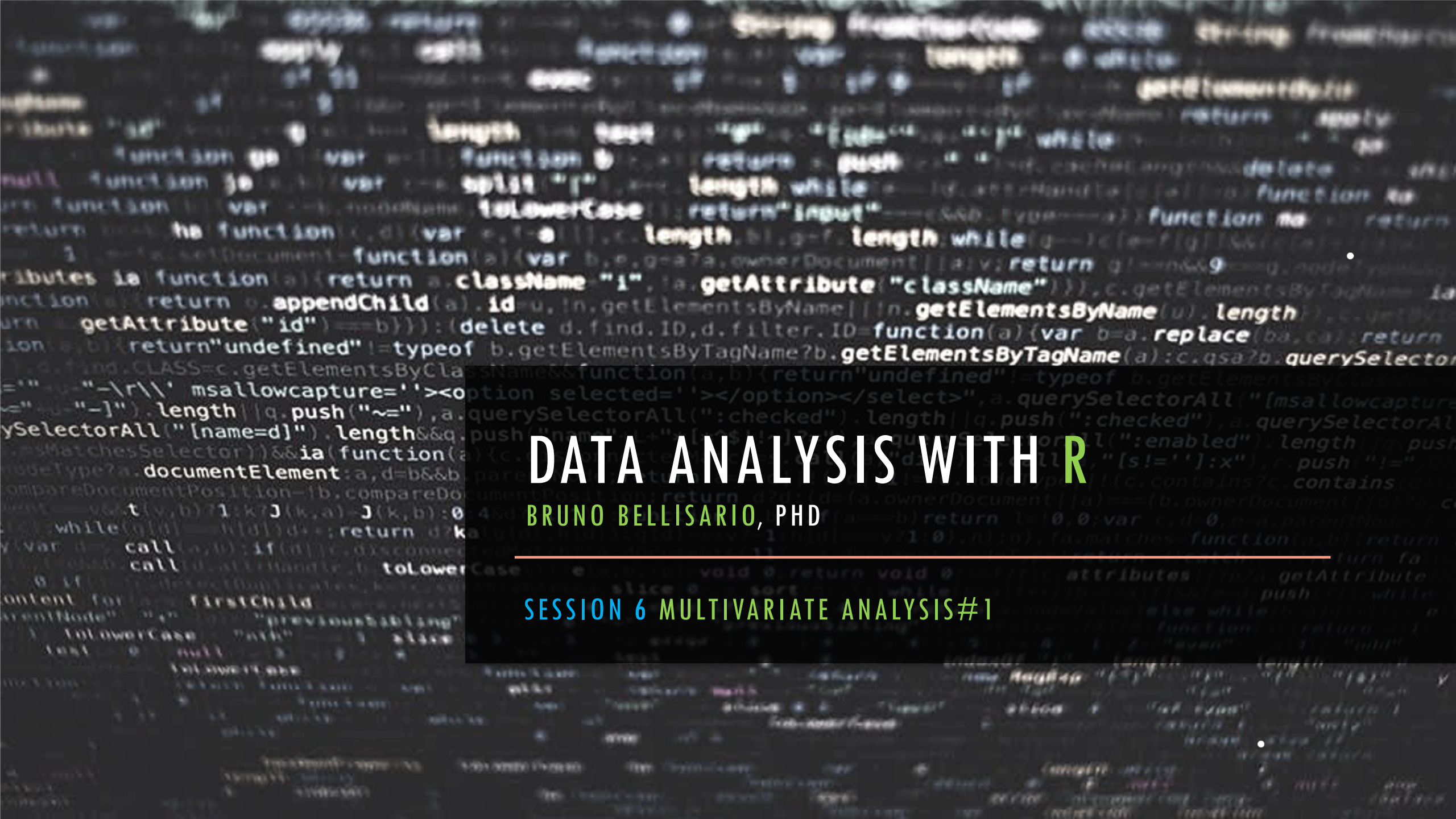

DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

SESSION 6 MULTIVARIATE ANALYSIS#1

MULTIVARIATE ANALYSIS

- Many of the statistical analyses encountered to date consist of a single response variable and one or more explanatory variables.
- Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable.
- There isn't an entirely clear "canon" of what is a multivariate technique and what isn't. However, we are going to consider the simultaneous analysis of a number of related variables.

