Worksheet 10 - Multivariate analysis

Multivariate analysis

- Quinn & Keough (2002) - Chpt 15, 17-18

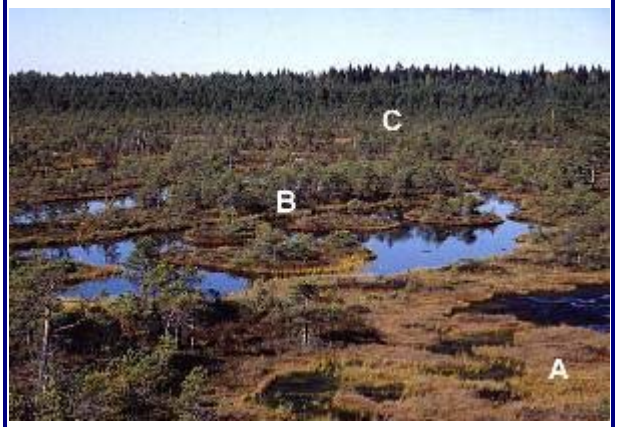
Question 1 - Principal Components Analysis (PCA)

Gittens(1985) measured the abundance of 8 species of plants from 45 sites within 3 habitat types. Essentially, the plant ecologist wanted to be able to compare the sites according to their plant communities.

Format of veg.csv data file

SITE	HABITAT	SP1	...	SP8
1	A	4	..	68
2	B	92	..	4
3	A	9	..	68
4	A	52	..	24
5	C	99	..	0
6	A	12	..	68
7	C	72	..	8
8	C	80	..	8
9	B	80	..	0
10	C	92	..	0

SITE	A number or name given to each quadrat (site)
HABITAT	A letter or name given to each habitat type
SP1, SP2, .., SP8	Number of individuals of each plant species found in each quadrat



Open the veg data file (HINT). Note the format of the file, with variables in columns and samples/sites in rows.

Q1-1. Treating the data set as univariate (only a single response variable), examine the patterns between habitats based on the following single species;

a. SP1 (HINT). What does this plot illustrate about the relationships between habitats based on this plant species?

b. SP2 (HINT). What does this plot illustrate about the relationships between habitats based on this plant species?

c. SP5 (HINT). What does this plot illustrate about the relationships between habitats based on this plant species?

d. SP8 (HINT). What does this plot illustrate about the relationships between habitats based on this plant species?

The ecologist was not interested in teasing out the patterns based on each individual species in isolation. The ecologist wanted to see patterns between the plant communities, rather than individual species. Hence a multivariate approach was taken. You may have noticed that the patterns between sites (and habitats) based on SP1 and SP8 were very similar. The abundances of SP1 and SP2 appear to be correlated to one another. It is not surprising that different species might be correlated, as they are likely to respond similarly to similar conditions.
If two or more species reveal exactly the same patterns (hypothetically), we could easily combine them into a single group that characterises the sites. Species are never likely to be exactly correlated, however we can still generate new groups that are the combinations of multiple species abundances. If we were to attempt to combine three species, two of which were highly correlated to one another and the other not correlated to either, then the two correlated species will contribute a lot to the new group and the other variable will contribute only little.
In the example, lets say we wanted to reduce the 8 species variables down to just 2 groups. Based on how much each species is correlated to each other species, each species will contribute something to each of the two new groups. So each new group is a linear combination of original species variables. This sort of data combining (or reduction) can be done in a number of ways, however for it to work meaningfully, there must be a reasonable degree of correlation between the species.

Q1-2. Examine the degree of correlation between each of the species (HINT). Does it appear that some species are correlated to others, which have the greatest degree of correlation?

Q1-3. SP1 and SP8 appeared to be highly negatively correlated. To examine this correlation, create a scatterplot of SP1 against SP8 (HINT) and fit a smooth line through these data(HINT). If we were purely trying to combine SP1 and SP2 into a new group, the position of each site (point) along this effectively becomes the sites new value in the new group.

Q1-4. Use principal components analysis (PCA, HINT) to generate new groups (components) and explore the trends in plant communities amungst sites (and habitats)

a. Examine the Eigenvalues for each new component (group) (HINT). The sum of these values should add up to the number of original variables (species). If there were absolutely no correlations between the species, then you would exect each new component to represent a single original species and it would have an eigenvalue of 1. The more correlated the species were, the more they will group together into the first few newly generated compenents (groups) and thus the higher the eigenvalues of these earlier groups. What do the eigenvalues indicate in this case?

b. Calculate the percentage of total variation is explained by each of the new principal components (HINT). How much of the total original variation is explained by principal component 1 (as a percentage)?

c. Calculate the cumulative sum of these percentages (HINT). How much of the total variation is explained by the first three principal components (as a percentage)?

d. Using the Eigenvalues and a screeplot, determine how many principal compenents are necessary to represent the original variables (species) (HINT). How many principal components are necessary?

