Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Appendix A Offprint

Aspects of Theory

First published: December 2013
ISBN: 978-84-92937-50-9

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

Fundación BBVA

Appendix A

Aspects of Theory

This appendix summarizes the theory described in this book. The treatment is definitely not exhaustive and the bibliography in Appendix B gives some pointers to additional reference material. We deal with the theory in more or less the same order as the corresponding methods appeared in the text, although some topics might be grouped slightly differently.

Contents

Transformations and standardization 279
Measures of distance and dissimilarity 281
Cluster analysis 282
Multidimensional scaling 283
Principal component analysis, correspondence analysis and log-ratio analysis 284
Supplementary variables and points 286
Dimension reduction with constraints 287
Permutation testing and bootstrapping 288
Statistical modelling 290

Transformations and standardization

The most common measurements scales are:

- *Continuous interval:* differences between values are measured and interpreted; variables on this scale can have negative values; we also say an *additive* scale. For example, time, temperature.
- *Continuous ratio:* ratios between values are measured and interpreted (i.e., percentage differences); variables on this scale have positive values; we also say a *multiplicative* scale. For example, heavy metal concentration, weight.
- *Categorical* (or *discrete*) *nominal:* only a few categories are possible and they have no particular order. For example, region, phylogenetic group.
- *Categorical* (or *discrete*) *ordinal:* only a few categories are possible and they do have an inherent ordering. For example, month, sediment class.

- *Count:* Variable that takes positive integer values, including 0; we also say a *frequency*. For example, abundance count, number of offspring.
- *Compositional:* this refers to a set of variables with the unit-sum constraint, proportions that add up to 1 (or 100% if percentages). For example, a set of fatty acid compositions, relative abundances of a set of species.

Standardization is often applied to put different variables on the same scale. For data $x_1, \ldots, x_n$ on an interval-scale variable the most common is to make them all mean 0, variance 1, by *centering* (i.e., subtracting) with respect to the mean $\bar{x}$ and *normalizing* (i.e., dividing) with respect to the standard deviation s.

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, \ldots, n \tag{A.1}$$

Other forms of standardization might be deemed more suitable, such as centering with respect to the median and normalizing by the range.

For positive data $x_1, \ldots, x_n$ on a ratio-scale variable, a convenient transformation is the logarithm:

$$z_i = \log(x_i) \tag{A.2}$$

because it converts the ratio-scale variable x to an interval-scale variable z, which needs no further normalization.

Nominal and ordinal categorical data are often converted to *dummy variables*, which are as many variables as there are categories, taking the values 0 and 1.

Count data as well as compositional data are similar to ratio-scale variables and are usually logarithmically transformed, or root-transformed (square root, fourth root…). If there are zero count values, then they are often transformed as $\log(1 + x)$. In the case of compositional data, we prefer to replace zeros with small values equal to the detection limit in the context of the data.

The Box-Cox transformation is a general power transformation for ratio-scale, count and compositional data:

$$z = \frac{1}{\lambda}(x^{\lambda} - 1) \tag{A.3}$$

usually for powers λ less than 1, and where for zero values of x, $x^{\lambda} = 0$. As the power λ tends to 0 (we say as the root transformation gets stronger) the transformation gets closer and closer to the log-transformation $\log(x)$.

Differences between values of a single interval variable are computed simply by subtraction, while for a ratio variable or a count, multiplicative differences can be computed by taking ratios, or differences on the log-scale. For multivariate samples difference is measured by a distance or dissimilarity which combines differences across the variables. A *distance* has all the properties of a well-defined metric, including the triangular inequality property. A *dissimilarity* is an acceptable measure of inter-sample difference but does not obey the triangular inequality.

