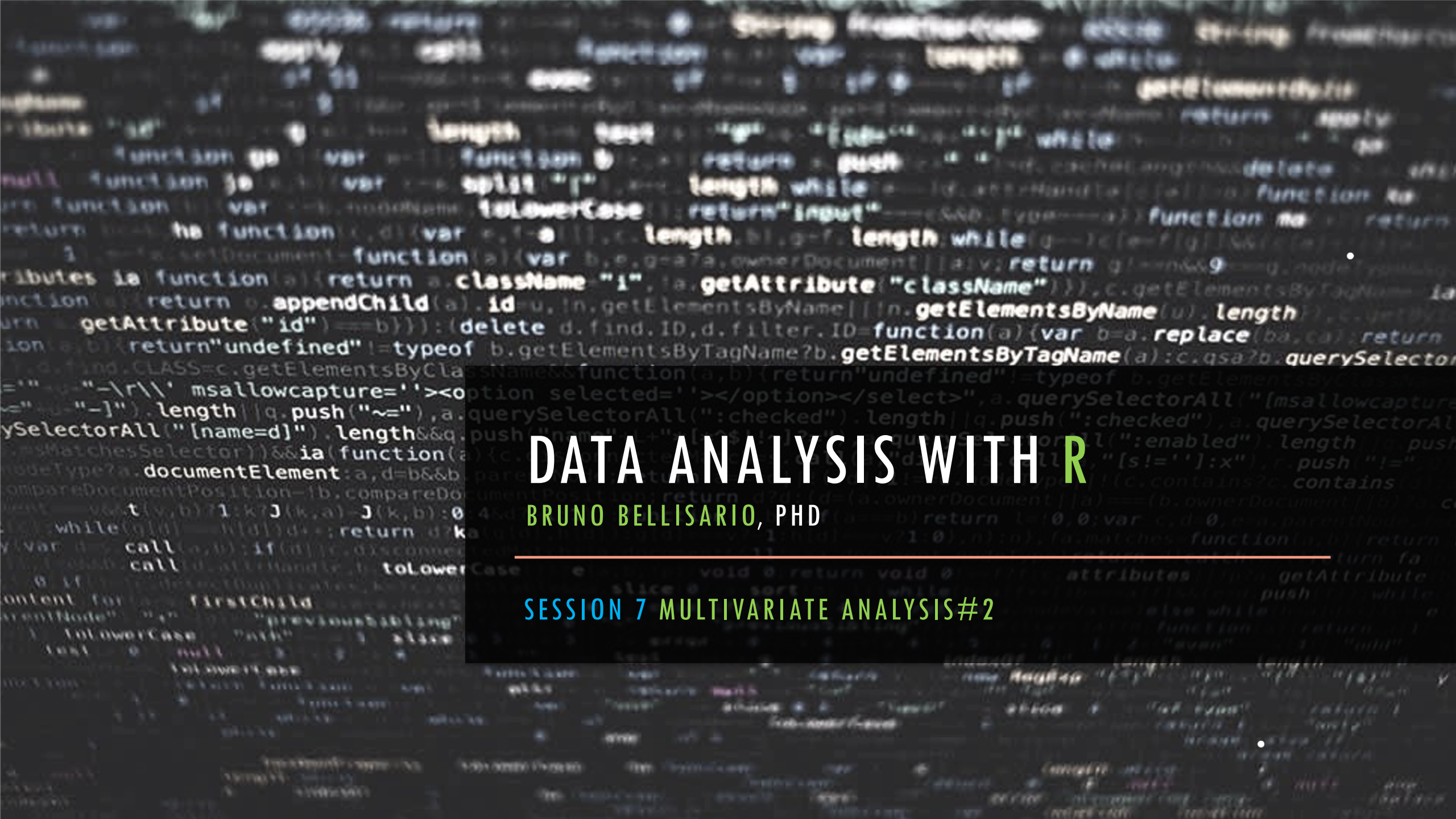


DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

SESSION 7 MULTIVARIATE ANALYSIS#2

UNCONSTRAINED ORDINATION

Unconstrained ordinations are a type of 'indirect gradient analysis' - any interpretation of potential effects of other factors that generated the patterns can only be made indirectly because those factors were not explicitly included in the analysis.