THE NATURE OF MULTIVARIATE DATA

- Much of the data examined is observational rather than collected from designed experiments.
- Multivariate data consist of individual measurements that are acquired as a function of more than two variables, for example, kinetics measured at many wavelengths and as a function of temperature, or as a function of pH, or as a function of initial concentrations, and so forth, of the reacting solutions.

From: Comprehensive Chemometrics, 2009

MULTIVARIATE ANALYSIS

The multivariate problem can be approached in (at least) two ways:

→The first group of problems relates to classification, where attention is focused on individuals who are more alike.

→The other group of problems concerns inter-relationships between variables.

COMMON GOALS

- Describe the p-dimensional distribution
- Reduce the number of variables without losing significant information
- Investigate dependence between variables
- Statistical inference
- Clustering and Classification

CLUSTER ANALYSIS

Cluster analysis describes a class of techniques tasked with placing objects in groups, called clusters.

Dissimilarities between objects within these groups should be smaller than those between groups. The definition of a cluster varies, and different cluster analysis techniques may approach the problem very differently.

Two widely-used approaches: hierarchical vs non-hierarchical cluster analysis

DISSIMILARITY MATRIX

The dissimilarity matrix (also called distance matrix) describes pairwise distinction between M objects. It is a square symmetrical M x M matrix with the $(ij)^{th}$ element equal to the value of a chosen measure of distinction between the $(i)^{th}$ and the $(j)^{th}$ object.

DISSIMILARITY MATRIX

(Dis)similarity, distance, and dependence measures are powerful tools in determining ecological association and resemblance.

→Choosing an appropriate measure is essential as it will strongly affect how your data is treated during analysis and what kind of interpretations are meaningful.

→Non-metric dimensional scaling, principal coordinate analysis, and cluster analysis are examples of analyses that are strongly influenced by the choice of (dis)similarity measure used.

DISSIMILARITY MATRIX

Key terminology

→Q mode analysis

Focuses on relationships between objects.

→R mode analysis

Focuses on relationships between variables.

→Euclidean space

Sometimes called metric space, Euclidean spaces have axes which are quantitative and may be used for standard measurement. In other words, distances between points in Euclidean space are literal distances and are directly interpretable.

DISSIMILARITY MATRIX

Q mode dissimilarity and distance measures

Metric coefficients

→The distance between identical objects is 0

→If objects are not identical, a metric coefficient will have a positive value

→Symmetry

→Conformance to the triangle inequality.

DISSIMILARITY MATRIX

Q mode dissimilarity and distance measures

Metric coefficients

→Euclidean distance - A simple, symmetrical metric using the Pythagorean formula. Sensitive to double zeros (missing information in pairwise objects). Unsuitable for most of ecological applications without ecologically-motivated data transformations.

→Hellinger distance - This distance measure performs well in linear ordination. Variables with few non-zero counts are given lower weights (robust enough to double zeros).

→Jaccard distance - The one complement of the Jaccard similarity, which measures the similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. Given a "sites x species" matrix, the Jaccard coefficient can be used to express species/OTU turnover.

DISSIMILARITY MATRIX

Q mode dissimilarity and distance measures

(Semi)Metric coefficients

→Semimetric measures do not always satisfy the triangle inequality and hence cannot be fully relied upon to represent dissimilarities in a Euclidean space without appropriate transformation.

→They often do behave metrically and can be used in principal coordinates analysis (following an adjustment for negative eigenvalues if necessary) and non-metric dimensional scaling.

DISSIMILARITY MATRIX

Q mode dissimilarity and distance measures

(Semi)Metric coefficients

→Bray-Curtis dissimilarity - This is an asymmetrical measure often used for raw count data and a very popular measure of dissimilarity in ecology. This measure treats differences between high and low variable values equally.

→Sørensen dissimilarity - The one complement of the Sørensen similarity coefficient, which is similar to the Jaccard coefficient, although gives double weight to non-zero agreements.

HIERARCHICAL CLUSTER ANALYSIS

Hierarchical cluster analysis is an algorithmic approach to find discrete groups with varying degrees of (dis)similarity in a data set represented by a (dis)similarity matrix.

These groups are hierarchically organised as the algorithms proceed and may be presented as a dendrogram.

Many of these algorithms are greedy (i.e. the optimal local solution is always taken in the hope of finding an optimal global solution) and heuristic, requiring the results of cluster analysis to be evaluated for stability by, for example, bootstrapping procedures.

HIERARCHICAL CLUSTER ANALYSIS

→Hierarchical clustering may be useful in discretising largely continuous phenomena to aid structure detection and hypothesis generation.