A general distance is the *weighted Euclidean distance*, computed between two samples $x_1,\ldots,x_p$ and $y_1,\ldots,y_p$ observed on p variables, with weights on the variables $w_1,\ldots,w_p$:

$$d_{x,y} = \sqrt{\sum_{j=1}^{p} w_j (x_j - y_j)^2} \qquad \text{(A.4)}$$

Well-known special cases are:

- Euclidean distance, when $w_j = 1$; applicable to set of interval variables all on the same scale that do not need normalization, or a set of ordinal variables, all on the same scale (e.g., five-point ordinal scales of plant coverage) for which the inter-category differences are accepted as interval measures.
- Standardized Euclidean distance, for a set of interval-scale variables: $w_j = 1/s_j^2$, the inverse of the variance of the j-th variable; this is the distance function computed by standardizing all the variables first and then applying the regular unweighted Euclidean distance.
- Chi-square distance, for abundance, relative abundance, and compositional data: $w_j = 1/c_j$, where c_j is the mean for variable j.

The Bray-Curtis (or Sørensen) dissimilarity (which is not a true distance function, since it does not obey the triangle inequality) is a popular choice for measuring differences between samples when the data are abundances, or other positive amounts such as biomasses:

$$b_{x,y} = \frac{\sum_{j=1}^{p} |x_j - y_j|}{\sum_{j=1}^{p} (x_j + y_j)} \qquad \text{(A.5)}$$

For one/zero data, for example presence/absence data, there are many possibilities and we only summarize the two presented in this book, the *matching coefficient* and *Jaccard dissimilarity*. For p variables observed on two samples, we define

a = number of variables matched with a 1 in both samples, d = number of matches of 0 in both samples, b = number of variables "mismatched" with 1s in the first sample, 0s in the second, c = number of mismatches with 0s in the first sample, 1s in the second, so $a + b + c + d = p$. Then:

- Matching: $(b + c)/p$ (actually, this is a measure of mismatching) (A.6)

- Jaccard: $(b + c)/(p - d)$ (A.7)

Both the above dissimilarities lie between 0 and 1, with 0 when there are no mismatches. For matching the maximum value of 1 is attained when $a + d = 0$ (no 1s or 0s matched), while for Jaccard, which ignores matching 0s, the maximum of 1 is reached when $a = 0$ (no 1s matched). Jaccard is preferable for presence/absence data when the co-occurrence of absences is not interesting, only the co-occurrence of presences.

For *mixed-scale* multivariate data, usually continuous and categorical mixed, some form of normalization or homogenization is required so that it makes sense to combine them into a measure of inter-sample difference. The *Gower index of dissimilarity* (not discussed in the book) involves applying a standardization on the continuous variables to make them comparable to the categorical ones that are dummy coded, after which Euclidean distance is applied. The alternative that is presented in this book is to *fuzzy code* the continuous variables into sets of fuzzy categories. Fuzzy categories corresponding to a continuous variable look like a set of dummy variables except that they have any values between 0 and 1, not exactly 0 or 1, and in this way preserve the exact value of the continuous variable in categorical form. With the categorical variables coded as dummy variables and the continuous variables coded as fuzzy categorical variables, Euclidean distance can be applied, possibly with weights to adjust the contributions of each variable to the measure of distance.

Cluster analysis

To define a method of cluster analysis one defines the algorithm used to implement the method. Two approaches are of interest, hierarchical and nonhierarchical clustering, both of which rely on a matrix of proximities (distances or dissimilarities) between pairs of *objects* to be clustered, where objects can be sampling units such as sites or variables such as species.

Hierarchical cluster analysis creates a *dendrogram*, or binary tree, in a stepwise fashion, successively aggregating objects, two at a time, and eventually aggregating groups of objects as well, according to their proximities. Assuming a decision about the measure of proximity has been made, the crucial decision is then how to measure proximity between groups of objects formed in the previous stage of the stepwise procedure. The main options in practice are: (1) complete linkage, where the

maximum distance or dissimilarity value between groups is used; (2) average linkage, where the average value is used; or (3) Ward clustering, a different ANOVA-like approach which maximizes the overall between-group variance at each step of the clustering, equivalently minimizing within-group variance. The final result is a dendrogram, which is then cut at a certain level to create a small number of groups of objects, designed to be internally homogeneous and distinct from one another.

Nonhierarchical cluster analysis is used when the number of objects is very large, say greater than100, when the dendrogram becomes unwieldy to interpret. The most popular example is *k-means clustering*, which has the same objective as Ward clustering, to maximize between-group variance while minimizing within-group variance. The number of groups k is pre-specified and the algorithm proceeds from a random start to create k groups iteratively, at each iteration assigning objects to the group with the closest mean. The solution is seldom globally optimum and several random starts are recommended, and the best final solution accepted.

