BIO 682
Multivariate Statistics
Spring 2009

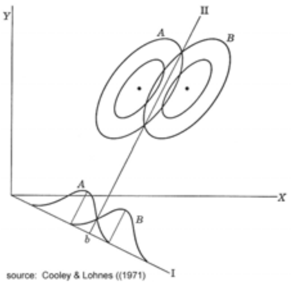
Steve Shuster

http://www4.nau.edu/shustercourses/BIO682/index.htm

Lecture 12

Canonical Discriminant Analysis (CDA)

- Canonical discriminant analysis is used as a means of distinguishing among a group of samples from potentially different populations.



source: Cooley & Lohnes ((1971)

Figure 9.1.

Canonical Discriminant Analysis (CDA)

- The goals are to:

(1) find the axis of greatest discrimination between groups identified ***a priori***,

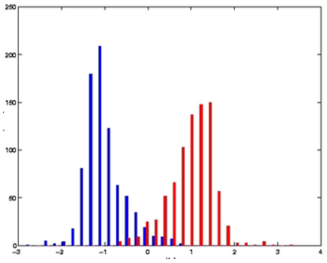
(2) test whether the means of those groups along that axis are significantly different, and

(3) attempt to assign individual specimens to groups.

Canonical Discriminant Analysis (CDA)

- When there are two groups to be separated, the technique is known as discriminant function analysis (DFA);
- with more than two groups, the same question is addressed through canonical variates analysis (CVA).
- Often these terms are used interchangeably.

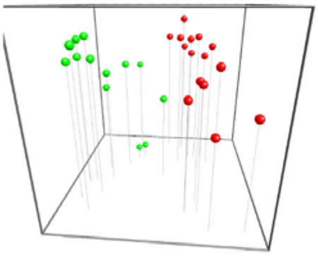
Canonical Discriminant Analysis (CDA)



- The key assumption of canonical discriminant analysis is that all individuals can be assigned to one and only one group ***in advance***, through some means external to the data being analyzed;

Canonical Discriminant Analysis (CDA)

- in this way it is different from PCA and factor analysis, which assume that any substructuring in the data is ***unknown*** prior to the analysis.



History of Canonical Discriminant Analysis

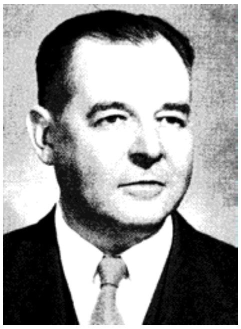
- Discriminant analysis is a graphical version of MANOVA,
- It "looks" for combinations of observed variables that indicate a significant difference in the means of treatment groups.

History of Canonical Discriminant Analysis

The techniques were developed in the 1930s, by 3 people working on the same problem from different perspectives.

- R.A. Fisher (UK) was interested in a technique for distinguishing between two groups on the basis of multivariate data; his contribution was the Fisher linear discriminant function.

History of Canonical Discriminant Analysis



- Harold Hotelling (US) developed the Hotelling's T^2 test as a means of testing for significant differences in the position of the centroids between two multivariate samples.

History of Canonical Discriminant Analysis

- Mahalanobis (India) was attempting to find a way of measuring the multivariate distance between the centroids of two samples; his Mahalanobis D^2 distance is an extension of the Pythagorean theorem for sets of correlated variables.

History of Canonical Discriminant Analysis

- These techniques were united in **canonical discriminant analysis** (CDA),
- It measures:

  1. the distance between the means from two samples through the Mahalanobis distance,

  2. determines whether that distance is significantly different from zero using Hotelling's T2 or other similar test, and

  3. develops a regression equation (linear discriminant function) allowing us to assign new specimens to one of the two groups.

History of Canonical Discriminant Analysis

- In the case of more than two groups, we also determine the primary axes of among-group variation, known as canonical variates.

Principles of Canonical Discriminant Analysis

- PCA is used when the data all come from one sample (group, population),
- CDA (discriminant analysis) assumes there are multiple groups that can be unambiguously defined in advance.

