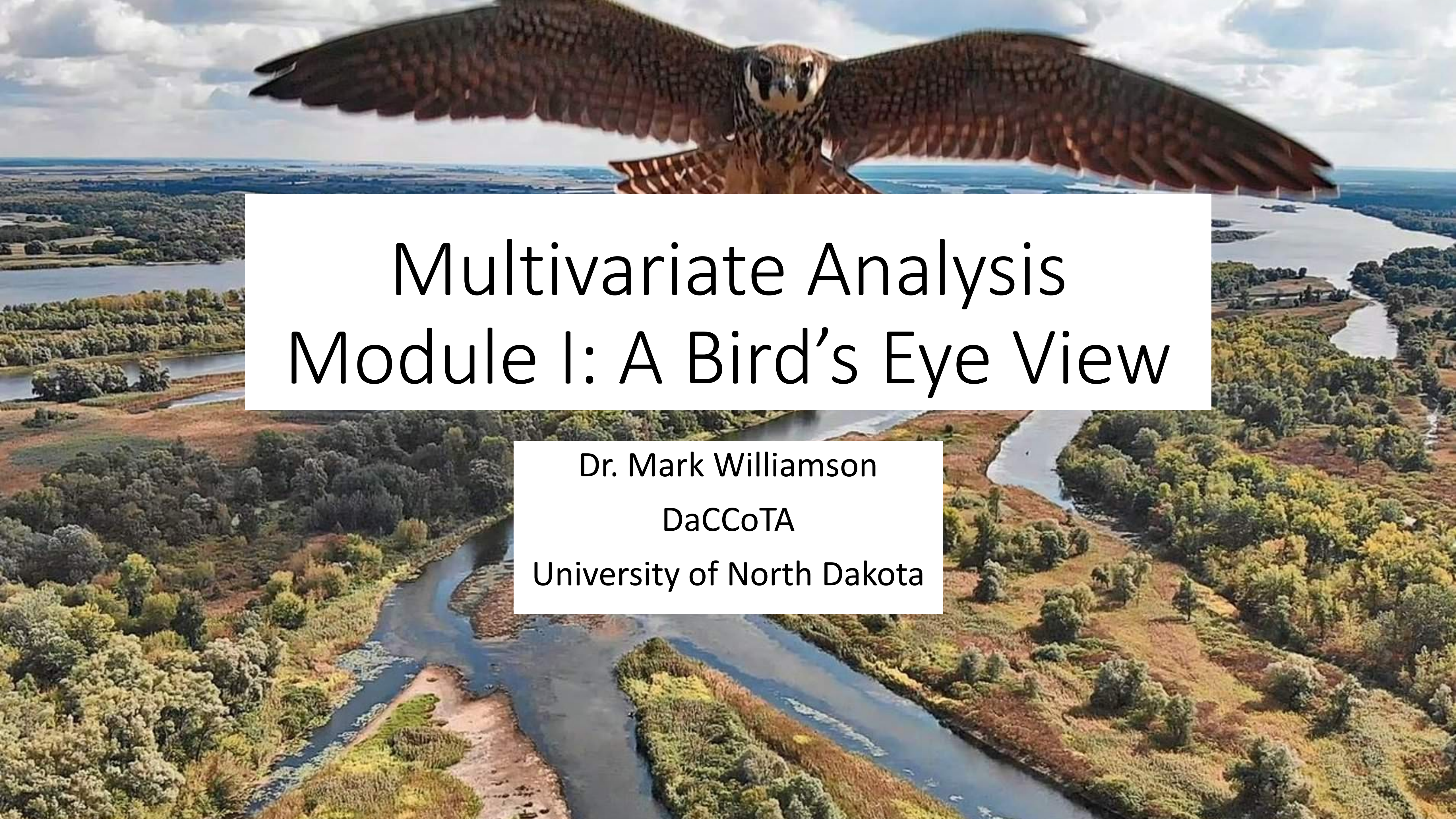


Multivariate Analysis Module I: A Bird’s Eye View

Dr. Mark Williamson
DaCCoTA
University of North Dakota

Introduction

What is Multivariate analysis?