Multidimensional scaling

While clustering results in a grouping of objects, multidimensional scaling (MDS) results in an ordination map of the objects. Given a matrix of inter-object proximities, MDS finds a configuration of the objects in a space of specified dimensionality, almost always a two-dimensional plane, such that the displayed inter-object distances are as close as possible to the given proximities. Different ways of measuring the fit between the displayed distances, gathered in a matrix $\mathbf{D}$, and the given proximities, gathered in a matrix Δ, lead to different MDS techniques.

Classical MDS, also called *principal coordinate analysis*, relies on the eigenvalue-eigenvector decomposition, called *eigen-decomposition*, of a square matrix of scalar products to obtain a solution. Initially, the elements of the given proximity matrix are squared – this matrix of squared distances or dissimilarities is denoted by $\Delta^{(2)}$. To give the most general form of classical MDS, we assume that there is a set of positive weights $w_1,\ldots,w_n$ assigned to the n objects, where $\sum_i w_i = 1$, so that the objective is to optimize a weighted fit where objects of higher weight are displayed more accurately in the solution. An operation of double-centering and multiplying by $-½$ is applied to $\Delta^{(2)}$ to obtain the scalar product matrix $\mathbf{S}$:

$$\mathbf{S} = -½(\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})\Delta^{(2)}(\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})^{\mathsf{T}} \qquad \text{(A.8)}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{1}$ is the $n \times 1$ vector of 1s and $\mathbf{w}$ the $n \times 1$ vector of weights. Centering of the values in the columns of $\Delta^{(2)}$ is performed by premultiplying by the *centering matrix* $(\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})$, while post-multiplying by the transposed centering matrix centers the values in the rows. The eigen-decomposition is then obtained on a weighted form of $\mathbf{S}$, where $\mathbf{D}_w$ is the diagonal matrix of weights:

$$\mathbf{D}_w^{½}\,\mathbf{S}\,\mathbf{D}_w^{½} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^{\mathsf{T}} \tag{A.9}$$

U contains the eigenvectors of **S** in its columns and $\mathbf{D}_\lambda$ is a diagonal matrix with the eigenvalues of **S** down the diagonal, in decreasing order. The *principal coordinates* of the objects are finally given by:

$$\mathbf{F} = \mathbf{D}_w^{-½}\mathbf{U}\mathbf{D}_\lambda^{½} \tag{A.10}$$

The rows of **F** refer to the objects and the columns to the principal axes of the solution in decreasing order of importance. For a two-dimensional display the first two columns provide the coordinate pairs (f_{i1}, f_{i2}) for displaying the i-th object. The sum of the eigenvalues are a measure of the total variance and each eigenvalue a measure of explained variance by a principal axis, hence the quality of display in a two-dimensional solution, which can be interpreted like an R^2 in regression, is $(\lambda_1 + \lambda_2)/\sum_k \lambda_k$.

When all the nonzero eigenvalues are positive, the given matrix of proximities is *Euclidean embeddable*, which means that it is possible to represent the objects in a Euclidean space, with dimensionality equal to the number of positive eigenvalues. When there are some negative eigenvalues, their absolute values quantify the part of variance that is impossible to represent in a Euclidean space. For example, a matrix of chi-square distances is always Euclidean embeddable, while a matrix of Bray-Curtis dissimilarities is not. This fact has led to practitioners preferring nonmetric MDS to display Bray-Curtis dissimilarities.

Nonmetric MDS relaxes the measure of fit between the displayed distances and the given proximities. A perfect fit in nonmetric MDS would be when the order of all the displayed distances is the same as the order of all the given proximities. Specifically, if the $½n(n-1)$ displayed distances are listed next to the $½n(n-1)$ given proximities, a perfect fit would give a Spearman rank correlation between the two lists of 1. Rather than measure quality of fit, nonmetric MDS measures error of fit using a quantity called *stress*, so a perfect fit would be a stress of 0. The measure of stress will always appear more optimistic than the measure of unexplained variance in classical MDS, but this does not imply that nonmetric MDS is an improvement – classical MDS has a stricter objective, and thus more error in achieving it.

