Chapter 17. Redundancy analysis (RDA) and canonical correspondence analysis (CCA)

1. Both PCA and CA apply to one data table. PCA mainly treats site×environmental data table, while CA treats site×species table. Different from PCA and CA, RDA and CCA are used to analyze two data tables.

2. Suppose there are two data tables: X and Y. In a typical ecological study, X represents environmental data and Y represents species data. In general, however, X and Y can represent any type of data. For example, a classic statistic example is the study of bone and skull measurements of White Leghorn fowl (Dunn, L.C. 1928. The effect of inbreeding on the bones of the fowl. Storrs Agricultural Experimental Station Bulletin 52:1-112).

$$\text{Head}\,(X): \begin{cases} X_1 = \text{Skull length} \\ X_2 = \text{skull breadth} \end{cases}$$

$$\text{Leg}\,(Y): \begin{cases} Y_1 = \text{femur length} \\ Y_2 = \text{tibia length} \end{cases}$$

3. RDA, or called Canonical Correlation Analysis (another CCA!) in statistics, seeks to identify and quantify the association between these two sets (head and leg) of variables. The basic steps of RDA include:

   (1) PCA – Find a linear combination (a principal axis) of the variables in the first data set (X_1, X_2, …) and a linear combination (a principal axis) of the variables in the second data set (Y_1, Y_2, …). This step is equivalent to PCA.

   (2) Correlation – The first principal axis pair of these two data sets must have the largest correlation.

   (3) The second principal pair must have the second largest correlation and this pair must uncorrelated with any other principal pairs.

(4) The pairs of the linear combinations are called the **canonical variables**, and their

correlations are called **canonical correlations**.

4. In matrix notation, we have

$U = a'X$ # principal axis: linear combination of $(X_1, X_2, \ldots)$

$V = b'Y$ # principal axis: linear combination of $(Y_1, Y_2, \ldots)$

where $a = \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_p \end{bmatrix}$, $X = \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_p \end{bmatrix}$, and $b = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_q \end{bmatrix}$, $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_q \end{bmatrix}$

What RDA seeks is the coefficient vectors a and b (PCA loadings) that make the

following correlation as large as possible:

$$Cor(U,V) = \frac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}a}\sqrt{b'\Sigma_{22}b}}$$

Further condition is that canonical pairs must be uncorrelated with each other, i.e., U_1,

$U_2, \ldots, U_p$ must be uncorrelated. Similarly, $V_1, V_2, \ldots, V_q$ must be uncorrelated with

each other.

5. R implementation using **vegan**:   **> rda**

6. CCA (Canonical correspondence analysis) in principle also seeks for the maximum

correlation between respective principal axes of two data sets. Similar to CA, CCA is

also based on χ^2 distance.

7. R implementation using **vegan**: **> cca**.

8. It is useful to point out that **rda** and **cca** in **vegan** allows a third data table for

conditioning. This can be useful if there is a need to control or remove the effect of

the third data table.

9. Variance partition using **vegan: >varpart.**
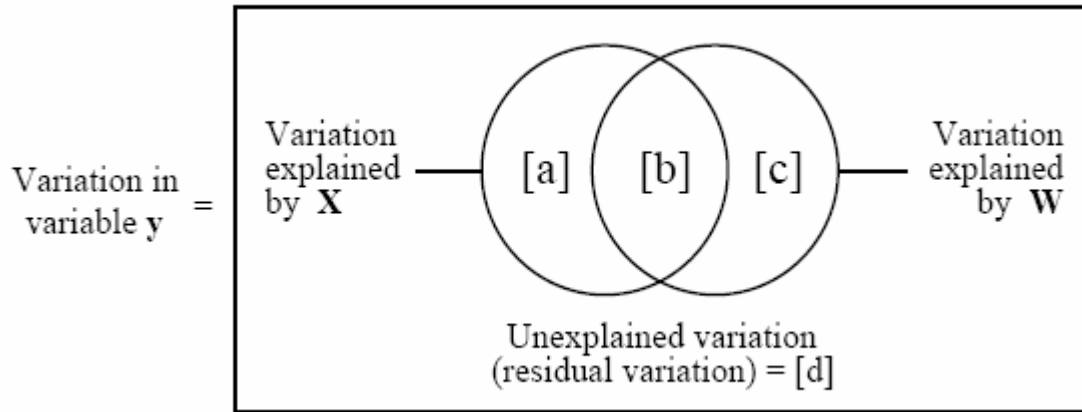
10. Use Guotianshan plot as an example to partition variation in species data into topographical and space components.

>toto=varpart(guo.sp10,guo.env2,guo.pcnm.179)
> toto

Partition of variation in RDA

Call:
varpart(Y = Guo.sp10, X = Guo.env2, Guo.pcnm.179)

Explanatory tables:
X1: Guo.env2
X2: Guo.pcnm.179

No. of explanatory tables: 2
Total variation (SS): 2143507
      Variance: 3578.5

No. of observations: 600

Partition table:

	Df	R.squared	Adj.R.squared	Testable
[a+b] = X1	7	0.26137	0.25264	TRUE
[b+c] = X2	179	0.73640	0.62405	TRUE
[a+b+c] = X1+X2	186	0.75360	0.64263	TRUE
Individual fractions				
[a] = X1|X2	7		0.01858	TRUE
[b]	0		0.23406	FALSE
[c] = X2|X1	179		0.38999	TRUE
[d] = Residuals			0.35737	FALSE

Use function 'rda' to test significance of fractions of interest

===========