→Non metric Multi Dimensional Scaling (nMDS) - Distance based

→Principal Component Analysis (PCA)– Linear based

UNCONSTRAINED ORDINATION

NMDS

- The main aim is to locate samples in low dimensional ordination space (two or three axes) so as the Euclidean distances between these samples correspond to the dissimilarities represented by the original dissimilarity index.
- The method is non-metric, because it does not use the raw dissimilarity values, but converts them into the ranks and use these ranks in the calculation.
- The algorithm is iterative - it starts from the initial distribution of samples in the ordination space, and by the iterative reshuffling of samples it searches for optimal final distribution.
- Due to the iterative nature of the algorithm, each run may result in a different solution.

UNCONSTRAINED ORDINATION

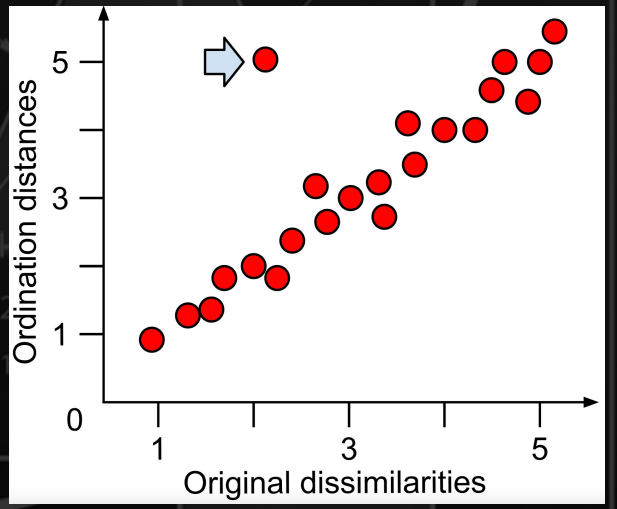
NMDS

The so-called stress value give us an estimation of how good the ordination is.

As a rule of thumb, an ordination with a stress value around or above 0.2 is deemed suspect and a stress value approaching 0.3 indicates that the ordination is arbitrary.

Stress values equal to or below 0.1 are considered fair, while values equal to or below 0.05 indicate good fit.

Plotting the observed stress values against the number of dimensions used in a series of NMDS runs can guide your selection of an appropriate number of dimensions.



Shepard stress plot showing the relationship between the actual dissimilarities between objects (from the original dissimilarity matrix) and the ordination distances (i.e. the distances on the final plot). If these are well correlated, the ordination stress will be low and the visualisation trustworthy. If there is a large amount of scatter (i.e. a poor linear relationship), then the ordination is not representative of the original distances. Occasionally, specific objects may be ordinated poorly (blue arrow), despite the overall solution being acceptable.

UNCONSTRAINED ORDINATION

NMDS

→Reading NMDS plots is quite straightforward: objects that are ordinated closer to one another are likely to be more similar than those further apart. However, the scale of the axes is arbitrary as is the orientation of the plot. Solutions with higher stress values (usually above 0.20) should be interpreted with caution and those with stress above 0.30 are highly suspect.

→Tight clusters of points that are well-separated from other clusters may indicate sub-populations in the data. The stress of the solution would be minimally affected by rearranging points in a tight cluster. Re-running an NMDS with only those objects in a given cluster may reveal more informative patterns.

→NMDS is suited to indirect gradient analysis. If the patterns in an ordination corroborate existing knowledge or a hypothesis, there may be grounds for a direct gradient analysis, a hypothesis test, or the design of a new sampling campaign targeting that variation.

UNCONSTRAINED ORDINATION

PCA

→Principal component analysis (PCA) is a linear unconstrained ordination method.

→It is implicitly based on Euclidean distances among samples, which is suffering from double-zero problem. As such, PCA is not suitable for heterogeneous compositional datasets with many zeros (so common in case of ecological datasets with many species missing in many samples).

→It can be applied to quantitative variables (these could be also negative), and also presence-absence data, but it cannot handle qualitative variables.

However, up-to-date methods in R allow to handle bot qualitative and quantitative data in a mixedPCA approach.

