

# Scientific Computing for Biologists

## Lecture 7

### Testing for Group Effects and Contrasting Groups: ANOVA and Discriminant Analysis

Instructor: Paul M. Magwene

09 October 2012

# Outline of Lecture

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

# Testing the Regression Effects

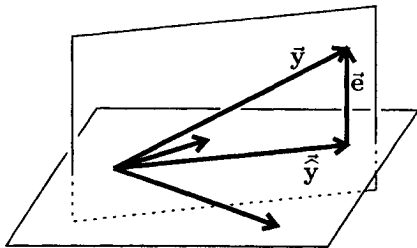


Figure : Geometry of multiple regression.

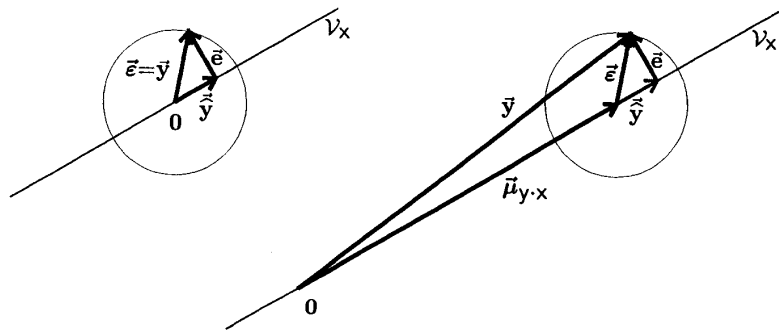
$$|\vec{y}|^2 = SS_{\text{total}}$$

$$|\vec{\hat{y}}|^2 = SS_{\text{regression}} = R^2 SS_{\text{total}}$$

$$|\vec{e}|^2 = SS_{\text{residual}} = (1 - R^2) SS_{\text{total}}$$

Is my regression significant?  $\Rightarrow$  Is  $|\vec{\hat{y}}|^2$  large relative to  $|\vec{e}|^2$ ?

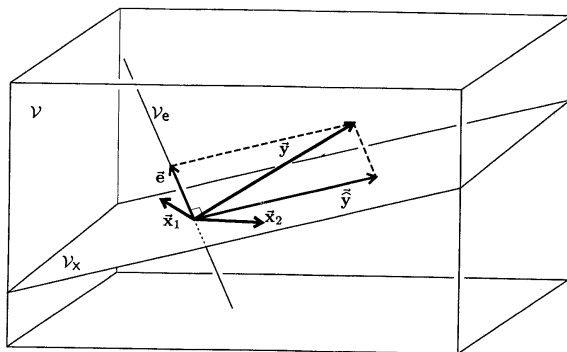
# Geometry of the Population Regression Model



**Figure :** **Left:** null hypothesis of no regression effects is true; **Right:** null model is false.

**Null Hypothesis:**  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$

# Dimensionality of Regression Subspaces



$$\dim(V_{\text{total}}) = N \quad \text{(total)}$$

$$\dim(V_1) = 1 \quad \text{(mean effect)}$$

$$\dim(V_x) = p \quad \text{(effect space)}$$

$$\dim(V_e) = N - p - 1 \quad \text{(error space)}$$

## Comparing the Effect Space and the Error Space

To compare the squared length of  $|\vec{\hat{y}}|^2$  and  $|\vec{e}|^2$  we divide them by the dimension of the subspaces in which they lie.

$$M(\vec{\hat{y}}) = \frac{|\vec{\hat{y}}|^2}{\dim(\mathcal{V}_x)}$$
$$M(\vec{e}) = \frac{|\vec{e}|^2}{\dim(\mathcal{V}_e)}$$

We compare these by defining a statistic,  $F$ :

$$F = \frac{M(\vec{\hat{y}})}{M(\vec{e})} = \frac{\dim(\mathcal{V}_e)|\vec{\hat{y}}|^2}{\dim(\mathcal{V}_x)|\vec{e}|^2}$$
$$= \frac{(N - p - 1)R^2}{p(1 - R^2)}$$

When null hypothesis is true,  $F \approx 1$ ; when it is false,  $F \gg 1$ .

## Two-group ANOVA as Regression

We can also use a geometric perspective to test whether the mean of a variable differs between two groups of subjects.