- "multiple dependent variables resulting in one outcome" [1]
- "statistical models that have two or more dependent or outcome variables" [2]
- "matrices" [3]

Confusions

- Sometimes uses interchangeably with multivariable [2]
- Other classifications ANOVA, logistic regression, etc.
- 'Multivariate regression' also thrown around wildly

<u>Multivariate:</u>
multiple dependent variables or other more complicated structures such as ordination or non-linearity

$$(1)\ y = \alpha + x\beta + \varepsilon$$

$$(2)\ y = \alpha + x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k + \varepsilon$$

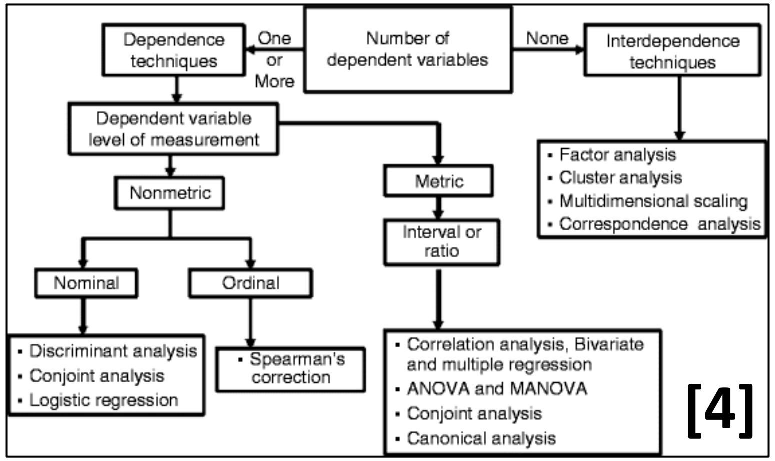
$$(3)\ Y_{n\times p} = X_{n\times(k+1)}\,\beta_{(k+1)\times p} + \varepsilon$$

- **Single response and single predictor (simple regression)**
- **Single response and multiple predictors (multiple regression)**
- **Multiple responses and predictors (multivariate regression)**

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Classification Problems

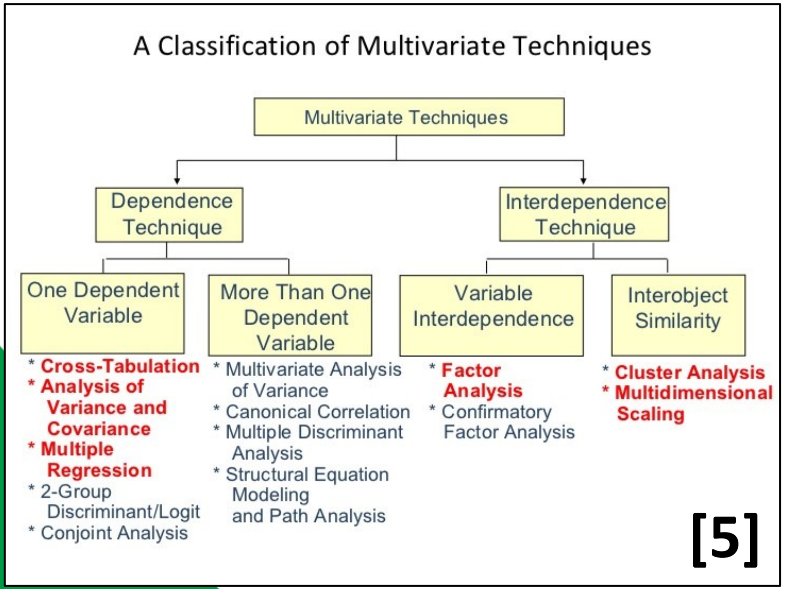

Number of dependent variables
One or More
None
Dependence techniques
Interdependence techniques
Dependent variable level of measurement
Nonmetric
Metric
Nominal
Ordinal
Interval or ratio
- Discriminant analysis
- Conjoint analysis
- Logistic regression

- Spearman's correction

- Correlation analysis, Bivariate and multiple regression
- ANOVA and MANOVA
- Conjoint analysis
- Canonical analysis

- Factor analysis
- Cluster analysis
- Multidimensional scaling
- Correspondence analysis

[4]


A Classification of Multivariate Techniques

Multivariate Techniques
Dependence Technique
Interdependence Technique
One Dependent Variable
More Than One Dependent Variable
Variable Interdependence
Interobject Similarity