Principal component analysis, correspondence analysis and log-ratio analysis

These three methods, abbreviated as PCA, CA and LRA, are variations of the same theme, so we treat them together. All three methods start with a rectangular data matrix, prepared according to the method for being decomposed by the singular-value decomposition (SVD). The SVD is similar to the eigen-decomposition but applicable to rectangular rather than square matrices. All three methods can be defined using eigen-decompositions as well, but the SVD approach is more

elegant and brings out clearly the features of the eventual joint display, for example, whether the display is a biplot or not.

The SVD is defined as follows, for a rectangular matrix $\mathbf{A}$ $(I \times J)$:

$$\mathbf{A} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^\mathsf{T} \tag{A.11}$$

where $\mathbf{U}$ $(I \times R)$ and $\mathbf{V}$ $(I \times R)$ have orthonormal columns: $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$, and $\mathbf{D}_\sigma$ is a diagonal matrix of positive values in descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_R > 0$. R is the rank of $\mathbf{A}$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are called *left* and *right singular vectors*, respectively, corresponding to the *singular values* σ_r. (A.11) can be written equivalently as the sum of R terms, each of which involves a singular value and associated pair of singular vectors:

$$\mathbf{A} = \sigma_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} + \sigma_2\mathbf{u}_2\mathbf{v}_2^\mathsf{T} + \cdots + \sigma_R\mathbf{u}_R\mathbf{v}_R^\mathsf{T} \tag{A.12}$$

Since each matrix $\mathbf{u}_r\mathbf{v}_r^\mathsf{T}$ has sum of squared elements equal to 1 and the singular values are in descending order, this already suggests that the first terms of (A.12) come close to reproducing the matrix $\mathbf{A}$. In fact, the famous Eckart-Young theorem states that the first R^* terms constitute a rank R^* least-squares matrix approximation of $\mathbf{A}$ – if we take the first two terms, for example, which is the most popular choice, then we have a rank 2 approximation of $\mathbf{A}$, and this will provide us with coordinates of points representing the rows and columns of $\mathbf{A}$ in a two-dimensional plot.

We need a slightly more general form of the SVD to take into account weights assigned to the rows and columns. Suppose $r_1,\ldots,r_I$ and $c_1,\ldots,c_J$ are, respectively, two such sets of weights, all positive and each set adding up to 1. Then, the weighted form of the SVD, which gives weighted least-squares approximations to $\mathbf{A}$, is obtained by first multiplying the elements a_{ij} of the matrix by the square roots of the weights, $(r_ic_j)^{1/2}$, then decomposing this reweighted matrix by the SVD, and finally "de-weighting" the final result. In matrix formulation these three steps are as follows, where $\mathbf{D}_r$ and $\mathbf{D}_c$ denote diagonal matrices of the row and column weights:

- Weight rows and columns: $\mathbf{D}_r^{1/2}\mathbf{A}\mathbf{D}_c^{1/2}$ (A.13)
- Compute SVD: $\mathbf{D}_r^{1/2}\mathbf{A}\mathbf{D}_c^{1/2} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^\mathsf{T}$ (A.14)
- "De-weight" to get the solution: $\mathbf{A} = (\mathbf{D}_r^{-1/2}\mathbf{U})\mathbf{D}_\sigma(\mathbf{D}_c^{-1/2}\mathbf{V})^\mathsf{T}$ (A.15)

Solutions of PCA, CA and LRA can be found by specifying the input matrix $\mathbf{A}$ and the weights. In all cases there is some type of centering of the original data matrix

to obtain **A**. Centering of the columns, for example, as seen already in (A.8), is performed by pre-multiplying by $(\mathbf{I} - \mathbf{1r}^\mathsf{T})$, where **r** is the vector of row weights. Centering of the values in the rows involves post-multiplying by $(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}$, where **c** is the vector of column weights. Here follow the three variants:

PCA: The data matrix **Y** contains interval-scale data, cases in rows, variables in columns. Usually case and variable weights are equal, i.e. $\mathbf{r} = (1/I)\mathbf{1}$ and $\mathbf{c} = (1/J)\mathbf{1}$, where **1** denotes an appropriate vector of ones. The columns are centered and optionally standardized, for example in the unstandardized case, $\mathbf{A} = (\mathbf{I} - (1/I)\mathbf{11}^\mathsf{T})\mathbf{Y}$. For the standardized case, divide the values in each column of **Y** by their respective standard deviation.