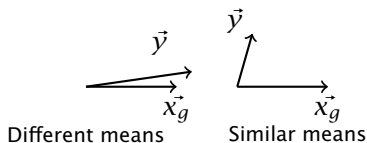
- Setup a 'dummy variable' as the predictor  $X_g$ . We assign all subjects in group 1 the value 1 and all subjects in group 2 the value -1 on the dummy variable. We then regress the variable of interest,  $Y$ , on  $X_g$ .

$$y = X_g b + e$$

Group	Raw		Centered	
	$Y_i$	$X_i$	$y_i$	$x_i$
1	2	-1	-3	$-\frac{4}{3}$
	3	-1	-1	$-\frac{4}{3}$
2	5	1	0	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	7	1	2	$\frac{2}{3}$
Mean	5	$\frac{1}{3}$	0	0

## Two-group ANOVA as Regression, cont

- When the means are different in the two groups,  $X_g$  will be a good predictor of the variable of interest, hence  $\vec{y}$  and  $\vec{x}_g$  will have a small angle between them.
- When the means in the two groups are similar, the dummy variable will not be a good predictor. Hence the angle between  $\vec{y}$  and  $\vec{x}_g$  will be large.





# Multi-way ANOVA as Regression

- Exactly the same idea applies to  $g$  groups, except now instead of one grouping variable, we define  $g - 1$  grouping variables,  $\dim(X_g) = g - 1$ .
- Then we calculate the multiple regression as we did before:

$$y = Xb + e$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} ; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1g} \\ 1 & x_{21} & x_{22} & \cdots & x_{2g} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ng} \end{bmatrix} ;$$

Estimate  $b$  as:

$$b = (X^T X)^{-1} X^T y$$

# How Do We Construct the Grouping Matrix, $X_g$ ?

Two common methods are:

- 1 Dummy coding – define a set of  $g$  grouping variables, where values take either 0 or 1, depending on group membership, but *use only the first  $g - 1$  columns*:

$$U_j = \begin{cases} 1, & \text{for every subject in group } j, \\ 0, & \text{for all other subjects.} \end{cases}$$

and

$$X_g = [U_1, U_2, \dots, U_{g-1}]$$

- 2 Effect coding – define the  $U_j$  as above, and set:

$$X_g = [U_1 - U_g, U_2 - U_g, \dots, U_{g-1} - U_g]$$

In general, effect coding is more similar to standard ANOVA contrasts.

## ANOVA: Example Data Set

	$g_1$	$g_2$	$g_3$	$g_4$	
	20	21	17	8	
	17	16	16	11	
	17	14	15	8	
$M_{g.}$	18	17	16	9	$M_{..} = 15$

$$y = \begin{bmatrix} 20 \\ 17 \\ 17 \\ 21 \\ 16 \\ 14 \\ 17 \\ 16 \\ 15 \\ 8 \\ 11 \\ 8 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

## ANOVA: Example Data Set, cont

Solving for  $\mathbf{b}$  we find:

$$\mathbf{b} = \begin{bmatrix} 15 \\ 3 \\ 2 \\ 1 \end{bmatrix}, \quad |\hat{\mathbf{y}}|^2 = 150, \quad |\mathbf{e}|^2 = 40$$

Since,  $\dim(\mathcal{V}_x) = 3$ , and  $\dim(\mathcal{V}_e) = 8$ , we get:

$$F = \frac{\dim(\mathcal{V}_e) |\vec{\hat{\mathbf{y}}}|^2}{\dim(\mathcal{V}_x) |\vec{\mathbf{e}}|^2} = 10$$

Here's the more conventional ANOVA table for the same data:

Source	df	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Pr(F)</i>
Experimental	3	150	50	10	.0044
Error	8	40	5		
Total	11	190			

# Overview of Discriminant Analysis

## Discrimination

Given an  $n \times p$  data matrix,  $X$ , and a grouping of the  $n$  specimens into  $g$  groups, find the linear combination of the variables,  $a'x$  that best discriminates between the groups (using  $'$  to indicate transpose).

## 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  $a$  that maximizes the ratio of the squared group mean difference to within-group variance:

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

where

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

# Geometry of the Two-Group Discriminant Function

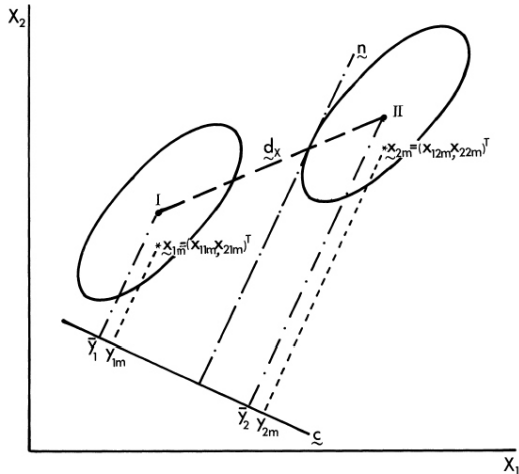


FIG. 4.—Representation of the discriminant function for two groups and two variables, showing the group means and associated 95% concentration ellipses. The vector  $e$  is the discriminant vector. The points  $\bar{y}_1$  and  $\bar{y}_2$  represent the discriminant means for the two groups.

