Multivariate analysis

One matrix of variables
"unsupervised methods"
Ressemblance between :

Y (=descriptors)
= "R mode analysis"

Covariance, Pearson correlation :
Y quantitative (and linear relation)
Spearman + Kendal Tau correlation :
Y semi-quantitative (and/or monotonic relation)
Contingency tables metrics (Chi sq etc...):
Y qualitative (OK for non monotonic relation)
Some distances indices from the Q mode...
+ others ie : niche similarity, niche overlap,...

Observations
= "Q mode analysis"

Similarity / Distance:
> 30 indices
(depending on the type of data) :
Euclidean, Bray-Curtis,
Sorensen, Jaccard, Gower,
Chi squared, Hellinger, Chord,...

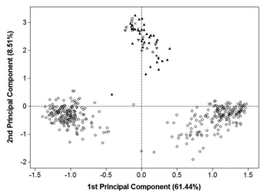
+ (other descriptive stats : ie biodiversity indices,...)

Ordination in reduced space

- Reducing the number of descriptors
- Finding (indirrect, "unknown") gradients along observations explaining most of their variation

Principal component Analysis (PCA)
Transformation based PCA (tbPCA)
Correspondance Analysis (CA)
Principal Coordinale Analysis (PCoA = MDS)
Non Metric Multidimensionnal Scaling (nMDS)

PCA : sensitive to double 0, Cor matrix, Euclidian dist
CA, tbPCA : unsensitive to double 0,
CA : chisq dist - tbPCA : Chord, Hellinger dist,...
PCoA, NMDS : any distance matrix



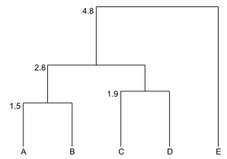
Clustering

partitionning observations
(or sometimes descriptors)
into groups based on
similarity/distance matrices

Dendrograms
(hierarchical clustering)

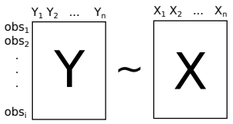
Non hierarchical clustering :
K-means partitionning
Medoids partitionning

Description of the groups :
ie indicator species
(**Twinspan, Indval**)



Discriminant Analysis
Predict appartenance to
a priori defined groups
based on several variables

Relation between 2 Matrices
"supervised methods"



Canonical Ordination

= dirrect gradient analysis
= Ordination on predicted values from GLM for each Y

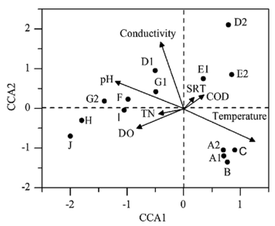
Redundancy Analysis (RDA + tbRDA)
= canonical PCA (or tbPCA)
Strictly equivalent to MANOVA/MANCOVA

Canonical Correspondance Analysis (CCA)
= canonical CA
distance based RDA (dbRDA)
= PCoA + RDA

+ Variance Partitionning, Partial Regression

Multivariate Correlations

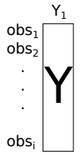
Canonical Correlation, Mantel tests,...



Other Approaches :

Multivariate Regression Trees (MRT)
Repeated GLM on each Y
Multivariate GLM

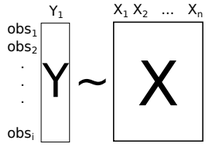
Univariate analysis



One variable

Descriptive statistics :
- position : mean, median quantiles, mode,...
- variability : variance, standard deviation
do not confound with Standard error, Confidence Intervals

Different approaches for inference :
Null Hypothesis testing, parametric Confidence intervals
Permutation tests
Bootstraping
Montecarlo simulations, parametric bootstrap
MCMC simulations + Bayesian approaches,...



Relation between 1 dependent variable (Y, response) and a matrix of explanatory variables (X, predictors)
"supervised methods"
Predict the values of Y based on the values of X

When number of X grow : you need **Model selection**
ie by Information criterion (AIC & co),
model inference, model averaging
ridge regression, lasso regression,...

"Regression based approaches"

Type of data	***"Classical" statistics***	***"Unified" statistics***
Quantitative Y vs Quantitative X	**Linear regression**, **multiple regression**	**(General) Linear Models**
Compare the means of 2 groups = Quantitative Y vs qualitative X (2 levels)	**t test**, paired t test Mann-Whithey, Wilcoxon	
Compare the means of n groups = Quantitative Y vs qualitative X (n levels)	**ANOVA** (Analysis of Variance) Kruskal-Wallis test, Friedman test	
Quantitative Y vs quantitative + qualitative X	**ANCOVA** (Analysis of Covariance)	
Binary Y vs Quantitative and/or qualitative X	**Logistic regression**, Probit regression	**Generalized Linear Models**
Qualitative Y vs Qualitative X = Contingency tables	Chi square test, **G test**, Fisher exact test	
Y = counts or % (nbr of success/nbr trials)	Often analized with ANOVA or regression after variable transformation (ArcSin, log,...)	
Groups of observations, repeated observations	Nested ANOVA, random ANOVA, mixed ANOVA Repeated measures ANOVA, Randomized block designs, Latin squares designs, etc...	**(Generalized) Linear Mixed Models**

Acronyms : LM (R), GLM (SAS)
Estimation method : Sum of Squares
Residuals with a Gaussian distribution

Acronyms : GLM (R), GENMOD (SAS), GLIM, ...
Estimation method : Maximum likelihood
Several residual distribution available :
Gaussian, Poisson, Binomial, Gamma,...

Acronyms : LMM, GLMM, lme
Estimation method :
(Restricted) Maximum likelihood

LM extensions

Partial regression, Variance partitioning
Path analysis
Structural Equations Modelling

GLM & Mixed models extensions

Generalized Additive (Mixed) Models : GAM
Multivariate Adaptive Regression Splines : MARS
Modelize non linear relationship

Possibility to **modelize** (spatial, temporal)
residual correlation & variance heterogeneity
With Mixed models or generalized least squares

Zero Inflated Poisson models (ZIP) , **Hurdle models**,
Modelisation of overdispersion (quasilikelihood or other methods)

Other non standard GLMM / Mixture models:
Site occupancy models, state space models

Generalized Estimating Equations (GEE)
Another way to modelise repeated measurements,
and residuals correlation - used in TRIM software

MCMC simulations + Bayesian Approches :
ideal for high complexity models (fitting+ inference)

Other approaches "not regression based"

Machne learning, Data mining, Computer intensive, algorithmic,... approaches

Tree based methods :
Classification and Regression Trees (CART)
Random forests (RF)
Boosted Regression Trees (BRT)
= Generalized Boosted Models (GBM)
No need to specify non linearity, interactions, etc...

Neural Networks
Support Vector Machines (SVM)

Niche/Entropy modelling
ENFA, MAXENT,...
Based on species presences only + pseudo-absences
(ie need a full caracterisation of the available environment)