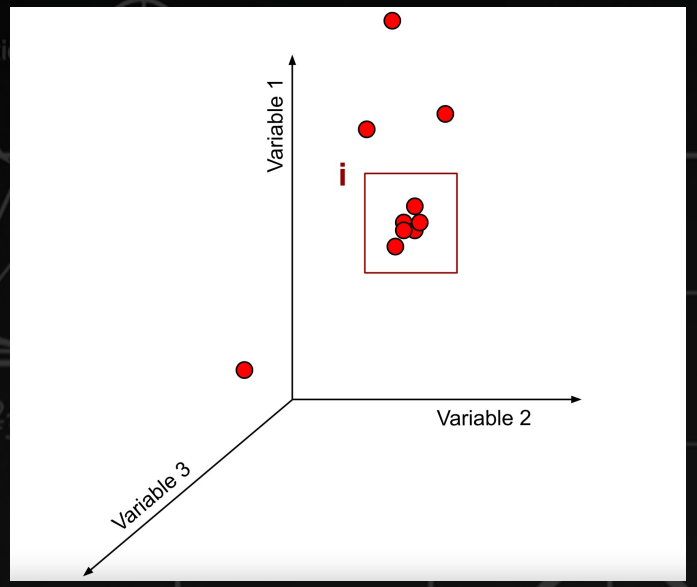
UNCONSTRAINED ORDINATION

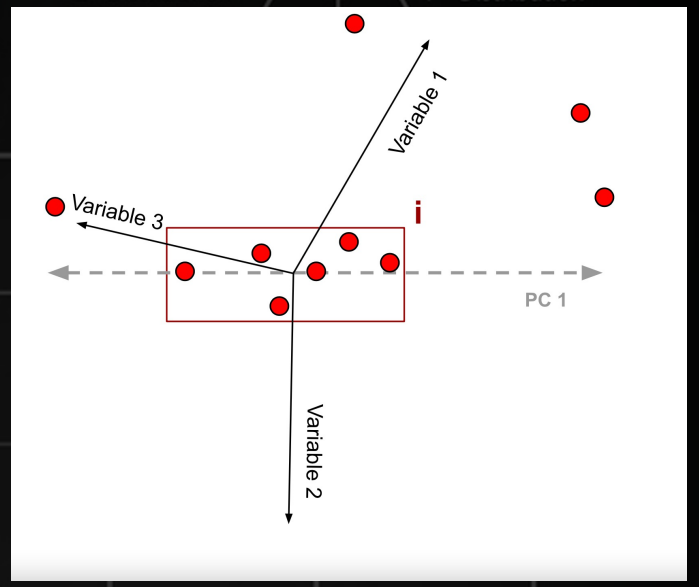
PCA

- PCA is a method to summarise, in a low-dimensional space, the variance in a multivariate scatter of points, providing an overview of linear relationships between your objects and variables.
- Imagine to have a data table with 50 objects described by 5 variables.
- A 5-dimensional scatter plot (i.e. a plot with 5 orthogonal axes) with each object's coordinates in the form (x1, x2, x3, x4, x5) is impossible to visualise and interpret.
- PCA attempts to express most of the variability in this 5-dimensional space by rotating it in such a way that a lower-dimensional representation will bring out most of the variability of the higher-dimensional space.
- A new set of axes (known as principal components) is created as a basis of the lower-dimensional representation.

UNCONSTRAINED ORDINATION

PCA





Panel on the left shows a 3-dimensional scatter plot in which the variability between the six points in box *i* is obscured. Panel on the right shows a rotation of the original axes to maximise the variability in a two dimensional space. The first principal component would be constructed in the direction of maximum scatter (i.e. maximum variability; dashed line). Subsequent PCs would be constructed in the same manner, however, must be orthogonal to (have no correlation with) all other PCs. The original variables would be rescaled as needed and may be represented in a biplot