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

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

Maximizing  $F$  gives:

$$a = cW^{-1}(\bar{x}_1 - \bar{x}_2)$$

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
  - $y_2 = -n_1/(n_1 + n_2)$  for observations in group 2
- Solve the standard multivariate regression,  $y = Xb + e$
- 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:

$$F = \frac{a'Ba}{a'Wa}$$

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

# Geometry of CVA

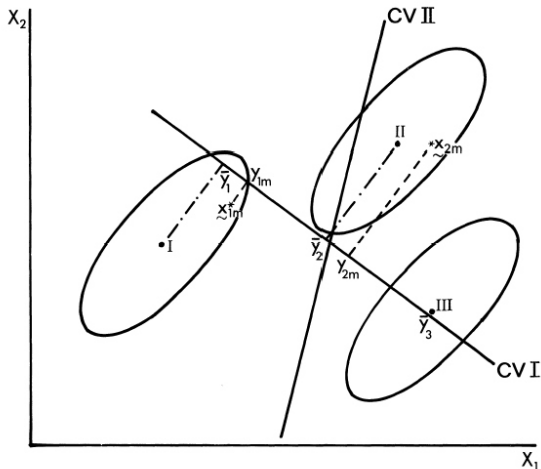


FIG. 2.—Representation of the canonical vectors for three groups and two variables. The group means (I, II and III) and 95% concentration ellipses are shown. The vectors CV I and CV II are the two canonical vectors. In the text, CV I =  $c$ . The points  $y_{1m}$  and  $y_{2m}$  represent the canonical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

Maximizing  $F$  leads to the following:

$$(B - lW)a = 0$$

- $l$  is an eigenvalue of  $W^{-1}B$
- $a$  is an eigenvector of  $W^{-1}B$

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

Organize the eigenvectors,  $a_i$ , as columns of a  $p \times s$  matrix  $A$ .

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

# CVA as a two-stage rotation I

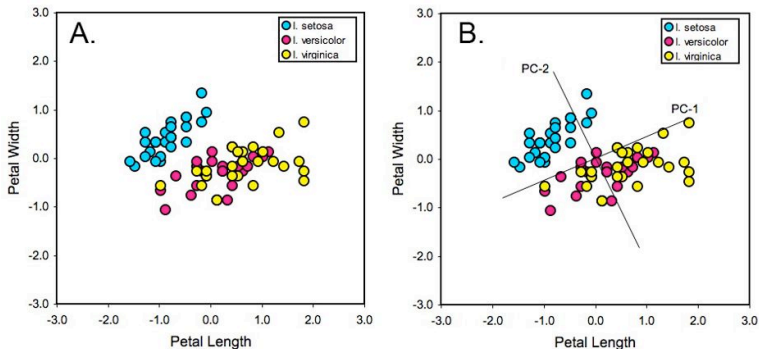


Figure 2. Stage 1 CVA implicit rotation. A. Scatterplot of first two *Iris* variables for example dataset. B. Orientation of the two pooled-sample principal components of the within-groups SSQCP matrix ( $W$ ).

## CVA as a two-stage rotation II

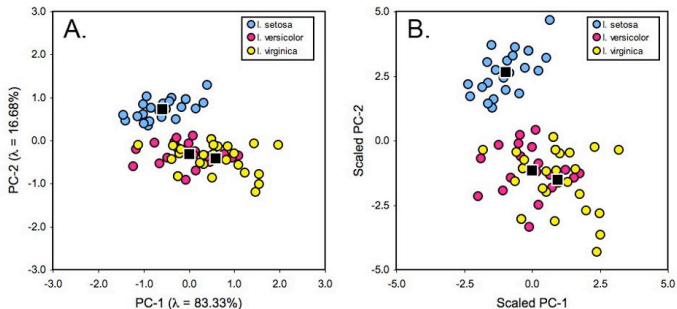


Figure 3. Intermediate scaling operation of a CVA. A. Scatterplot of *Iris* PC scores for the Stage 1 rotation (see Fig. 2). B. Result of scaling the two within-groups principal components by the square roots of their associated eigenvalues. Note difference in separation of the group centroids (black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

## CVA as a two-stage rotation III

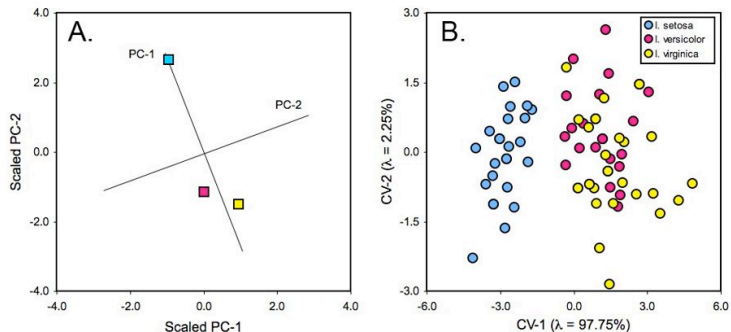


Figure 4. Stage 2 CVA implicit rotation. A. *Iris* group centroids plotted in the within-groups orthogonal-orthonormal space (see Fig. 3B) with between groups PC (= CVA) axes. B. Reduced *Iris* dataset plotted in the space defined by the CVA axes.