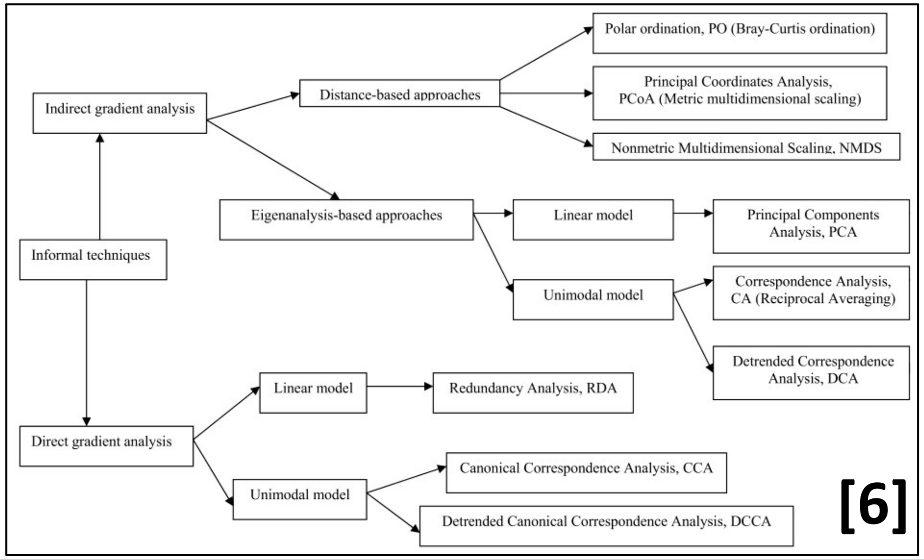
* Cross-Tabulation
* Analysis of Variance and Covariance
* Multiple Regression
* 2-Group Discriminant/Logit
* Conjoint Analysis

* Multivariate Analysis of Variance
* Canonical Correlation
* Multiple Discriminant Analysis
* Structural Equation Modeling and Path Analysis

* Factor Analysis
* Confirmatory Factor Analysis

* Cluster Analysis
* Multidimensional Scaling

[5]


Informal techniques
Indirect gradient analysis
Distance-based approaches
Polar ordination, PO (Bray-Curtis ordination)
Principal Coordinates Analysis, PCoA (Metric multidimensional scaling)
Nonmetric Multidimensional Scaling, NMDS
Eigenanalysis-based approaches
Linear model
Principal Components Analysis, PCA
Unimodal model
Correspondence Analysis, CA (Reciprocal Averaging)
Detrended Correspondence Analysis, DCA
Direct gradient analysis
Linear model
Redundancy Analysis, RDA
Unimodal model
Canonical Correspondence Analysis, CCA
Detrended Canonical Correspondence Analysis, DCCA

[6]

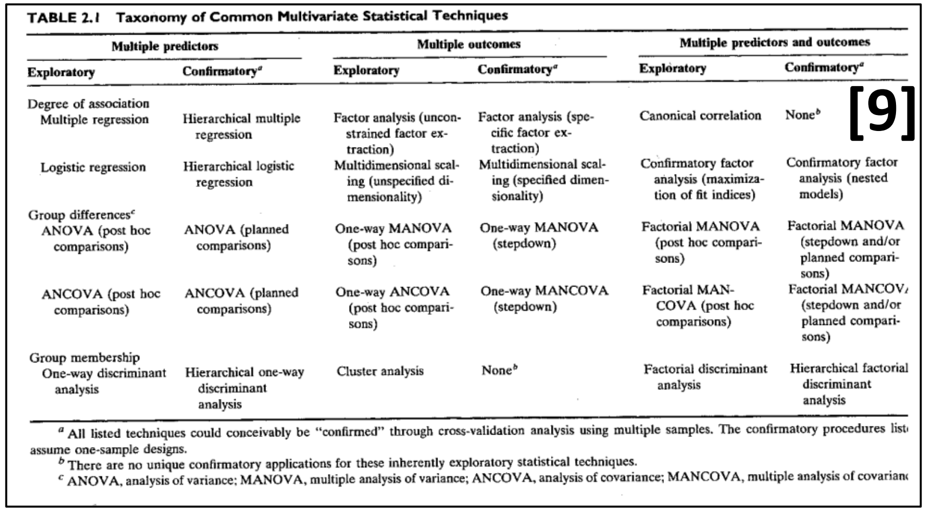
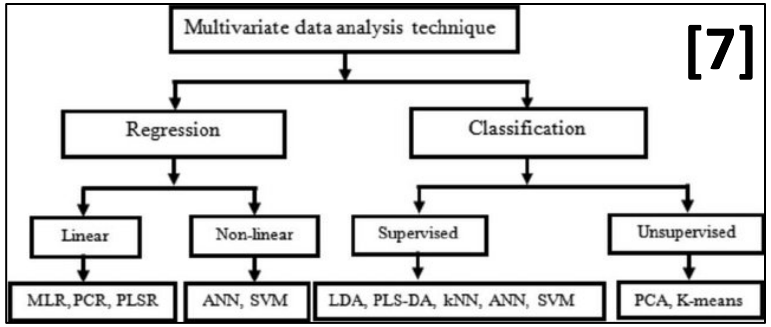