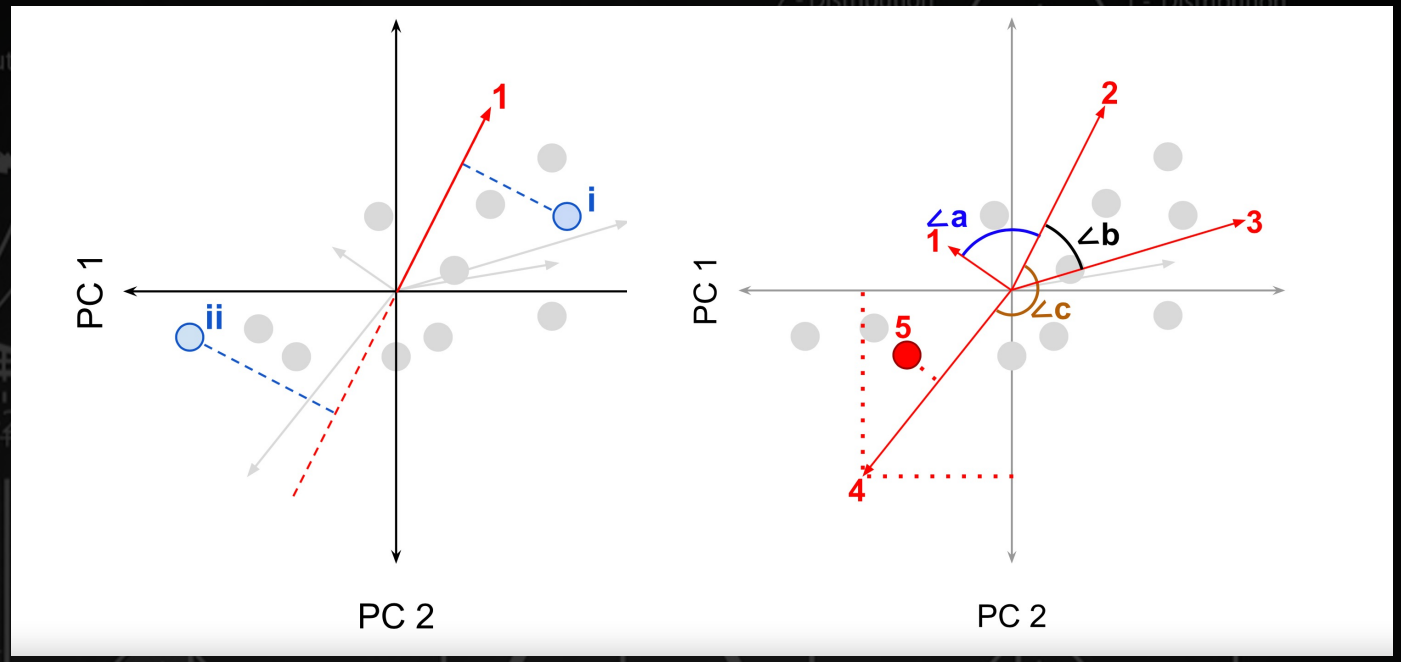
UNCONSTRAINED ORDINATION

PCA

- The total variance of the data set, that is, the total variance of all variables across all objects.
- The eigenvalues associated with each PC. Examining the proportion of variance explained attributed to each PC is useful in determining how much variation that PC is able to 'explain'.
- Objects and variables will have a score on each PCs calculated. The scores act as a new set of coordinates in the space described by PC axes.
- The variable loadings may be understood as how much each variable 'contributed' to building a PC. The absolute value of the loadings should be considered as the signs are arbitrary.

UNCONSTRAINED ORDINATION

PCA



The projection of an ordinated point onto a variable vector, as shown for point i in the left panel, approximates the variable's value realised for that object.

Cosines of angles between vectors (right panel) approximate the covariance between the variables they represent. In this case, ∠a is approaching 90, which suggests that variables "1" and "2" show very little covariance (i.e. they are almost orthogonal). ∠b is less than 90, suggesting positive covariance between variables "2" and "3" while ∠c is approaching 180, suggesting strong negative covariance between variables "2" and "4" (i.e. the directions of increase of variables "2" and "4" oppose one another).

CONSTRAINED ORDINATION

Constrained analysis is a form of direct gradient analysis, which attempts to explain variation in a data table directly through the variation in a set of explanatory variables (e.g. environmental factors) stored in a corresponding table or tables.

→Canonical Correspondence Analysis (CCA)

→Redundancy Analysis (RDA)

CONSTRAINED ORDINATION

CCA

→Canonical correspondence analysis (CCA) is a form of direct gradient analysis, wherein a matrix of explanatory variables intervenes in the calculation of the solution.

→Only correspondence that can be 'explained' by the matrix of explanatory variables is represented in the final results.