# CVA as a two-stage rotation IV

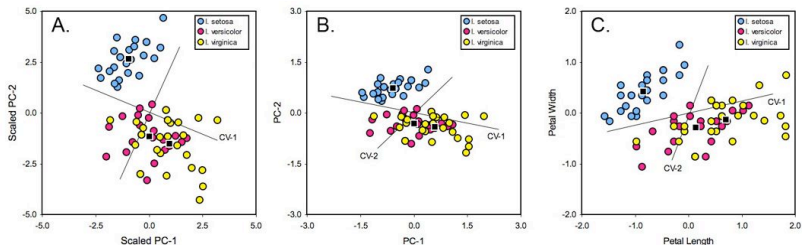


Figure 5. Back-calculation of final CVA axis orientation through the intermediate stages of the canonical rotations and scalings. A. Orientation of final CVA axes in the space of the scaled within-groups principal components (compare to Fig. 3A). B. Orientation of final CVA axes in the space of the raw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



# Similarities and Differences between CVA and PCA

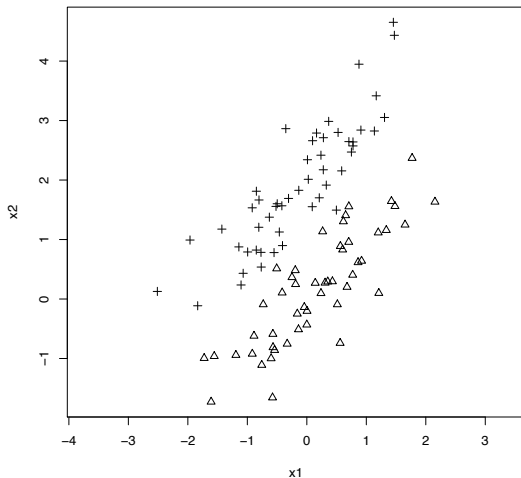
## PCA:

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

## CVA:

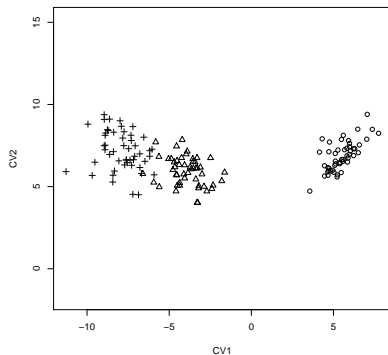
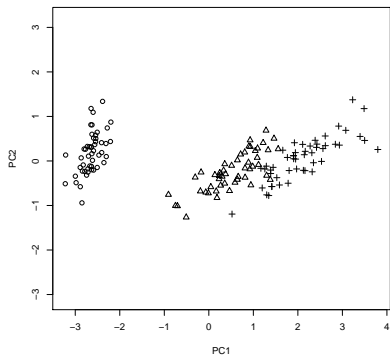
- Canonical variates are uncorrelated both *within* and *between* groups
- Canonical variates have equal variance *within* groups, but in decreasing order *between* groups
- non-orthogonal transformation, CV axes *not* at right angle to each other in the original frame of reference.

# PCA vs CVA: A Motivating Example



What is the direction of PC1? What is the direction of CV1?

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

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

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

Within groups the canonical variates are:

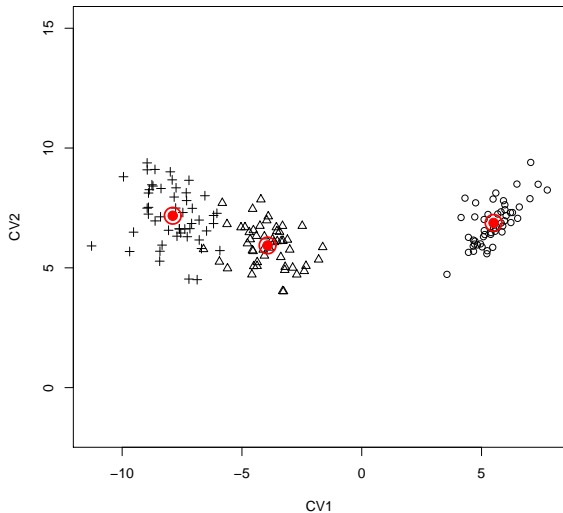
- uncorrelated
- have unit variance

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  $\mu_i$  is given by:

- hypersphere centered at  $\bar{y}_i$
- 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  $a_i$

- 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  $W$ ].