

Marti J. Anderson and Trevor J. Willis. 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* 84:511-525.

Appendix A. Statistical details of the canonical analysis of principal coordinates (CAP).

We provide here a detailed description of the canonical analysis of principal coordinate axes (CAP). First, let $\mathbf{Y} = (y_{ik})$ be a matrix of $i = 1, \dots, N$ real-valued observations on each of $k = 1, \dots, p$ variables. In an ecological context, these are commonly counts of individuals of each of p species. Consider also that the observations are classified a priori into g groups with sample sizes n_1, n_2, \dots, n_g and $n_1 + n_2 + \dots + n_g = N$. Let $\mathbf{D} = (d_{ij})$ be an $(N \times N)$ square symmetric matrix of distances or dissimilarities calculated between every pair of observation units.

The first step is to calculate principal coordinates from the distance matrix. These are orthogonal axes that place the points into Euclidean space as well as possible, while preserving the original distances or dissimilarities among the points (e.g., Gower 1966).

For principal coordinate analysis, let matrix $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$, then center this matrix on its rows and columns to produce matrix \mathbf{G} as follows:

$$\mathbf{G} = (g_{ij}) = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..} \quad (\text{A.1})$$

where $\bar{a}_{i.}$ is the average for row i , $\bar{a}_{.j}$ is the average for column j and $\bar{a}_{..}$ is the average value for the entire matrix \mathbf{A} . To get principal coordinate axes (Gower 1966), we do a spectral decomposition (eigenanalysis) of matrix \mathbf{G} to obtain

$$\mathbf{G} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}' \quad (\text{A.2})$$

where $\mathbf{\Lambda}$ is a diagonal matrix of ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ and $\mathbf{Q} = (q_{ij})$ is an $N \times N$ matrix of corresponding orthonormal eigenvectors. At least one eigenvalue will be zero, as only $N - 1$ axes are needed to describe the position of N points in Euclidean space. The column vectors of \mathbf{Q} are the orthonormal vectors we require. Commonly, these axes are each normalized so that their variance is equal to their corresponding eigenvalue. Then, the first two (or more) axes may be plotted for an unconstrained ordination. Our purpose here, however, is to use principal coordinate analysis in order to obtain an orthonormal Euclidean (i.e., Mahalanobis) representation of the data cloud. Thus, no such normalization is done.

The next step is to do a traditional canonical analysis on these orthonormal axes. Here is where our hypothesis comes into the analysis. In the case of a hypothesis concerning a priori groups, let \mathbf{X} be an $N \times (g - 1)$ design matrix containing orthogonal variables coding for group contrasts (e.g., see Appendix C of Legendre and Anderson 1999). In the case of a hypothesis concerning relationships of species data with another

set of quantitative variables (such as environmental variables), then \mathbf{X} contains this set of variables. Then let $\mathbf{H} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'$ be the $N \times N$ projection matrix of rank r (note that $r = (g - 1)$ if \mathbf{X} specifies groups). This is the usual idempotent “hat” matrix used for linear models (e.g., Neter et al. 1996).

To do the canonical analysis, we need to use only a subset, say the first m axes, of matrix \mathbf{Q} . The $N \times m$ matrix that contains the first m columns of \mathbf{Q} shall be denoted \mathbf{Q}_m . The canonical analysis is accomplished by spectral decomposition (eigenanalysis) of the $m \times m$ matrix $\mathbf{Q}'_m\mathbf{H}\mathbf{Q}_m$, thus:

$$\mathbf{Q}'_m\mathbf{H}\mathbf{Q}_m = \mathbf{U}'\Delta^2\mathbf{U} \quad (\text{A.3})$$

where Δ^2 is a diagonal matrix of ordered eigenvalues $\delta_1^2 \geq \delta_2^2 \geq \dots \geq \delta_s^2$ which are the squared canonical correlations, and \mathbf{U} is an $m \times s$ matrix of corresponding orthonormal eigenvectors. The number of canonical axes will be $s = \min(r, m)$. The canonical variable scores for ordination are then obtained as $\mathbf{Q}^* = \mathbf{Q}_m\mathbf{U}$. For plotting, the canonical variable scores are standardized by the square root of their corresponding eigenvalue (i.e., δ_i).

The following test statistics for differences among groups, based on the CAP approach, have been suggested (Anderson and Robinson, *in press*). First, we have the trace statistic

$$tr[\mathbf{Q}'_m\mathbf{H}\mathbf{Q}_m] = tr[\Delta^2] = \sum_{k=1}^s \delta_k^2, \quad (\text{A.4})$$

which is equal to the sum of squared canonical correlations. Second, we have the maximum root statistic

$$\delta_1^2, \quad (\text{A.5})$$

which is the first squared canonical correlation. Each of these can be tested by permutation of the observation vectors of \mathbf{Y} (or, equivalently, by permutation of the row vectors of \mathbf{Q}_m), the only assumption being that the observation vectors are exchangeable under the null hypothesis of no relationship with \mathbf{X} or no differences among the groups (Anderson 2001). Note that in the case of Euclidean distances being used to calculate \mathbf{D} and when $m = p < N$, then these two test statistics correspond to the trace and maximum root statistics for traditional CDA or CCorA (Anderson and Robinson, *in press*).

