

# Scientific Computing for Biologists

## Lecture 11: Discriminant Analysis and Classification

Instructor: Paul M. Magwene

25 October 2011

# Outline of Lecture

- Fisher's Discriminant Function
- Canonical Variates Analysis (CVA)
  - Geometric and Algebraic View
  - Similarities and differences between CVA and PCA
  - Interpreting CVA

# Overview of Discriminant Analysis

## Discrimination

Given an  $n \times p$  data matrix,  $\mathbf{X}$ , and a grouping of the  $n$  specimens into  $g$  groups, find the linear combination of the variables,  $\mathbf{a}'\mathbf{x}$  that best discriminates between the groups.

## Classification

Given  $g$  groups, define a function that assigns an object with unknown assignment to the 'best' group.

# Fisher's Discriminant Function

- Applies to the two-group case.
- Solution: find  $\mathbf{a}$  that maximizes the ratio of the squared group mean difference to within-group variance:

$$F = \frac{(\mathbf{a}'\bar{\mathbf{x}}_1 - \mathbf{a}'\bar{\mathbf{x}}_2)^2}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

where

- $\bar{\mathbf{x}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_{1i}$
- $\bar{\mathbf{x}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{x}_{2i}$
- $\mathbf{W} = \frac{1}{n_1+n_2-2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$  (w/in-group pooled covariance matrix)
- $n_i$  indicates the number of observations in the  $i$ th group and the  $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$  represent the specific observations (as vectors).

# Geometry of the Two-Group Discriminant Function

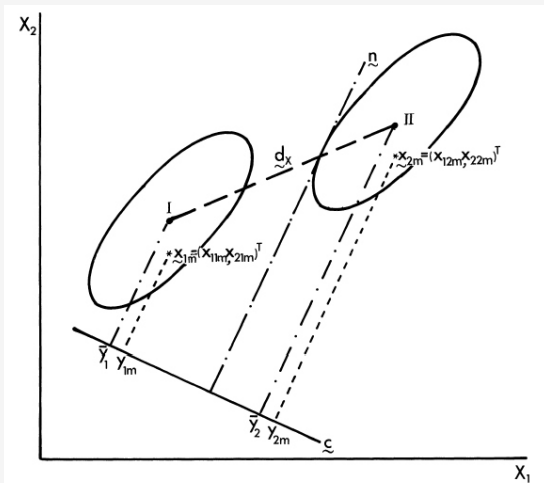


FIG. 4.—Representation of the discriminant function for two groups and two variables, showing the group means and associated 95% concentration ellipses. The vector  $e$  is the discriminant vector. The points  $\bar{y}_1$  and  $\bar{y}_2$  represent the discriminant means for the two groups.

The discriminant vector can be constructed by drawing the tangent  $n$  to the concentration ellipse at the point of intersection with the line  $d$  joining the group means; the discriminant vector is orthogonal to the tangent  $n$ .

# Fisher's LDF

$$F = \frac{(\mathbf{a}'\bar{\mathbf{x}}_1 - \mathbf{a}'\bar{\mathbf{x}}_2)^2}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

Maximizing  $F$  gives:

$$\mathbf{a} = c\mathbf{W}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

where  $c$  is an arbitrary constant (usually taken to be 1).

# Fisher's LDF as Classification

Fisher's solution can also be setup as a classification solution using regression.

- setup a dummy variable,  $y$  that takes the values:
  - $y_1 = n_2/(n_1 + n_2)$  for observations in group 1
  - $y_1 = -n_1/(n_1 + n_2)$  for observations in group 2
- Solve the standard multivariate regression,  $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$
- Allocate unknown individual to group 1 if it's predicted  $y$  is closer to  $y_1$  than to  $y_2$ , otherwise assign to group 2.

## What if there are more than two groups?

The multi-group equivalent of Fisher's LDF is called 'Canonical Variate Analysis' (CVA).