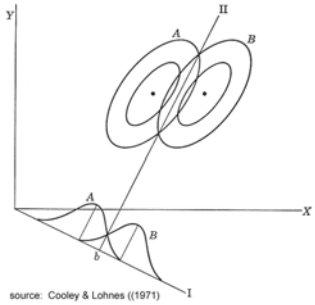
Principles of Canonical Discriminant Analysis

While PCA maximizes the total variation explained by each principal component, DFA/CVA maximizes the among-group variance explained by each canonical variate.

- As such, it focuses not on the overall variation in the data, but on the extent to which that variation is partitioned among groups (maximizing the separation of groups).

Principles of Canonical Discriminant Analysis

- This separation of groups (for two groups) is accomplished by finding a linear combination of the original variables for which the F value between groups is maximized.



source: Cooley & Lohnes ((1971)

Figure 9.1.

Principles of Canonical Discriminant Analysis

- In the cases of more than two groups, we get a set of canonical discriminant functions (canonical variates),
- These are multiple axes that separate sets of groups
- Assuming there are more variables than groups, there will be (m-1) canonical variates, where m is the number of groups.

Principles of Canonical Discriminant Analysis

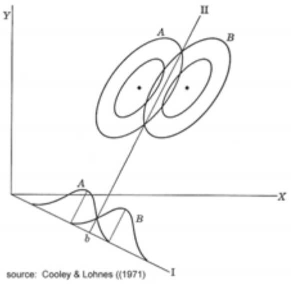
- The difference between Fisher discriminant functions and canonical discriminant functions;
- discriminant functions connect pairs of centroids, while canonical variates summarize the major axes of among-group variation.

Principles of Canonical Discriminant Analysis

- Canonical discriminant functions fit the equation
- $Z_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n$
- where the a values are coefficients derived from eigenanalysis of the matrix of between group variation, and the Z values are scores (coordinates) along the derived axis.
- If we bisect this axis, we can use the scores to assign individuals to groups at each end of the axis.

Canonical Discriminant Analysis (CDA)

- Canonical discriminant analysis is used as a means of distinguishing among a group of samples from potentially different populations.



source: Cooley & Lohnes ((1971)

Figure 9.1.

Nonmetric Multidimensional Scaling (NMDS)

- This is sometimes labeled "Multidimensional Scaling (MDS)", although this term has been used for PCoA (Principal Coordinates Analysis).

Nonmetric Multidimensional Scaling (NMDS)

- NMDS is very computer intensive; it has only recently become feasible for large data sets on the microcomputer.
- NMDS Maximizes rank-order correlation between distance measures and distance in ordination space.

Nonmetric Multidimensional Scaling (NMDS)

- Points are moved in ordination space to minimize "stress". Stress is a measure of the mismatch between the two kinds of distance.

Nonmetric Multidimensional Scaling (NMDS)

- NMDS assumes is that dissimilarity is monotonically related to ecological distance (i.e., order is ***preserved***).
- Gauch (1982) is WRONG is stating that NMDS assumes species have monotonic relationships to gradients.

Nonmetric Multidimensional Scaling (NMDS)

- The user must pre-specify a number of dimensions, or examine a plot of "stress" as a function of number of axes, and select the number of dimensions ***a posteriori***.

Nonmetric Multidimensional Scaling (NMDS)

- The configuration will change as the number of axes change.
- There is no guarantee that the correct (lowest stress) solution will be found, though it is widely assumed that this is not a big problem.

Stress

$$S = \left\{\left[\sum_{i<j}[d_{ij} - \hat{d}_{ij}]^2\right]\Big/\left[\sum_{i<j} d_{ij}{}^2\right]\right\}^{1/2}.$$

Where d_{ij} is the distance in ordination space between samples i and j, and

for each sample pair, i and j; the regression between the distance measure and distance produces a value, d^ij

Stress

$$S = \left\{\left[\sum_{i<j}[d_{ij} - \hat{d}_{ij}]^2\right]\Big/\left[\sum_{i<j} d_{ij}{}^2\right]\right\}^{1/2}.$$

“The smaller the stress, the closer the relation between dissimilarities and distance approaches monotonicity.”