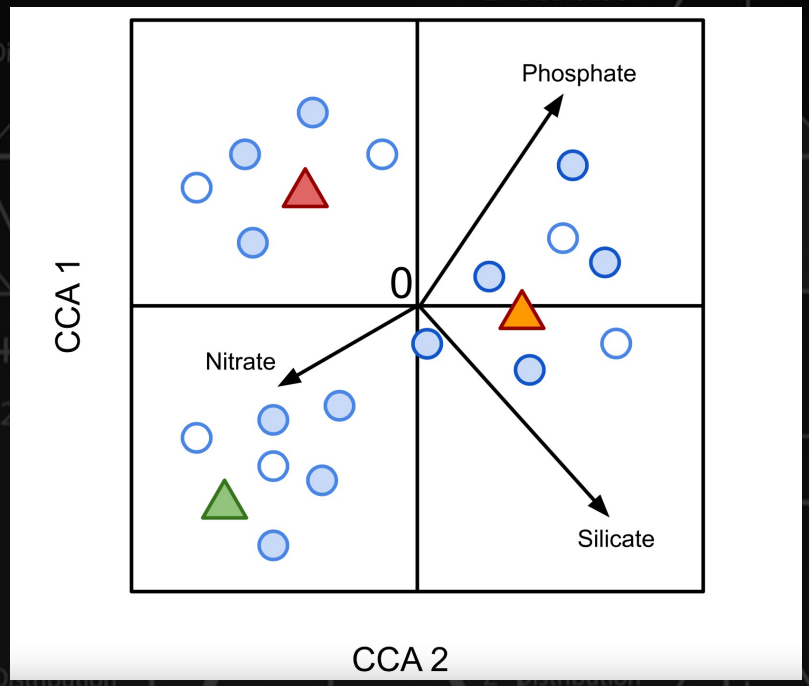
→CCA is suitable for response variables showing unimodal distributions and preserves χ2 (chi-squared) distances

→The correlation of the explanatory variables to the final ordination determines their 'importance'.

→CCA can be used to relate a qualitative explanatory variable to unimodal response data. The qualitative variable is recoded as a dummy variable and CCA is run.

CONSTRAINED ORDINATION

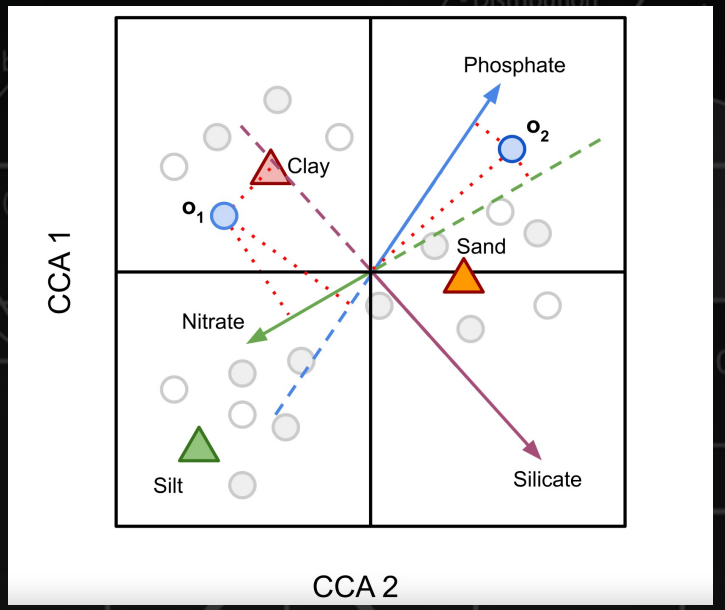
CCA



An illustrative schematic of a CCA triplot. Filled circles represent objects (e.g. sampling sites). Hollow circles represent response variables (e.g. OTU abundances). Arrows represent quantitative explanatory variables (here, nutrient concentrations) with arrowheads indicating their direction of increase. Filled triangles represent the states of a categorical explanatory variable (e.g. sand, silt, or clay sediment type).

CONSTRAINED ORDINATION

CCA



Two objects ("o1", "o2"), three quantitative explanatory variables ("Nitrate", "Phosphate", "Silicate") represented by vectors (arrows) pointing in the direction of increase and extended for clarity (dashed lines), and two states of a nominal (qualitative) variable, sediment type ("Sand", "Silt", "Clay"). Orthogonal projections are shown as dotted red lines. Object "o1" is very likely to be found in clay sediments while object "o2" is more likely to be found in sand sediments. Perpendicular projections of object "o1" onto quantitative explanatory variables suggests it realises high values of nitrate concentration, mid-to-low values of phosphate concentration, and low values of silicate concentration. Object "o2" realises high values of phosphate concentration, mid-range values of silicate concentration and low values of nitrate concentration.

CONSTRAINED ORDINATION

CCA

- Response variables show unimodal distributions across objects. This suggests that a sampling gradient must be long enough to allow the increase and decrease of a given species across the sites sampled.
- Gradients that are too short may manifest linear responses and may be better handled by redundancy analysis (RDA), although CCA may also handle linear relationships.
- Explanatory variables show linear, causal relationships to the response data. If one is unsure if their is a causal relationship between an explanatory variable and the response data, interpretation should be performed with care.