- straight forward extension of Fisher's solution
- Find **a** that maximizes the ratio of between-group to within-group variance:

$$F = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

- **W** is within-group matrix (as defined previously)
- **B** is the between-group covariance matrix
  - $\mathbf{B}_w = \frac{1}{g-1} \sum_{i=1}^g n_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$  where  $n_i$  is the sample size in group  $i$  (weighted version)
  - $\mathbf{B}_u = \frac{1}{g-1} \sum_{i=1}^g (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$  (unweighted version)



# Geometry of CVA

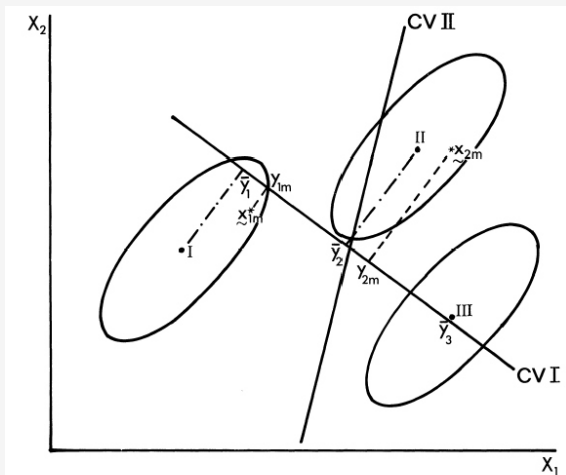


FIG. 2.—Representation of the canonical vectors for three groups and two variables. The group means (I, II and III) and 95% concentration ellipses are shown. The vectors CV I and CV II are the two canonical vectors. In the text, CV I =  $\mathbf{c}$ . The points  $y_{1m}$  and  $y_{2m}$  represent the canonical variate scores corresponding to the first canonical vector for the observations  $\tilde{x}_{1m}$  and  $\tilde{x}_{2m}$ .

# CVA Solution

Maximizing  $F$  leads to the following:

$$(\mathbf{B} - \lambda \mathbf{W})\mathbf{a} = \mathbf{0}$$

- $\lambda$  is an eigenvalue of  $\mathbf{W}^{-1}\mathbf{B}$
- $\mathbf{a}$  is an eigenvector of  $\mathbf{W}^{-1}\mathbf{B}$

There will be  $s = \min(p, g - 1)$  non-zero eigenvalues.

Organize the eigenvectors,  $\mathbf{a}_i$ , as columns of a  $p \times s$  matrix  $\mathbf{A}$ .

- The **canonical variates** are given by  $\mathbf{y} = \mathbf{A}'\mathbf{x}$
- The mean of the  $i$ th group in the canonical variates space is given by  $\bar{\mathbf{y}}_i = \mathbf{A}'\bar{\mathbf{x}}_i$

# CVA as a two-stage rotation I

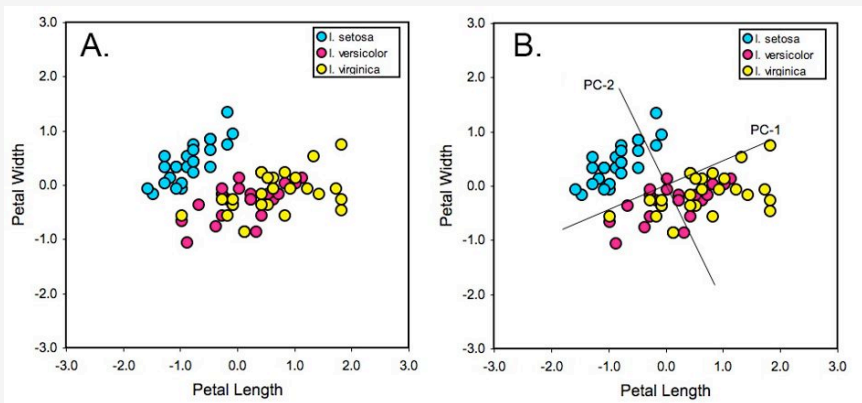


Figure 2. Stage 1 CVA implicit rotation. A. Scatterplot of first two *Iris* variables for example dataset. B. Orientation of the two pooled-sample principal components of the within-groups SSQCP matrix ( $W$ ).

From MacLeod 2010, *PalaeoMath* newsletter

## CVA as a two-stage rotation II

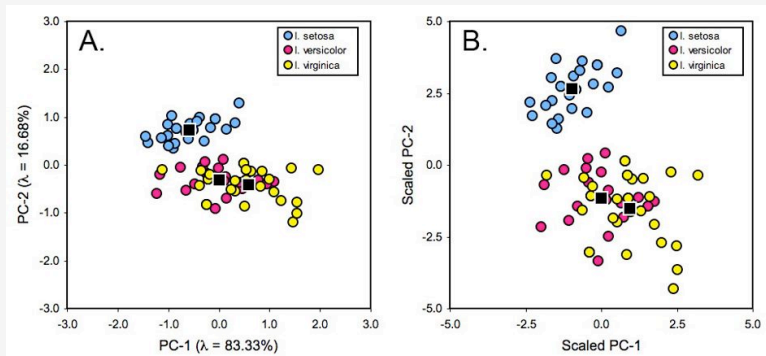


Figure 3. Intermediate scaling operation of a CVA. A. Scatterplot of *Iris* PC scores for the Stage 1 rotation (see Fig. 2). B. Result of scaling the two within-groups principal components by the square roots of their associated eigenvalues. Note difference in separation of the group centroids (black squares) after scaling.