“Stress cannot reliably be used as a measure of ordination efficiency (the stress value will in fact ***always decrease*** with ***increasing dimensionality*** of the ordination or solution space irrespective of the true dimensionality of the data)."

Fasham et al. 1977

Nonmetric Multidimensional Scaling (NMDS) a là Whitham et al.

- NMDS is based upon pair-wise community dissimilarities generated by the Bray-Curtis dissimilarity coefficient.
- Thus, in our analyses, each clonal replicate within each tree genotype generates a single NMDS score, which summarizes the composition of each community.

The Bray-Curtis Dissimilarity Coefficient

$$d_{ij} = \frac{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|}{\sum_{k=1}^{n} \left(x_{ik} + x_{jk} \right)}$$

Where i and j refer to the cell values in each k-th dissimilarity matrix and dij is the dissimilarity score for that matrix.

The Bray-Curtis Dissimilarity Coefficient

	Value1	Value2	Value3	Value4
Object A	**0**	**3**	**4**	**5**
Object B	**7**	**6**	**3**	**-1**

$$d_{BA} = \frac{|0-7| + |3-6| + |4-3| + |5+1|}{(0+7) + (3+6) + (4+3) + (5-1)}$$

$$= \frac{7+3+1+6}{7+9+7+4} = \frac{17}{27} = 0.630$$

The Bray-Curtis Dissimilarity Coefficient is appropriate for ecological data because:

(1) its value is 1 when samples have no species in common,

(2) its value is 0 when samples are identical,

(3) species that are jointly absent from samples do not affect the dissimilarity value among samples,

(4) the addition of samples does not affect the dissimilarity values for other pairs of samples,

(5) it registers differences in the total abundance among samples when the relative abundances are identical (Clarke and Warwick 2001), and

(6) it successfully recovers simulated ecological gradients in ordination (Faith et al. 1987; Minchin 1987).

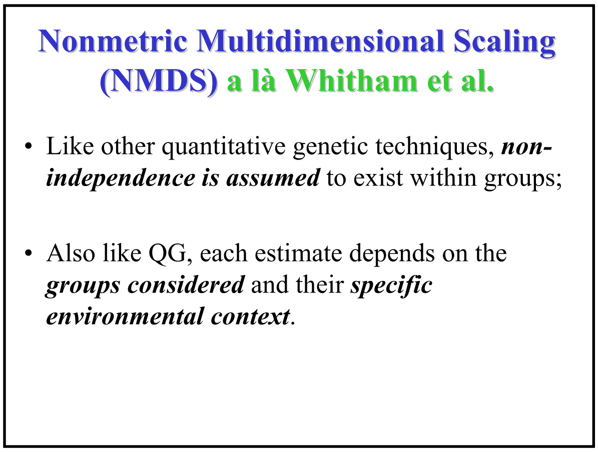
Nonmetric Multidimensional Scaling (NMDS) a là Whitham et al.

- Performs well when handling data with high beta diversity (Fasham 1977), but also efficiently handles data in which beta diversity is low (Minchin 1987).
- This point is critical in the cottonwood system in which turnover rates approach 80% among trees of the same cross type (Wimp, unpublished data).

Nonmetric Multidimensional Scaling (NMDS) a là Whitham et al.

- It provides a robust ordination technique for community analysis because:

1. It captures the consequences of trait interactions among host cottonwoods and their dependent arthropods, and
2. it summarizes them as a single community phenotype.



Nonmetric Multidimensional Scaling (NMDS) a là Whitham et al.

- Like other quantitative genetic techniques, ***non-independence is assumed*** to exist within groups;

- Also like QG, each estimate depends on the ***groups considered*** and their ***specific environmental context***.