→If data were collected along a gradient, for example, cluster analysis may help to identify (relatively) distinct regions therein which may correspond to an ecologically meaningful grouping.

→As always, one must consider carefully whether clustering is suitable and meaningful for the task at hand. If a very smooth response gradient (i.e. very even changes in (dis)similarity between objects) is being scrutinised, ordination procedures may be more appropriate.

HIERARCHICAL CLUSTER ANALYSIS

Clustering methods

Agglomerative clustering is a widespread approach to cluster analysis. Agglomerative algorithms successively merge individual entities and clusters that have the highest similarity values. Agglomerative algorithms end when all the individual entities and clusters have been merged into a single cluster.

Single-linkage - When a new cluster is formed, the (dis)similarities between it and the other clusters and/or individual entities are computed based on the (dis)similarity between the nearest two members of each group

Complete-linkage - When a new cluster is formed, the (dis)similarities between it and the other clusters and/or individual entities present are computed based on the (dis)similarity between the farthest two members of each group

Average-linkage - When a new cluster is formed, the (dis)similarities between it and the other clusters and/or individual entities present are computed based on the average (dis)similarity between all members in each group

HIERARCHICAL CLUSTER ANALYSIS

→Hierarchical clustering methods can become computationally intensive for large data sets and bootstrapping results may be prohibitively expensive.

→Many hierarchical clustering approaches are sensitive to outliers.

→Being greedy, heuristic algorithms, mis-groupings that occur at early stages are not corrected at later stages. Many implementations support bootstrapping or other resampling techniques to assess the stability of a clustering solution and suggest a consensus grouping.

→Be aware of caveats and known issues associated with the wide range of clustering methods available. Some may have profound consequences on interpretability.

NON-HIERARCHICAL CLUSTER ANALYSIS

Non-hierarchical cluster analysis aims to find a grouping of objects which maximises or minimises some evaluating criterion.

Many of these algorithms will iteratively assign objects to different groups while searching for some optimal value of the criterion.

NON-HIERARCHICAL CLUSTER ANALYSIS

K-means

→K-means clustering aims to assign objects to a user-defined number of clusters (k) in such a way that maximises the separation of those clusters while minimising intra-cluster distances relative to the cluster's mean or centroid

→The algorithm typically defaults to Euclidean distances, however, alternate criteria, such as different distance or dissimilarity measures, can be accepted by many implementations.

→The user is usually expected to set 3 parameters: 1) the number of clusters expected (k), 2) the initialisation method, and 3) the distance metric to be used.

NON-HIERARCHICAL CLUSTER ANALYSIS

- The k-means algorithm can become 'stuck' in local optima. Repeating the clustering algorithm and adding noise to the data can help evaluate the robustness of the solution.
- The k-means algorithm will favour higher values of k. This is not necessarily desirable and users should consider carefully which values of k are sensible for their data set.
- The full k-means algorithm is computationally intensive.
- Solutions may vary with different distance measures.
- The algorithm is insensitive to other features of an object population. Thus, verify that the populations of interest can be classified well by their distance to their multivariate mean.
- Objects must be assigned to one cluster, problematic in cases where an object is equidistant from two or more centroids. It is also problematic if there exist outliers.
- If populations of objects overlap, the k-means algorithm may provide biased estimates of their centroids, being 'pulled' towards regions where no (or less) overlap occurs.
- Standardising variables may change the solution.
- Comparing clustering solutions with different values of k may be problematic, as many solutions for each value would have to be examined to prevent evaluating solutions that represent local optima.
- When using Euclidean distances, variables which are correlated will naturally influence object positioning in similar ways.

HIERARCHICAL VS. NON-HIERARCHICAL CLUSTERING

Hierarchical Clustering:	Non Hierarchical Clustering:
Involves creating clusters in a predefined order from top to bottom	Involves formation of new clusters by merging or splitting the clusters instead of following a hierarchical order.
It is considered less reliable than Non Hierarchical Clustering.	It is comparatively more reliable than Hierarchical Clustering.
It is considered slower than Non Hierarchical Clustering.	It is comparatavely more faster than Hierarchical Clustering.
It is very problematic to apply this technique when we have data with high level of error.	It can work better then Hierarchical clustering even when error is there.
It is comparatively easier to read and understand.	The clusters are difficult to read and understand as compared to Hierarchical clustering.
It is relatively unstable than Non Hierarchical clustering.	It is a relatively stable technique.