *biplot:* a joint display of points representing the rows and columns of a table such that the *scalar product* between a row point and a column point approximates the corresponding element in the table in some optimal way.

— *biplot axis:* a line in the direction of a *biplot vector* onto which points can be projected in order to estimate values in the table being analysed.

— *biplot vector:* a vector drawn from the origin to a point in a *biplot*, often representing a variable of the data matrix, or a *supplementary variable*.

— *bootstrapping:* a computer-based method of investigating the variability of a statistic, by generating a large number of replicate samples, with replacement, from the observed sample.

— *Burt matrix:* a particular matrix of *concatenated tables*, consisting of all two-way cross-tabulations of a set of categorical variables, including the cross-tabulations of each variable with itself.

— *calibration:* the process of putting a scale on a *biplot axis* with specific tic-marks and values.

— *canonical correspondence analysis (CCA):* extension of *correspondence analysis* to include external explanatory variables; the solution is constrained to have dimensions that are linearly related to these explanatory variables.

— *centroid:* weighted average point.

— *chi-square distance:* weighted *Euclidean distance* measure between *profiles*, where each squared difference between profile elements is divided by the corresponding element of the average profile; the distance function used in *correspondence analysis*.

— *classical scaling:* a version of *multidimensional scaling* which situates a set of points in multidimensional Euclidean space, based on their interpoint distances or dissimilarities, and then projects them down onto a low-dimensional space of representation.

— *concatenated table:* a number of tables, usually based on cross-tabulating the same individuals, joined together row-wise or column-wise or both.

— *contingency table:* a cross-tabulation of a set of cases or objects according to two categorical variables; hence the grand total of the table is the number of cases.

— *contribution to variance/inertia:* component of variance (or *inertia*) accounted for by a particular point on a particular *principal axis*; these are usually expressed relative to the corresponding *principal variance* (or *principal inertia*) on the axis (giving a diagnostic of how the axis is constructed) or relative to the variance (or *inertia*) of the point (giving a measure of how well the point is explained by the axis).

— *contribution (or standard) biplot: biplot* in which one set of points, usually the variables, is normalized so that their squared coordinates are the parts of variance (or *inertia*) on the respective axes; this scaling facilitates the interpretation of the solution space.

— *correspondence analysis (CA):* a method of displaying the rows and columns of a table as points in a spatial map, with a specific geometric interpretation of the positions of the points as a means of interpreting the similarities and differences between rows, the similarities and differences between columns and the association between rows and columns. Fundamental concepts in the definition of CA are those of *mass* and *chi-square distance*.

— *covariance biplot: asymmetric biplot* where the variables (usually columns) are in *principal coordinates*, thus approximating the covariance structure of the variables (i.e., lengths of *biplot vectors* approximate standard deviations, angle cosines between biplot vectors approximate correlations), and the rows (usually cases) in *standard coordinates*.

— *dimension:* in the context of biplots, a synonym for axis.

— *dimensionality:* the number of dimensions inherent in a table needed to reproduce its elements exactly in a *biplot* or *map*; in this context it is synonymous with the *rank* of the matrix being analyzed.

— *dimension reduction*: the action of finding fewer *dimensions* than the *dimensionality* of a matrix, which can reproduce the matrix optimally.

— *dissimilarity:* a measure of difference between objects which is like a *distance* but does not satisfy the *triangular inequality*.

— *distance:* a measure of difference between pairs of objects which is always positive or zero, and zero if and only if the objects are identical, and furthermore satisfies the *triangular inequality*.

— *double centring:* an operation applied to a data matrix which first subtracts the row means from each row of the matrix, and then subtracts the columns means from each column of the row-centred matrix. Often the centring includes weights on the rows and the columns (e.g., the *masses* in *correspondence analysis* and *log-ratio analysis*). The (weighted) means of the rows and the columns of a double-centred matrix are all 0.

— *dual biplots*: a pair of *asymmetric biplots* which are versions of the same *singular-value decomposition*; usually, the allocation of the singular values to the left or right matrix of singular vectors is what distinguishes the pair.

— *dummy variable:* a variable that takes on the values 0 and 1 only; used in one form of *multiple correspondence analysis* to code multivariate categorical data.

— *eigenvalue:* a quantity inherent in a square matrix, forming part of a decomposition of the matrix into the product of simpler matrices. A square matrix has as many eigenvalues and associated eigenvectors as its rank; in the context of *biplots*, eigenvalue is a synonym for the *principal variance* or *principal inertia.*

— *Euclidean distance:* distance measure between vectors where squared differences between corresponding elements are summed, followed by taking the square root of this sum.