TABLE 2.1 Taxonomy of Common Multivariate Statistical Techniques

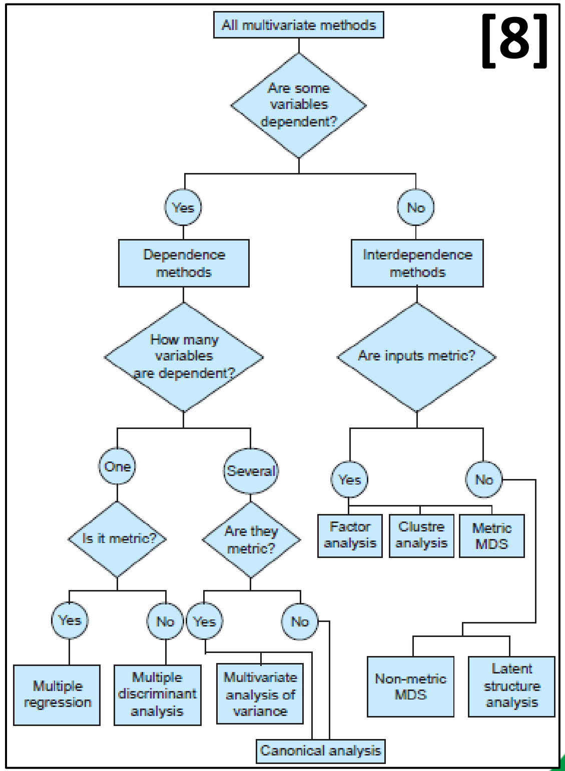
Multiple predictors		Multiple outcomes		Multiple predictors and outcomes	
Exploratory	Confirmatory[a]	Exploratory	Confirmatory[a]	Exploratory	Confirmatory[a]
Degree of association Multiple regression	Hierarchical multiple regression	Factor analysis (unconstrained factor extraction)	Factor analysis (specific factor extraction)	Canonical correlation	None[b]
Logistic regression	Hierarchical logistic regression	Multidimensional scaling (unspecified dimensionality)	Multidimensional scaling (specified dimensionality)	Confirmatory factor analysis (maximization of fit indices)	Confirmatory factor analysis (nested models)
Group differences[c] ANOVA (post hoc comparisons)	ANOVA (planned comparisons)	One-way MANOVA (post hoc comparisons)	One-way MANOVA (stepdown)	Factorial MANOVA (post hoc comparisons)	Factorial MANOVA (stepdown and/or planned comparisons)
ANCOVA (post hoc comparisons)	ANCOVA (planned comparisons)	One-way ANCOVA (post hoc comparisons)	One-way MANCOVA (stepdown)	Factorial MANCOVA (post hoc comparisons)	Factorial MANCOV/ (stepdown and/or planned comparisons)
Group membership One-way discriminant analysis	Hierarchical one-way discriminant analysis	Cluster analysis	None[b]	Factorial discriminant analysis	Hierarchical factorial discriminant analysis

[a] All listed techniques could conceivably be "confirmed" through cross-validation analysis using multiple samples. The confirmatory procedures list assume one-sample designs.
[b] There are no unique confirmatory applications for these inherently exploratory statistical techniques.
[c] ANOVA, analysis of variance; MANOVA, multiple analysis of variance; ANCOVA, analysis of covariance; MANCOVA, multiple analysis of covarian

[9]