Q1-5. Generate an ordination (scatterplot of principal components) with principal component 1 on the x-axis and principal component 2 on the y-axis, but instead of using points (HINT), use the original habitat types as point labels (HINT). What patterns are revealed between habitats?

Q1-6. Examine the component loadings to determine the contribution of each of the original 8 species variables to the new principal components (groups) (HINT) and visualise this with a biplot (HINT).

Question 2 - Principal Components Analysis (PCA)

Peet & Loucks (1977) examined the abundances of 8 species of trees (Bur oak, Black oak, White oak, Red oak, American elm, Basswood, Ironwood, Sugar maple) at 10 forest sites in southern Wisconsin, USA. The data (given below) are the mean measurements of canopy cover for eight species of north American trees in 10 samples (quadrats). For this question we will explore the relationships between the quadrats, in terms of tree species abundances using PCA. That is, which quadrats are most similar/dissimilar to one another.

Format of wisc.csv data file

QUAD.	BUROAK	BLACKOAK	WHITEOAK	REDOAK	ELM	BASSWOOD	IRONWOOD	MAPLE
1	9	8	5	3	2	0	0	0
2	8	9	4	4	2	0	0	0
3	3	8	9	0	4	0	0	0
4	5	7	9	6	5	0	0	0
5	6	0	7	9	6	2	0	0

6	0	0	7	8	5	7	6	5
7	5	0	4	7	5	6	7	4
8	0	0	6	6	0	6	4	8
9	0	0	0	4	2	7	6	8
10	0	0	2	3	5	6	5	9

QUADRAT	A number or name given to each quadrat
BUROAK, BLACKOAK,....	Number of individuals of each tree species found in each quadrat

Open the wisc data file (HINT). Note the format of the file, with variables in columns and samples/sites in rows.

Q2-1. Exploring the data set, which quadrats appear to be most similar to one another with respect to tree communities?

Q2-2. Perform a PCA (HINT)

a. Examine the latent roots (Eigenvalues, HINT) and determine how much of the total variation (HINT) is explained by each new component

b. Construct a screen plot (HINT) and determine how many components are necessary to explain most of the variation. How many components would you retain in the data reduction procedure

c. Construct a PCA scatterplot of component 1 on the x-axis and component 2 on the y-axis (HINT) and use the quadrat names as labels(HINT). What does this plot illustrate about the relationships between sites?

d. Either examine the component loadings (HINT) or construct a PCA biplot of component 1 on the x-axis and component 2 on the y-axis (HINT) to indicate the contribution of each of the original variables to each of the new components.

Q2-3. Is there any evidence of a gradient in the landscape, reflected in the arrangement of points in the ordination?

Q2-4. How might you interpret independent data for each point, such as groundwater or soil characteristics, using ordination?

Question 3 - Dissimilarity

The following data are the abundances of 3 species of gastropods in 5 quadrats (ranging from high shore marsh, Quadrat 1, to low shore marsh, Quadrat 5) in a saltmarsh.

Format of gastropod.csv data file

	Salinator	Ophicardelus	Marinula
Q1	4	0	1
Q2	9	3	0
Q3	9	4	1
Q4	6	2	0
Q5	0	1	1

Salinator	Number of Salinator gastropods - variable
Ophicardelus	Number of Ophicardelus gastropods - variable
Marinula	Number of Marinula gastropods - variable
Q1-Q5	Quadrats - these are the samples (sites)

Q3-1. By hand, calculate the **Bray-Curtis(Czekanowski) dissimilarity coefficient** and **Euclidean distance** between all pairs of quadrats. Fill in the matrix below. Note that the lower left-hand section of a dissimilarity, correlation, distance, etc matrix is mirrored in the upper right-hand section and thus can be represented by a triangular matrix. To save space and assist comparisons, fill the lower left section with Bray-Curtis dissimilarity values and the upper right with Euclidean distances.

	Q1	Q2	Q3	Q4	Q5
Q1	0.00				
Q2		0.00			
Q3			0.00		
Q4				0.00	
Q5					0.00

a. Do the two measures correspond? In what ways are they similar/different?

Open the gastropod data file (HINT). Note the format of the file, with variables in columns and samples/sites in rows.

Q3-2. Now lets use R to calculate separately the **Bray-Curtis(Czekanowski) dissimilarity coefficients and Euclidean distances** (HINT), (HINT).

Question 4 - Mantel tests

Vare *et al.* (1995) measured the cover abundance of 44 plants from 24 sites so as to explore patterns in vegetation communities between these sites. They also measured a number of environmental variables (mainly concentration or various soil chemicals) from each site so as to also be able to characterise sites according to soil characteristics. Their primary interest was to investigate whether there was a correlation between the plant communities and the soil characteristics.