CA: The data matrix **Y** contains nonnegative ratio-scale data, usually counts such as abundances, or biomasses or percentages. Suppose **P** equals **Y** divided by its grand total, so that the sum of all elements of **P** is 1. The row and column sums of **P** are **r** and **c**, the row and column weights. Compute the matrix of ratios $p_{ij}/(r_i c_j)$, i.e. $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}$. Then **A** is the double-centered matrix of these ratios: $\mathbf{A} = (\mathbf{I} - \mathbf{1r}^\mathsf{T})\,\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1}(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}$.

LRA: The starting point of LRA is similar to CA, except that the data matrix **Y** must be strictly positive. Again the masses **r** and **c** are the row and column sums of **Y** relative to the grand total. Then **A** is the double-centered matrix of the logarithms of Y: $\mathbf{A} = (\mathbf{I} - \mathbf{1r}^\mathsf{T})\ \log(\mathbf{Y})\,(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}$.

After putting these options through steps (A.13)–(A.15), various coordinates can be computed:

Principal row coordinates: $\mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\sigma$	*Principal column coordinates:* $\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\sigma$	(A.16)
Standard row coordinates: $\mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}$	*Standard column coordinates:* $\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$	(A.17)
Contribution row coordinates: **U**	*Contribution column coordinates:* **V**	(A.18)

In each method the total variance, customarily called *inertia* in CA, is the sum of squared singular values computed in (A.14). This is identical to the sum of squared elements of the weighted matrix in (A.13). The part of variance explained by the first R^* dimensions of the solution (e.g., $R^* = 2$) is the sum of the first R^* squared singular values. The squared singular values are, in fact, eigenvalues in the equivalent definitions in terms of eigen-decompositions.

Supplementary variables and points

In all three methods there is the concept of a *supplementary variable* and a *supplementary point*. A supplementary variable is an additional continuous variable that is related to the low-dimensional solution afterwards, using multiple regression.

When the supplementary variable is standardized and the row standard coordinates are used as explanatory variables, the regression coefficients reduce to the correlation coefficients of the variable with the dimensions, thanks to the dimensions being uncorrelated. Hence the supplementary variable can be represented by coordinates equal to their correlation coefficients. If the rows are displayed in standard coordinates then they have a biplot relationship with these supplementary variables: rows can be projected onto the supplementary variable direction to line up the rows on that variable, the origin being the average. Notice that if there are row weights, then the regression and correlation calculations have to be weighted.

A supplementary point is an additional row (or column) that one wants to add to the existing map. This point differs from a supplementary variable in that it is comparable in scale to the data matrix that was analysed (called the *active* data matrix – sometimes a supplementary point is referred to as *passive*). For example, in a CA of abundance data, there might be additional species, or groups of species, that one wants to situate in the ordination. These have profiles just like the active data, and they can be projected onto the solution just like the active profiles were projected. The only difference is that the supplementary points have not been used to construct the solution, as if they were data points with zero weight. Supplementary points are often used as an alternative way of representing a categorical variable in an ordination. For example, again in CA, suppose the data were fish abundances, with columns as fish and classified into two types, pelagic and demersal. Aggregating all the columns corresponding to pelagic fish and all those corresponding to demersal fish gives two new columns labelled *pelagic* and *demersal*. These aggregated abundances have well-defined profiles in the column space and can be displayed on the ordination – in fact, their positions will be at the respective weighted average positions of the set of pelagic and set of demersal fish. In a similar way, fuzzy categories can be displayed. For example, the rows (e.g., sites) may have fuzzy categories for temperature, so aggregation of abundances is now performed over the rows to get four fictitious sites representing the fuzzy categories. The aggregation must be fuzzy as well, in other words, the abundances are multiplied by the fuzzy value and summed.

