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
    - Interpretting CVA

Testing the Regression Effects

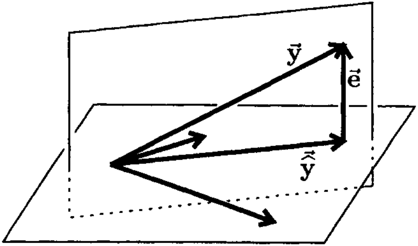


Figure : Geometry of multiple regression.

$$
\begin{aligned}
|\vec{y}|^2 &= SS_{\text{total}} \\
|\hat{\vec{y}}|^2 &= SS_{\text{regression}} = R^2 SS_{\text{total}} \\
|\vec{d}|^2 &= SS_{\text{resiudal}} = (1 - R^2) SS_{\text{total}}
\end{aligned}
$$

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

Geometry of the Population Regression Model

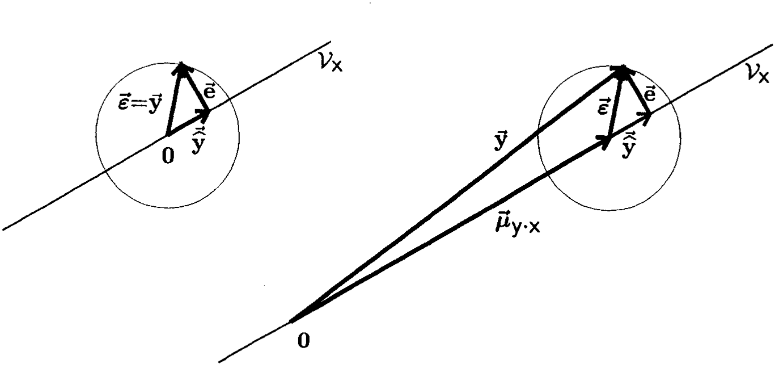
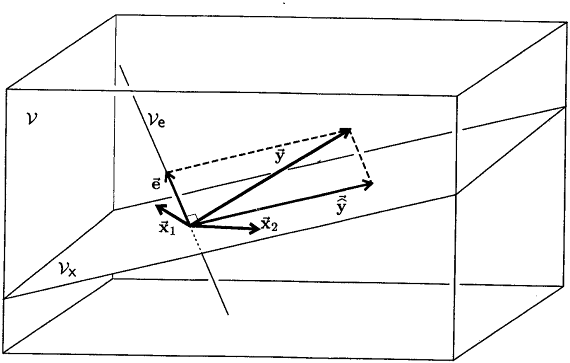


Figure : **Left:** null hypothesis of no regression effects is true; **Right:** null model is false.

Null Hypothesis: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

Dimensionality of Regression Subspaces



$$
\begin{aligned}
\dim(\mathcal{V}_{\text{total}}) &= N && \text{(total)}\\
\dim(\mathcal{V}_1) &= 1 && \text{(mean effect)}\\
\dim(\mathcal{V}_x) &= p && \text{(effect space)}\\
\dim(\mathcal{V}_e) &= N - p - 1 && \text{(error space)}
\end{aligned}
$$

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:

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

$$= \frac{(N-p-1)R^2}{p(1-R^2)}$$

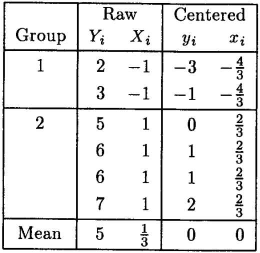
When null hypothesis is true, $F \approx 1$; when it is false, $F \gg 1$.

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.

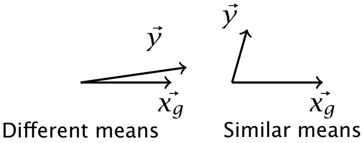
$$\mathsf{y} = \mathsf{X}_g\mathsf{b} + \mathsf{e}$$



	Raw		Centered	
Group	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:

$$\mathsf{y} = \mathsf{X}\mathsf{b} + \mathsf{e}$$

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

Estimate b as:

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

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

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

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

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

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

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, $\mathsf{a}'\mathsf{x}$ 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 a that maximizes the ratio of the squared group mean difference to within-group variance:

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

where

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

Geometry of the Two-Group Discriminant Function

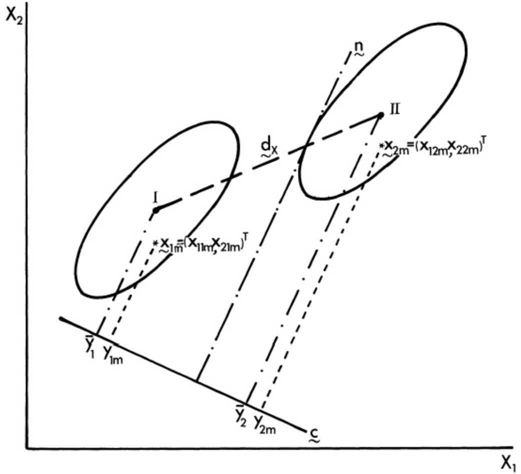


FIG. 4.—Representation of the discriminant function for two groups and two variables, showing the group means and associated 95% concentration ellipses. The vector **c** is the discriminant vector. The points $\bar{y}_1$ and $\bar{y}_2$ represent the discriminant means for the two groups.