Format of vareveg and vareenv data files

vareveg

SITE	C.vul	...	Cla.phy
18	0.55	...	0.00
15	0.67	...	0.00
24	0.10	...	0.00
27	0.00	...	0.00
23	0.00	...	0.00

SITE	Name or number of the 14 sites
C.val, ...Cl.phy	Cover abundance of 44 species of plants

vareenv

SITE	N	...	pH
18	19.8	...	2.7
15	13.4	...	2.8
24	20.2	...	3.0
27	20.6	...	2.8
23	23.8	...	2.7

SITE	Name or number of the 14 sites
N, ..., pH	Measurements for 14 environmental soil chemicals from each of the sites

Open the vareveg (HINT) and vareenv (HINT) data files.

Q4-1. Briefly examine the variables and determine whether or not standardisation is required and what sort of dissimilarity matrix is appropriate for the vegetation data and the environmental soil chemistry data

a. Vegetation data HINT)

b. Soil chemistry data HINT)

Q4-2. Perform any necessary standardisations and generate the appropriate distance matricies for the vegetation data and the soil chemistry data

a. Vegetation data (HINT)

b. Soil chemistry data (HINT)

Q4-3. Perform a Mantel test to calculate the correlation between the two matricies (vegetation and soil) (HINT).

a. What was the R value?

b. Was this significant?

Q4-4. Generate a correlogram (mutivariate correlation plot) (HINT).

Question 5 - Analysis of Similarities (ANOSIM)

Jongman *et al.* (1987) presented a data set from a study in which the cover abundance of 30 plant species were measured on 20 rangeland dune sites. They also indicated what the form of management each site experienced (either biological farming, hobby farming, nature conservation management or standard farming. The major intension of the study was to determine whether the vegetation communities differed between the alternative management practices.

Format of dune.csv data file

MANAGEMENT	Belper	..	Brohor
BF	3	..	4
SF	0	..	0
SF	2	..	3
SF	0	..	0
HF	0	..	0
..	..	..	..

Open the dune data file (HINT).

Q5-1. Generate a bray-curtis dissimilarity matrix for the 30 plant species (HINT).

Q5-2. Perform an ANOSIM to calculate whether the ratio of average rank dissimilarities between management strategies to average ranks within management strategies is significantly greater than you would expect from a random set of configurations (HINT).

a. What was the R value (HINT)? _____ ☐

b. Was this significant? _____ ☐

c. Examine the differences betwen groups. What patterns are revealed?

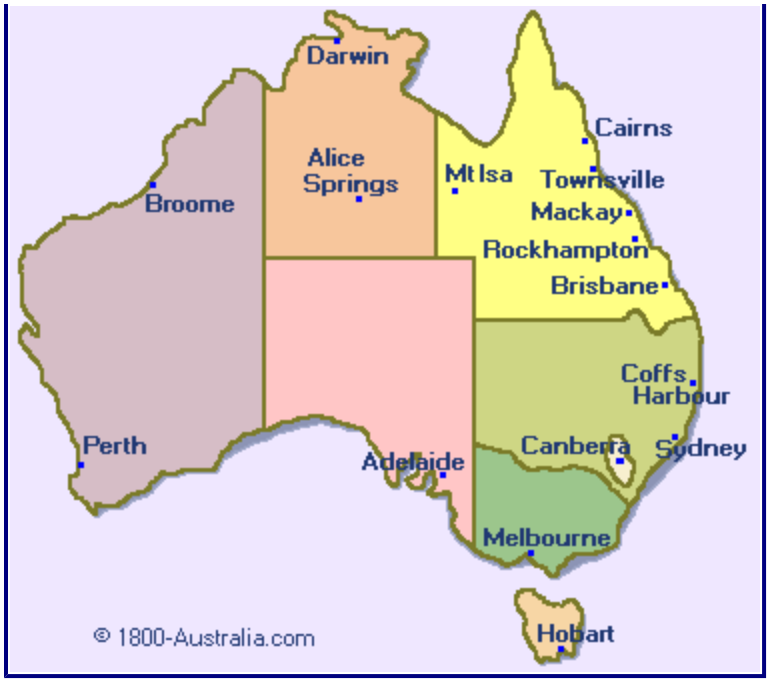
Q5-3. Summarise these findings with a plot (HINT).

Question 6 - Multidimensional scaling

The following example is designed to help you appreciate the link between distance measures and ordination space (MDS). The data set consists of distances (km) between major Australia cities (as the crow flies), and is in the form of a triangular matrix.