Dimension reduction with constraints

The three methods defined above lend themselves in exactly the same way to include a second data matrix $\mathbf{X}$ $(I \times K)$ of K explanatory variables, continuous and/or categorical in the form of dummy variables, that serve to constrain the solution. The data matrix $\mathbf{Y}$ is then regarded as responses to these explanatory variables, or predictors. Suppose $\mathbf{X}$ is standardized, always taking into account the weights assigned to the rows, in other words the columns of $\mathbf{X}$ have weighted means zero and weighted variances 1. The matrix $\mathbf{A}$ is first projected onto the space of the explanatory variables:

$$\mathbf{A}_X = [\mathbf{X}(\mathbf{X}^T\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_r]\mathbf{A} \quad (A.19)$$

and then the same three steps (A.13)–(A.15) are applied to $\mathbf{A}_X$ instead of $\mathbf{A}$, with the same options for the coordinates. This gives, respectively, redundancy analysis (PCA with constraints), canonical correspondence analysis (CCA, CA with constraints), and constrained log-ratio analysis. $\mathbf{A}_X$ is that part of the response data that is perfectly explained by the predictors. The matrix $\mathbf{A} - \mathbf{A}_X$ is the part of the response data that is uncorrelated with the predictors. If $\mathbf{X}$ includes variables that one wants to partial out, then $\mathbf{A} - \mathbf{A}_X$ is analysed using the same steps (A.13)–(A.15). In the case of CCA this is called *partial CCA*.

The total variance (or inertia) is now first partitioned into two parts, the part corresponding to the projected matrix $\mathbf{A}_X$, which is in the space of the predictors, and the part corresponding to $\mathbf{A} - \mathbf{A}_X$, which is in the space uncorrelated with the predictors. Otherwise, the computation of coordinates defined in (A.16)–(A.18) and the addition of supplementary variables and points follow in the same way.

Permutation testing and bootstrapping

The solutions obtained in all the multivariate analyses described in this book should be regarded as a complex *point estimate* – dendrograms and ordinations do not contain any information about the statistical significance of the results or whether the results would have been any different if the study were repeated in the same way. In order to perform hypothesis testing or to obtain intervals or regions of confidence, some standard multivariate tests exist for very special situations, which have quite restrictive assumptions, for example that data come from a multivariate normal distribution. We resort to computationally intensive methods to judge whether our solutions are nonrandom, reflecting some actual structure rather than random variation. In this book we have used permutation testing to obtain *p*-values associated with certain hypotheses, and bootstrapping to obtain measures of confidence, although this distinction is actually blurred (for example, one can do hypothesis testing using bootstrapping as well).

Permutation testing can be used for testing differences between groups. Under the null hypothesis that there is no inter-group difference, so that all the observations (e.g., sites) come from the same distribution, we can randomly assign the group labels to the observations and measure the inter-group difference by some reasonable statistic, such as the between-group sum of squares in multivariate space. Doing this a large number of times, obtaining a large number – say 9,999 – of values of the statistic, which defines its null distribution. Then, we see where the actual inter-group measure (in this case, the 10,000th) lies on this distribution and the estimated *p*-value is the proportion of all 10,000 values equal to or more

extreme than this value. The actual value is included in this proportion, so the smallest *p*-value obtainable would be 1/10,000 = 0.0001 in this case.

Permutation testing of inter-variable associations proceeds differently. In the case of a CCA, for example, there are two sets of variables, the response set and the explanatory set. We can measure how much inertia of the response data **Y** is explained by the explanatory data in **X** – this is the constrained inertia contained in the matrix $\mathbf{A}_X$ defined above. The null hypothesis is that there is no association, in which case every set of observations on the explanatory variables could be paired with any set of observations on the responses. So we randomize the order of one set of data, for example the rows of the explanatory data **X**, each time computing the amount of response inertia (or proportion) explained, doing this again thousands of times. The actual value of inertia explained is compared to the right tail of the null distribution to estimate the *p*-value.