— *form biplot: asymmetric biplot* where the cases (usually rows) are in *principal coordinates*, thus approximating distances between cases, and the columns (usually variables) in *standard coordinates*.

— *generalized linear model (GLM):* a generalization of linear regression, where there are several possible transformations of the mean of the response variable and several possible choices of the conditional probability distribution; examples are Poisson regression (transformation: logarithm; probability distribution: Poisson) and logistic regression (transformation: logit, or log-odds; probability distribution: binomial).

— *generalized linear model biplot:* similar to the *regression biplot* , except that the coefficients are obtained through a *generalized linear model*; hence any *calibrations* of the *biplot axes* are not at equal intervals as in regression biplots, but reflect the transformation of the mean of the corresponding response variable.

— *gradient:* in optimization theory, the vector of partial derivatives of a multivariable function with respect to its variables, indicating the direction of steepest ascent of the function; when the function is linear (e.g., a regression equation), then the gradient is simply the vector of coefficients of the variables.

— *indicator matrix:* the coding of a multivariate categorical data set in the form of *dummy variables*.

— *inertia:* weighted sum of squared distances of a set of points to their *centroid*; in *correspondence analysis* the points are *profiles*, weights are the *masses* of the profiles and the distances are *chi-square distances*.

— *interactive coding:* the formation of a single categorical variable from all the category combinations of two or more categorical variables.

— *joint correspondence analysis (JCA):* an adaptation of *multiple correspondence analysis* to analyse all unique two-way cross-tabulations of a set of categorical variables (contained in the *Burt matrix*) while ignoring the cross-tabulations of each variable with itself.

— *left matrix:* the first matrix in the decomposition of the *target matrix*, which provides the coordinates of the rows in a *biplot*.

— *linear discriminant analysis:* a dimension-reduction method which aims to optimally separate the *centroids* of groups of multivariate points, using the *Mahalanobis distance* to define distances between points.

— *link vector:* the vector in a *biplot* that joins two points and thus represents the difference vector between the two (e.g., the difference between two variables).

— *log-ratio:* given two elements in the same row or same column of a strictly positive data matrix, this is the logarithm of the ratio of the values.

— *log-ratio analysis:* a dimension-reduction method for a table of strictly positive data all measured on the same scale, based on log-transforming the data and *double-centring* before decomposing by the *singular value decomposition*. The rows and columns are preferably weighted, usually proportional to the margins of the table. Log-ratio analysis effectively analyzes all *log-ratios* in the rows and the columns of the table.

— *log-ratio distance:* the distance function underlying *log-ratio analysis*, based on the differences between all *log-ratios* in the rows or in the columns.

— *Mahalanobis distance:* a distance function used in *linear discriminant analysis*, which aims to de-correlate and standardize the variables within each of the groups being separated.

— *map:* a spatial representation of points with a *distance* or *scalar product* interpretation.

— *mass:* a weight assigned to a point; in *correspondence analysis* and *log-ratio analysis*, the row and column masses are the marginal totals of the table, divided by the grand total of the table.

— *monotonically increasing function:* a function that steadily increases as its argument increases; that is, its derivative (or slope) is always positive.

— *multidimensional scaling (MDS):* the graphical representation of a set of objects based on their interpoint *distances* or *dissimilarities*.

— *multiple correspondence analysis (MCA):* for more than two categorical variables, the *correspondence analysis* of the *indicator matrix* or *Burt matrix* formed from the variables.

— *nested principal axes:* a property of a *biplot* or *map* where solutions consist of a set of uncorrelated principal axes which combine in an ordered way: for example, the best three-dimensional solution consists of the two axes of the best two-dimensional solution plus the third axis.

— *normalization:* refers to the scale of a variable or a principal axis in terms of its variance; for example, a variable divided by its standard deviation is normalized to have variance 1, while the *principal coordinates* of a set of points on a particular axis have a normalization equal to the *eigenvalue* (*principal variance* or *principal inertia*) of that axis.

— *permutation test:* generation of data permutations, either all possible ones or a large random sample, assuming a null hypothesis, in order to obtain the null distribution of a test statistic and thus estimate the *p*-value associated with the observed value of the statistic.

— *principal axis:* a direction of spread of points in multidimensional space that optimizes the variance or *inertia* displayed; can be thought of equivalently as an axis which best fits the points in a least-squares sense, often weighted.