Format of austcites.csv data file

	Canberra	Sydney	Melbourne	..
Canberra	0	NA	NA	..
Sydney	246	0	NA	..
Melbourne	467	713	0	..
Adelaide	958	1160	653	..
Perth	3090	3290	2720	..
..	..	..	..	..



Open the austcities data file (HINT). Note the format of the file, it is a triangular distance matrix.

While the file is a distance matrix, at this stage R is unaware of it, we must manually make it aware (a round about way of saying that we must type a command to force R to treat the data set as a distance matrix. Convert the data frame into a distance matrix (HINT)

We are now ready to perform the MDS for the purpose of examining the ordination plot.

Q6-1. Perform an MDS with 2 dimensions on the city distances matrix (HINT).

a. What was the final stress value (as a percentage)? ☐

b. What does this stress value suggest about the success of the MDS?

c. Generate a Shepard diagram (HINT). The Shepard diagram (plot) represents the relationship between the original distances (y-axis) and the new MDS distances (x-axis). Does this and the stress value indicate that the patterns present in the original distance matrix (crow flies distances between cities) are adequately reproduced from the 2 new dimensions? ☐

d. Generate an ordination plot (HINT or HINT). The final ordination plot summarizes the relationship between the cities. Does this ordination plot approximate the true geographical arrangement of the cities? ☐

e. In this case, what might the two new MDS dimensions (variables) represent? (hint think of the ordination plot as a map)

Question 7 - Multidimensional scaling

MacNally (1989) studied geographic variation in forest bird communities. His data set consists of the maximum abundance for 102 bird species from 37 sites that where further classified into five different forest types (Gippsland manna gum, montane forest, woodland, box-ironbark and river redgum and mixed forest). He was primarily interested in determining whether the bird assemblages differed between forest types.

Format of macnally.csv data file

SITE	HABITAT	V1GST	..
Reedy Lake	Mixed	3.4	..
Pearcedale	Gippland Manna Gum	3.4	..
Varneet	Gippland Manna Gum	8.4	..
Cranbourne	Gippland Manna Gum	3.0	..
Lysterfield	Mixed	5.6	..
..	..	..	..

Open the macnally data file HINT.

Q7-1. Calculate the **Bray-Curtis(Czekanowski) dissimilarity coefficients** (HINT) amongst the sites using all 102 bird species abundances.

Q7-2. Perform an MDS on the dissimilarity matrix in which a number of random starts are attempted and the scaling in the result is standardised (HINT).

a. What was the final stress value (as a percentage)?

b. What does this stress value suggest about the success of the MDS?

c. The Sheppard diagram (plot) represents (HINT) the relationship between the original distances (y-axis) and the new MDS distances (x-axis). How would you describe the shape of this curve, and base on this is metric or non-metric MDS more appropriate?

d. The final ordination plot (HINT) summarizes the relationship between the sites. What would you conclude from this?

Question 8 - Cluster analysis

Peet & Loucks (1977) examined the abundances of 8 species of trees (Bur oak, Black oak, White oak, Red oak, American elm, Basswood, Ironwood, Sugar maple) at 10 forest sites in southern Wisconsin, USA. The data (given below) are the mean measurements of canopy cover for eight species of north American trees in 10 samples (quadrats). For this question we will explore the relationships between the quadrats via cluster analysis.

Format of wisc.csv data file

QUAD.	BUROAK	BLACKOAK	WHITEOAK	REDOAK	ELM	BASSWOOD	IRONWOOD	MAPLE
1	9	8	5	3	2	0	0	0
2	8	9	4	4	2	0	0	0
3	3	8	9	0	4	0	0	0
4	5	7	9	6	5	0	0	0
5	6	0	7	9	6	2	0	0
6	0	0	7	8	5	7	6	5
7	5	0	4	7	5	6	7	4
8	0	0	6	6	0	6	4	8
9	0	0	0	4	2	7	6	8
10	0	0	2	3	5	6	5	9

QUADRAT	A number or name given to each quadrat
BUROAK, BLACKOAK,....	Number of individuals of each tree species found in each quadrat

Open the wisc data file (HINT). Note the format of the file, with variables in columns and samples/sites in rows.

Q8-1. Calculate the **Bray-Curtis(Czekanowski) dissimilarity coefficients** (HINT) amongst the sites using all 8 species abundances.

Q8-2. Perform hierachical clustering using the following linkages;

a. Single linkage (nearest neighbour) (HINT)

b. Average linkage (UPGMA) (HINT)

Q8-3. For each of the above clusters, calculate the cophenetic correlation coefficient, and comment on which clustering linkage provides the better fit;

a. Single linkage (nearest neighbour) (HINT)

---

b. Average linkage (UPGMA) (HINT)

Welcome to the end of Worksheet 10!