Permutation testing can be used to give a guideline about the level at which a dendrogram should be cut to obtain significant clustering. Our approach has been to randomize the values within each column of data, that is shuffle them up randomly, assuming the columns contain the variables of the data, and recomputed the dendrogram each time. The node levels are stored for each dendrogram computed and this gives an idea of the null distribution of each level for data where there is no structure between the variables. The node levels of the actual dendrogram computed on the original data are then compared to these null distributions to obtain a *p*-value for each node. Here we are looking for values in the left tail of the respective null distributions, because significant clustering would be when the node levels are generally low in value. There can be several significant *p*-values in this case, and the final choice is based on these, substantive knowledge and the number of groups being sought.

Permutation testing can be similarly used for deciding on the dimensionality of the ordination solution. The columns of data are similarly randomized, each giving new parts of variance on the recomputed dimensions. This is done thousands of times, generating a null distribution of the parts of variance for the first dimension, second dimension, and so on. The original parts of inertia are compared to their corresponding null distributions to estimate a *p*-value for each dimension. In this case, *p*-values will generally increase for successive dimensions, and an obvious cut-off will appear, which usually coincides with the rule of thumb based on the *scree plot* of the eigenvalues.

To illustrate the use of bootstrapping for this last example, suppose we want a confidence region around the percentages of variance in a PCA, CA or LRA. The *I* rows of the data matrix are sampled, with replacement, until we have

a bootstrap sample, also of I rows. This means that some rows can be chosen more than once, others not at all – this differs from permutation testing where observations are simply re-arranged in a random order. For each bootstrap sample the multivariate method is recomputed and the percentages of inertia stored, and this is repeated thousands of times. This procedure results in an estimated distribution of percentages of inertia for each dimension, and a 95% confidence interval for each can be determined by cutting off 2.5% of the values on either tail of the distribution.

Statistical modelling

In the situation where we relate a single response to a set of explanatory variables, regular statistical modelling can be applied. Generalized linear modelling, generalized additive modelling and classification and regression trees, are alternative ways to model this relationship.

The most restrictive is generalized linear modelling (GLM), since it assumes that the effects of the explanatory variables are linear. But the way the linear effect translates to a change in the conditional mean of the response, called the *link function*, is different depending on the measurement scale of the response. The three most common types of responses are interval-scale continuous, ratio-scale count, and categorical binary:

Response variable	*Link function*	*Conditional distribution*	*Name of method*
Continuous:	Identity	Normal	Multiple linear regression
Count:	Logarithm	Poisson	Poisson regression
Categorical (binary):	Logit (log-odds)	Binomial	Logistic regression

The formulation of a generalized linear model is:

$$\eta(\overline{y}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \tag{A.20}$$

with inverse transformation

$$\overline{y} = \eta^{-1}(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots) \tag{A.21}$$

where η is the link function and the conditional distribution of the response is the one corresponding to it, with mean given by (A.21). The inverse function η^{-1} is $\exp(\cdot)$ for $\eta = \log$ and $\exp(\cdot)/[1 + \exp(\cdot)]$ for $\eta = \text{logit}$.

Generalized additive modelling (GAM) is like GLM, but with a freer and more flexible range of possibilities for the shape of the relationship between the

response and each explanatory variable. The linear model on the right of (A.20) is replaced by a sum of smooth terms: $\alpha + s(x_1) + s(x_2) + \cdots$. Each smooth function $s(\cdot)$ is quite general, and involves tying together several cubic functions called a *smoothing spline*. These functions have estimated degrees of freedom and their form can either confirm approximate linearity of the relationship or suggest a transformation of the explanatory variables to accommodate a nonlinear relationship.

Both GLMs and GAMs can include interactions between the explanatory variables. Classification and regression trees (CART) form an alternative nonparametric approach that uses simple rules for predicting the response by cutting up the range of the predictors, but specifically looking for interactions in the form of combinations of intervals of the predictors which maximize the fit to the response. The result is a decision tree that allows every case to be run down it, according to the conditions at each node, to arrive at a terminal node that predicts the response, either the mean or median for a continuous response, or a set of probabilities for a categorical response that lead to the prediction of the most likely category.