— *principal component analysis (PCA):* a method of dimension reduction which attempts to explain the maximum amount of variance in a data matrix in terms of a small number of *dimensions*, or components.

— *principal coordinates:* coordinates of a set of points projected onto a *principal axis*; the (weighted) sum of squared coordinates of the points along an axis equals the *principal inertia* on that axis.

— *principal inertia (or principal variance):* the *inertia* (or variance) displayed along a *principal axis*; often referred to as an *eigenvalue*.

— *profile:* a row or a column of a table divided by its total; the profiles are the points visualized in *correspondence analysis*.

— *projection:* given a point in a high-dimensional space, its projection onto a low-dimensional subspace refers to that point closest to the original point; the action of projection is usually perpendicular to the subspace.

— *projection matrix:* a matrix which when multiplied by a vector gives the projection of that vector on a low-dimensional subspace.

— *rank:* the rank of a matrix in a geometric context is the number of *dimensions* needed to reproduce the matrix exactly.

— *redundancy analysis (RDA):* extension of *principal component analysis* to include external explanatory variables; the solution is constrained to have *dimensions* that are linearly related to these explanatory variables.

— *regression biplot:* a *biplot* which has as its system of display axes a set of explanatory variables (in the simplest case, two variables), showing firstly a set of case points in terms of these variables and secondly a set of *biplot vectors* with coordinates defined by regression coefficients from the respective linear regressions of response variables on the explanatory variables. If the axes are standardized, then the biplot vectors are defined by the standardized regression coefficients.

— *right matrix:* the second matrix in the decomposition of the *target matrix*, which provides the coordinates of the columns in a *biplot*.

— *scalar product:* for two point vectors, the product of their lengths multiplied by the cosine of the angle between them; directly proportional to the projection of one point on the vector defined by the other.

— *scree plot:* a bar chart of the set of *eigenvalues* (*principal variances* or *inertias*) associated with a *biplot,* in descending order of magnitude.

— *simplex:* a triangle in two dimensions, a tetrahedron in three dimensions, and generalizations of these geometric figures in higher dimensions; in *correspondence analysis J*-element *profiles* lie inside a simplex defined by *J vertices* in (*J* –1)-dimensional space.

— *singular value decomposition (SVD):* the decomposition of a matrix into the product of three matrices with simple structure: the matrix of left singular vectors multiplied by the diagonal matrix of singular values (all positive and in descending order) multiplied by the transposed matrix of right singular vectors. The SVD is the natural generalization of the *eigenvalue*–eigenvector decomposition, but applicable more generally to rectangular matrices.

— *standard coordinates:* coordinates of a set of unit points projected onto *principal axes*—their (weighted) sum of squares along an axis equals 1.

— *standardized regression coefficient:* a regression coefficient that corresponds to a variable that has been normalized to have variance (or *inertia*) 1.

— *subset correspondence analysis:* a variant of *correspondence analysis* which allows subsets of rows and/or columns to be analysed, while maintaining the same *chi-square distance* function and point *masses* as for the full table.

— *supplementary point:* a point which has a position (e.g., a vector of data in principal component analysis or a *profile* in correspondence analysis) with *mass* set equal to zero; in other words, a supplementary point is displayed on the map but has not been used in the construction of the map.

— *supplementary variable:* a variable which is positioned in a map by (weighted) least-squares regression on the *principal axes*; the variable is usually depicted as a vector with coordinates equal to the regression coefficients.

— *symmetric map:* a simultaneous display of the *principal coordinates* of the rows of a matrix and the *principal coordinates* of its columns. While the distance geometry of both rows and columns is shown, this map is not a *biplot*, but approximates one if the *eigenvalues* of the axes are not too different.

— *target matrix:* a matrix which is decomposed into the product of two matrices, the *left* and *right matrices*, which provide the coordinates for the rows and columns respectively in a *biplot*.

— *transition formula:* the relationship between the row points and column points in a *map* or a *biplot*.

— *triangular inequality:* a property of a true distance function whereby the distance between two objects is necessarily less than or equal to the sum of the distances from the two objects to a third one.

— *triplot:* a *biplot* showing, in addition, a third set of points or vectors corresponding to the explanatory variables which constrain the solution, for example in *canonical correspondence analysis* or *redundancy analysis*.

— *vertex:* a unit point in multidimensional space, with all elements zero except one with value 1, usually a unit *profile* in CA which delimits the *simplex* within which the points in CA lie.

— *weighted Euclidean distance:* similar to *Euclidean distance*, but with a positive weighting factor for each squared difference term.