CONSTRAINED ORDINATION

RDA

- Redundancy analysis (RDA) is a method to extract and summarise the variation in a set of response variables that can be explained by a set of explanatory variables.
- RDA is a direct gradient analysis technique which summarises linear relationships between components of response variables that are "redundant" with (i.e. "explained" by) a set of explanatory variables.
- RDA extends multiple linear regression (MLR) by allowing regression of multiple response variables on multiple explanatory variables.
- A matrix of the fitted values of all response variables generated through MLR is then subject to principal components analysis (PCA).

CONSTRAINED ORDINATION

RDA

→RDA can also be considered a constrained version of principal components analysis (PCA), wherein canonical axes - built from linear combinations of response variables - must also be linear combinations of the explanatory variables (i.e. fitted by MLR).

→The RDA approach generates one ordination in the space defined by the matrix of response variables and another in the space defined by the matrix of explanatory variables

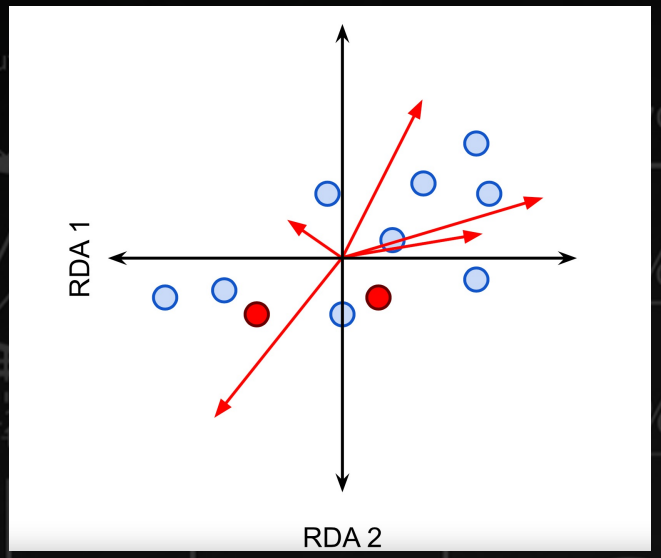
CONSTRAINED ORDINATION

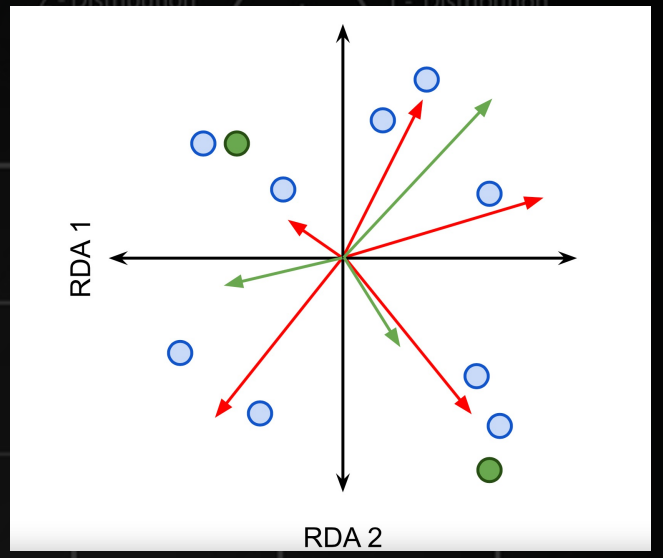
RDA

- If your response variables are not dimensionally homogeneous (i.e. if they have different base units of measurement), you may centre them on their means or standardise them .
- Ensure the number of explanatory variables is less than the number of objects in your data matrices. If not your system is overdetermined.
- If your explanatory variables are not dimensionally homogeneous (e.g. have different physical units), centre them on their means and standardise them.
- Examine the distribution of each variable in you explanatory and response matrix. If the relationships are markedly non-linear, apply transformations to linearise the relationships and reduce the effect of outliers.
- If you wish to represent non-Euclidean relationships (e.g. Hellinger distances) between objects in an RDA ordination, you should apply an ecologically-motivated transformation discussed on this page before analysis.

CONSTRAINED ORDINATION

RDA





Schematic representation of an RDA biplot (figure on the right) and an RDA triplot (figure on the left). An RDA biplot ordinates objects as points and either response or explanatory variables as vectors (red arrows). Levels of nominal variables are plotted as points (red). In a triplot, objects are ordinated as points (blue) while both response and explanatory variables (red and green arrows resp.) are plotted as vectors. Levels of nominal variables are plotted as points (green).