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

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

Maximizing F gives:

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

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, $y = \mathrm{X}\mathrm{b} + e$
- 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:

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

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

Geometry of CVA

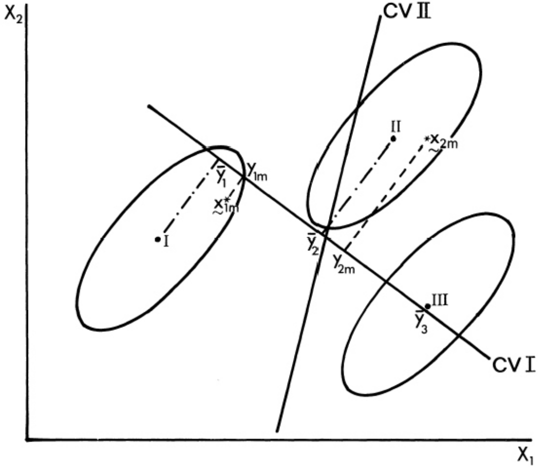


FIG. 2.—Representation of the canonical vectors for three groups and two variables. The group means (I, II and III) and 95% concentration ellipses are shown. The vectors CVI and CVII are the two canonical vectors. In the text, CVI = $\mathbf{e}$. The points y_{1m} and y_{2m} represent the canonical variate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

CVA Solution

Maximizing F leads to the following:

$$(\mathrm{B} - l\mathrm{W})\mathrm{a} = 0$$

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

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

CVA as a two-stage rotation I

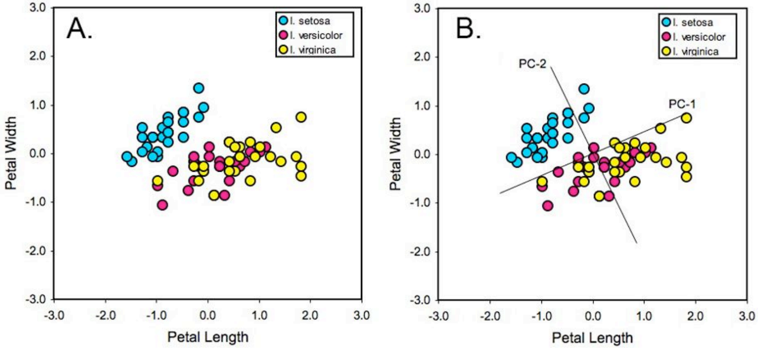


Figure 2. Stage 1 CVA implicit rotation. A. Scatterplot of first two *Iris* variables for example dataset. B. Orientation of the two pooled-sample principal components of the within-groups SSQCP matrix (*W*).

From MacLeod 2010, PalaeoMath newsletter

CVA as a two-stage rotation II

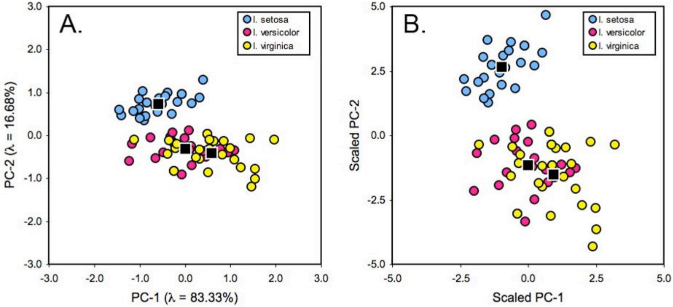


Figure 3. Intermediate scaling operation of a CVA. A. Scatterplot of *Iris* PC scores for the Stage 1 rotation (see Fig. 2). B. Result of scaling the two within-groups principal components by the square roots of their associated eigenvalues. Note difference in separation of the group centroids (black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

From MacLeod 2010, PalaeoMath newsletter

CVA as a two-stage rotation III

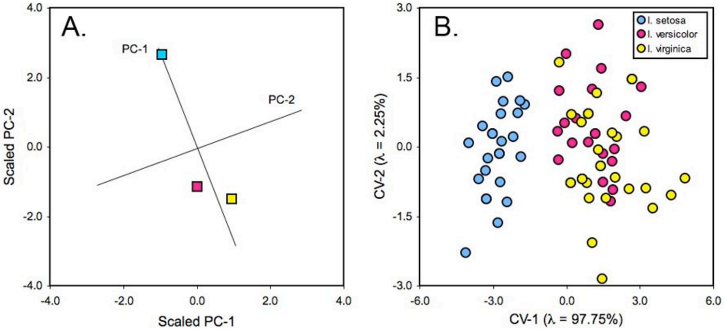


Figure 4. Stage 2 CVA implicit rotation. A. *Iris* group centroids plotted in the within-groups orthogonal-orthonormal space (see Fig. 3B) with between groups PC (= CVA) axes. B. Reduced *Iris* dataset plotted in the space defined by the CVA axes.