Formulation using singular value decomposition

An alternative formulation is obtained using singular value decomposition. A Cholesky decomposition of matrix \mathbf{H} gives matrix \mathbf{B} where $\mathbf{B}\mathbf{B}' = \mathbf{H}$. Consider the singular value decomposition of matrix $\mathbf{Q}'_m\mathbf{B}$ as follows:

$$\mathbf{Q}'_m \mathbf{B} = \mathbf{U} \Delta \mathbf{V}' \quad (\text{A.6})$$

where \mathbf{U} and \mathbf{V} are semi-orthogonal matrices, $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_s$, with $s = \min(r, m)$ and $\Delta = (\Delta^2)^{0.5}$ is a diagonal matrix of positive eigenvalues $\delta_1 \geq \delta_2 \geq \dots \geq \delta_s$, which are the canonical correlations. Then $\mathbf{Q}^* = \mathbf{Q}_m \mathbf{U}$ are the canonical variables, which, for plotting, are standardized by their corresponding eigenvalue (i.e., δ_i), as described above.

Classification

The classification of a new multivariate observation can be achieved by first considering the distances (or dissimilarities) of the new $(N + 1)$ th observation point from each of the N previously observed points. Then, using equation (A.1), the values it would take in the \mathbf{G} matrix would be

$$g_{(N+1)i} = -\frac{1}{2} \left[d_{(N+1)i}^2 - \frac{1}{N} \sum_{j=1}^N d_{ij}^2 - \frac{1}{N} \sum_{i=1}^N d_{(N+1)i}^2 + \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^2 \right]. \quad (\text{A.7})$$

The eigenvectors \mathbf{Q} obtained from the spectral decomposition of \mathbf{G} (shown in Eq. A.2 above) can also be written as

$$g_{ij} = \sum_{\ell=1}^N \lambda_{\ell} q_{\ell i} q'_{\ell j} \quad (\text{A.8})$$

and rearranging this in terms of the principal coordinate eigenvector for a new $(N + 1)$ th observation gives

$$q_{\ell(N+1)} = \sum_{i=1}^N g_{(N+1)i} q_{\ell i} / N \lambda_{\ell}. \quad (\text{A.9})$$

Then, the position of the new observation in the canonical space is obtained by multiplying the values of $q_{\ell(N+1)}$ for each of the $\ell = 1, \dots, m$ principal coordinates by the appropriate values in the matrix of canonical eigenvectors \mathbf{U} . That is,

$$q_{k(N+1)}^* = \sum_{\ell=1}^m q_{\ell(N+1)} u_{\ell k} \quad (\text{A.10})$$

for $k = 1, \dots, s$ canonical axes. We can then classify the observation into one of the groups by observing which group centroid is the closest to it in the canonical space, as measured by Euclidean distance and where the canonical axes have been standardized in the usual way according to the canonical correlations.

Goodness of fit

Consider that $\mathbf{Q}'_m \mathbf{U}$ and \mathbf{BV} are $N \times s$ arrays with correlation Δ . So, a least-squares estimate of \mathbf{BV} is $\mathbf{Q}'_m \mathbf{U} \Delta$ and we can estimate \mathbf{B} by $\mathbf{Q}'_m \mathbf{U} \Delta \mathbf{V}'$ or by $\mathbf{Q}_m \mathbf{Q}'_m \mathbf{B}$. Using a “leave-one-out” approach (e.g., Lachenbruch and Mickey 1968), we can calculate the position in the principal coordinate space of the i th observation that has been left out, using only its distances from each other point, as described in the previous section on classification. This will be a $1 \times m$ vector which we will call \dot{q}_i , whose corresponding row in matrix \mathbf{B} is the $1 \times r$ vector b_i . Next, consider a prediction of b_i using \dot{q}_i and matrices \mathbf{B} and \mathbf{Q}_m which have been calculated from the remaining $N - 1$ observations, which we shall call $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{Q}}_m$, respectively. The residual sum of squares for a given choice of m is then

$$R = \sum_{i=1}^N \text{tr}((b_i - \dot{q}_i \tilde{\mathbf{Q}}'_m \tilde{\mathbf{B}})'(b_i - \dot{q}_i \tilde{\mathbf{Q}}'_m \tilde{\mathbf{B}})) \quad (\text{A.11})$$

and a choice of m can be made where this criterion is minimized. In the case of a priori groups, the choice of m can also be made by calculating the “leave-one-out” misclassification error for increasing values of m , as described in the text. In practice, these two criteria will generally indicate the same or similar values for m , but note that they may differ for reasons given in the discussions by Williams (1983, p. 1290) and Seber (1984, Section 6.10, pp. 337-343).

Literature Cited

- Anderson, M. J. 2001. Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences* **58**:626-639.
- Anderson, M. J., and J. Robinson. *In press*. Generalised discriminant analysis based on distances. *Australian and New Zealand Journal of Statistics*.
- Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**:325-338.
- Lachenbruch, P. A., and M. R. Mickey. 1968. Estimation of error rates in discriminant analysis. *Technometrics* **10**:1-11.
- Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* **69**:1-24.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models, fourth edition. Irwin, Chicago, Illinois, USA.

Seber, G. A. F. 1984. Multivariate observations. John Wiley and Sons, New York, New York, USA.

Williams, B. K. 1983. Some observations of the use of discriminant analysis in ecology. *Ecology* **64**:1283-1291.