From MacLeod 2010, *PalaeoMath* newsletter

## CVA as a two-stage rotation III

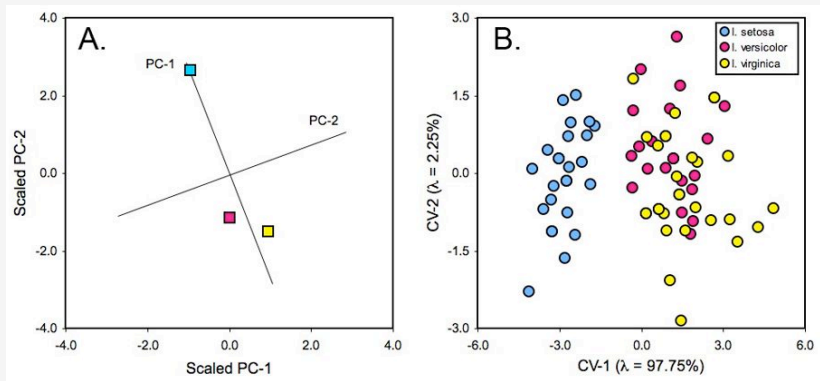


Figure 4. Stage 2 CVA implicit rotation. A. *Iris* group centroids plotted in the within-groups orthogonal-orthonormal space (see Fig. 3B) with between groups PC (= CVA) axes. B. Reduced *Iris* dataset plotted in the space defined by the CVA axes.

From MacLeod 2010, *PalaeoMath* newsletter

## CVA as a two-stage rotation IV

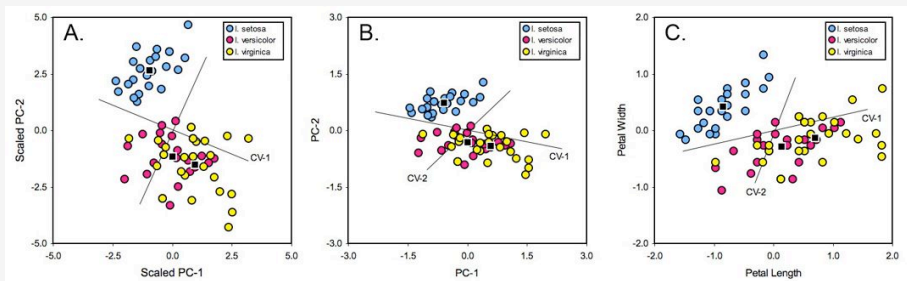


Figure 5. Back-calculation of final CVA axis orientation through the intermediate stages of the canonical rotations and scalings. A. Orientation of final CVA axes in the space of the scaled within-groups principal components (compare to Fig. 3A). B. Orientation of final CVA axes in the space of the raw within-groups principal components (compare to Fig. 3B). C. Orientation of final CVA axes in the space of the original variables (compare to Fig. 2).

*From MacLeod 2010, PalaeoMath newsletter*

# Similarities and Differences between CVA and PCA

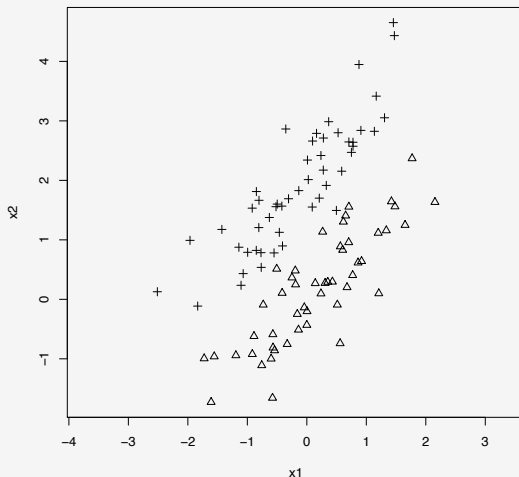
## PCA:

- Uncorrelated over the whole sample
- orthogonal transformation from the original variates,  $\mathbf{x}$ , to the new variates  $\mathbf{y}$ . PC axes at right angles to each other in the space of the original variables.