From MacLeod 2010, PalaeoMath newsletter

CVA as a two-stage rotation IV

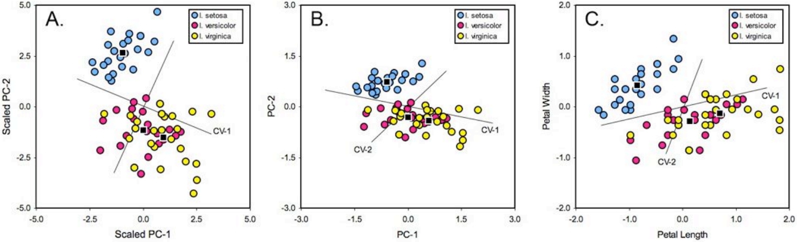


Figure 5. Back-calculation of final CVA axis orientation through the intermediate stages of the canonical rotations and scalings. A. Orientation of final CVA axes in the space of the scaled within-groups principal components (compare to Fig. 3A). B. Orientation of final CVA axes in the space of the raw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

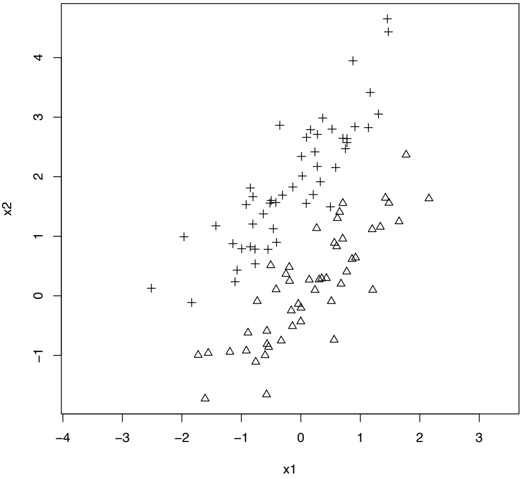
PCA:

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

CVA:

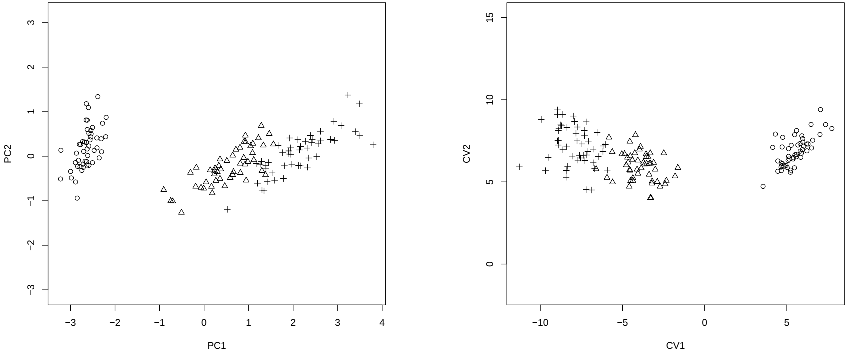
- Canonical variates are uncorrelated both *within* and *between* groups
- Canonical variates have equal variance *within* groups, but in decreasing order *between* groups
- non-orthogonal transformation, CV axes *not* at right angle to each other in the original frame of reference.

PCA vs CVA: A Motivating Example



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

PCA vs. CVA: Anderson's Iris Data



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

To test:

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

A couple of approaches:

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

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

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

Within groups the canonical variates are:

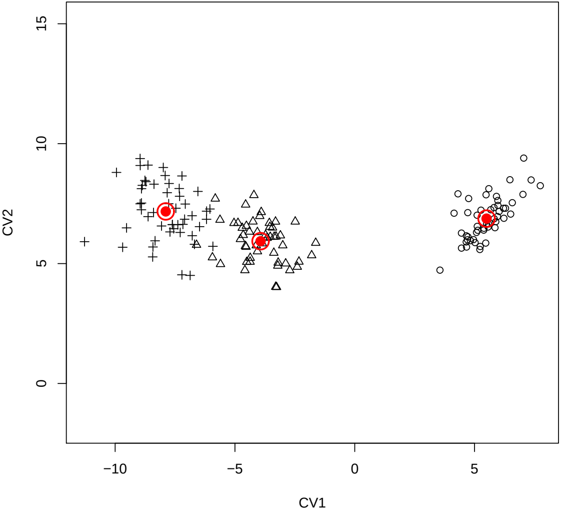
- uncorrelated
- have unit variance

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-α) percent confidence region for the true mean ν_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, $\mathrm{a}_i^* = (a_{i1}^*, \ldots, a_{ip}^*)$ where the a_{ij}^* are given by $a_{ij}^* = a_{ij}\sqrt{w_{jj}}$ [w_{jj} are the diagonal elements of W].