## CVA:

- Canonical variates are uncorrelated both *within* and *between* groups
- Canonical variates have equal variance *within* groups, but in decreasing order *between* groups
- non-orthogonal transformation, CV axes *not* at right angle to each other in the original frame of reference.

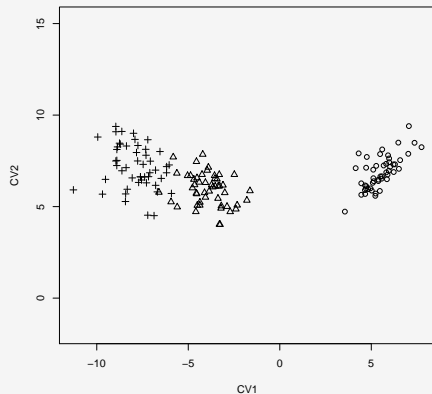
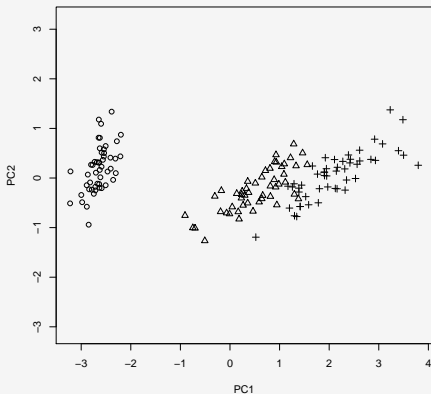
# PCA vs CVA: A Motivating Example



What is the direction of PC1? What is the direction of CVI?



# PCA vs. CVA: Anderson's Iris Data



# Are any of the groups significantly different in the canonical variate space?

To test:

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_3$
- $H_1$ : at least one  $\mu_i$  differs from the rest

A couple of approaches:

- Compare the largest eigenvalue,  $l_1$ , of  $\mathbf{W}^{-1}\mathbf{B}$  to critical values in a F-table.  $H_0$  is rejected for large values ( $> 1$ ).
- Likelihood approach:
  - Wilks' lambda,  $\Lambda = |\mathbf{W}|/|\mathbf{B} + \mathbf{W}| = \prod_{i=1}^p (1 + l_i)^{-1}$
  - there is an approximation that has a  $\chi^2$  distribution.

Both boil down to a consideration of eigenvalues of  $\mathbf{W}^{-1}\mathbf{B}$ .

## Which groups are different? Where does an unassigned observation belong?

Within groups the canonical variates are:

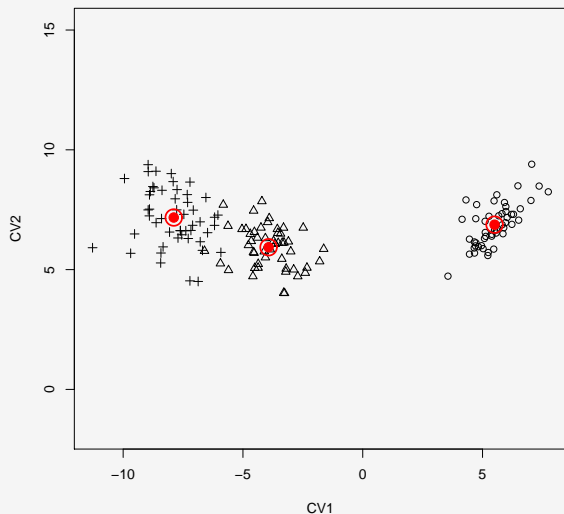
- uncorrelated
- have unit variance

If we assume multivariate normality of the data then we can exploit this to draw confidence intervals around the group means in the canonical variate space.

A  $100(1-\alpha)$  percent confidence region for the true mean  $\mathbf{v}_i$  is given by:

- hypersphere centered at  $\bar{\mathbf{y}}_i$
- with radius  $(\chi^2_{\alpha,r}/n_i)^{1/2}$  where  $r$  is the number of canonical variate dimensions considered

# Illustration of group means and tolerance regions



# Which variables are most important in CVA?

## Question

Which variables are most 'important' in distinguishing between the groups?

Consider the coefficients  $\mathbf{a}_i$

- large coefficients may be due to *either* large between-group variability *or* small within-group variability of the corresponding variable
- for interpretation it's better to consider modified coefficient,  $\mathbf{a}_i^* = (a_{i1}^*, \dots, a_{ip}^*)$  where the  $a_{ij}^*$  are given by  $a_{ij}^* = a_{ij} \sqrt{w_{jj}}$  [ $w_{jj}$  are the diagonal elements of  $\mathbf{W}$ ].