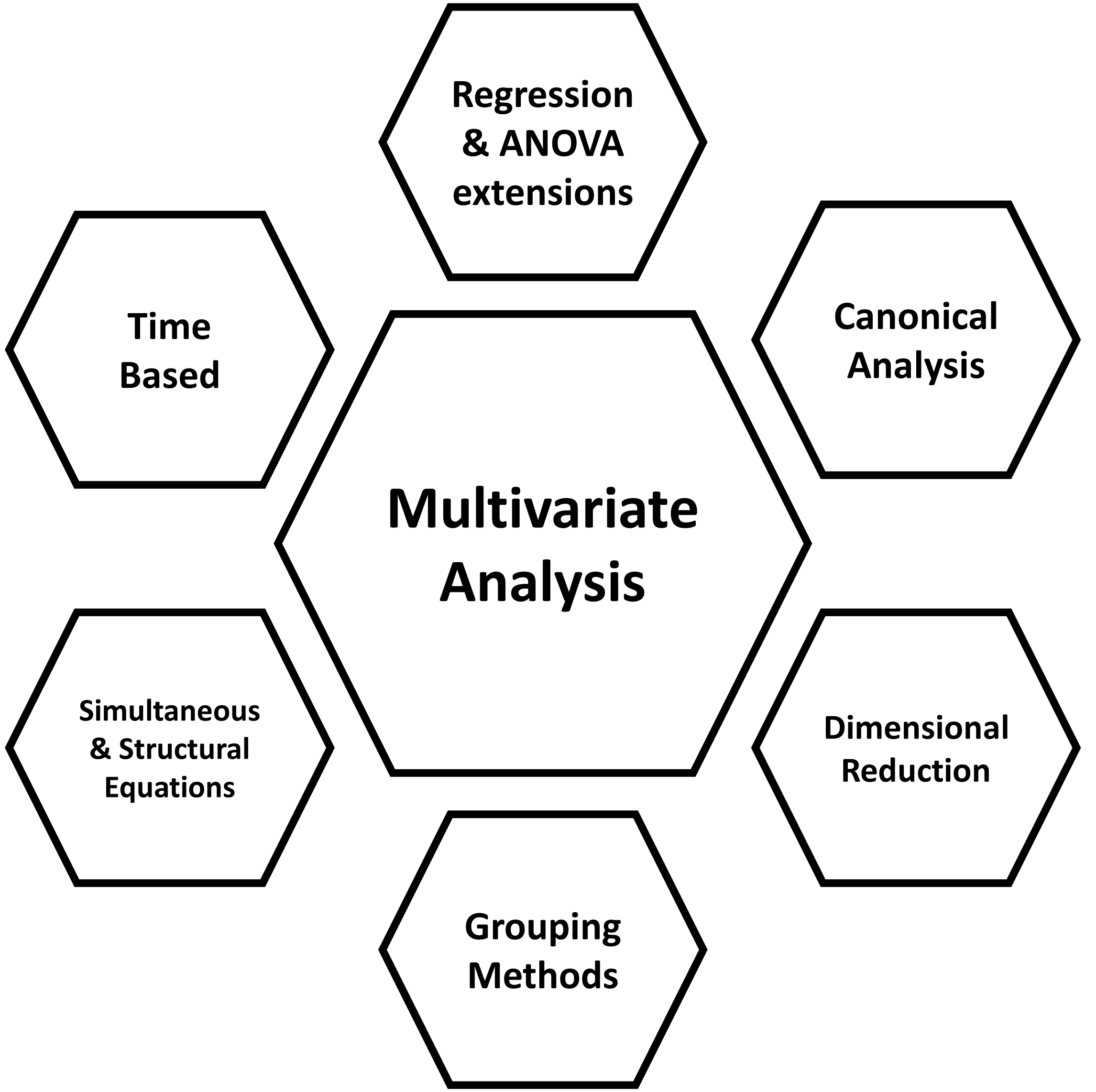

Multivariate data analysis technique
Regression
Classification
Linear
Non-linear
Supervised
Unsupervised
MLR, PCR, PLSR
ANN, SVM
LDA, PLS-DA, kNN, ANN, SVM
PCA, K-means

[7]


All multivariate methods
Are some variables dependent?
Yes
No
Dependence methods
Interdependence methods
How many variables are dependent?
Are inputs metric?
One
Several
Yes
No
Is it metric?
Are they metric?
Factor analysis
Clustre analysis
Metric MDS
Yes
No
Yes
No
Multiple regression
Multiple discriminant analysis
Multivariate analysis of variance
Canonical analysis
Non-metric MDS
Latent structure analysis

[8]

Landscape



Multivariate Analysis

Regression & ANOVA extensions

Canonical Analysis

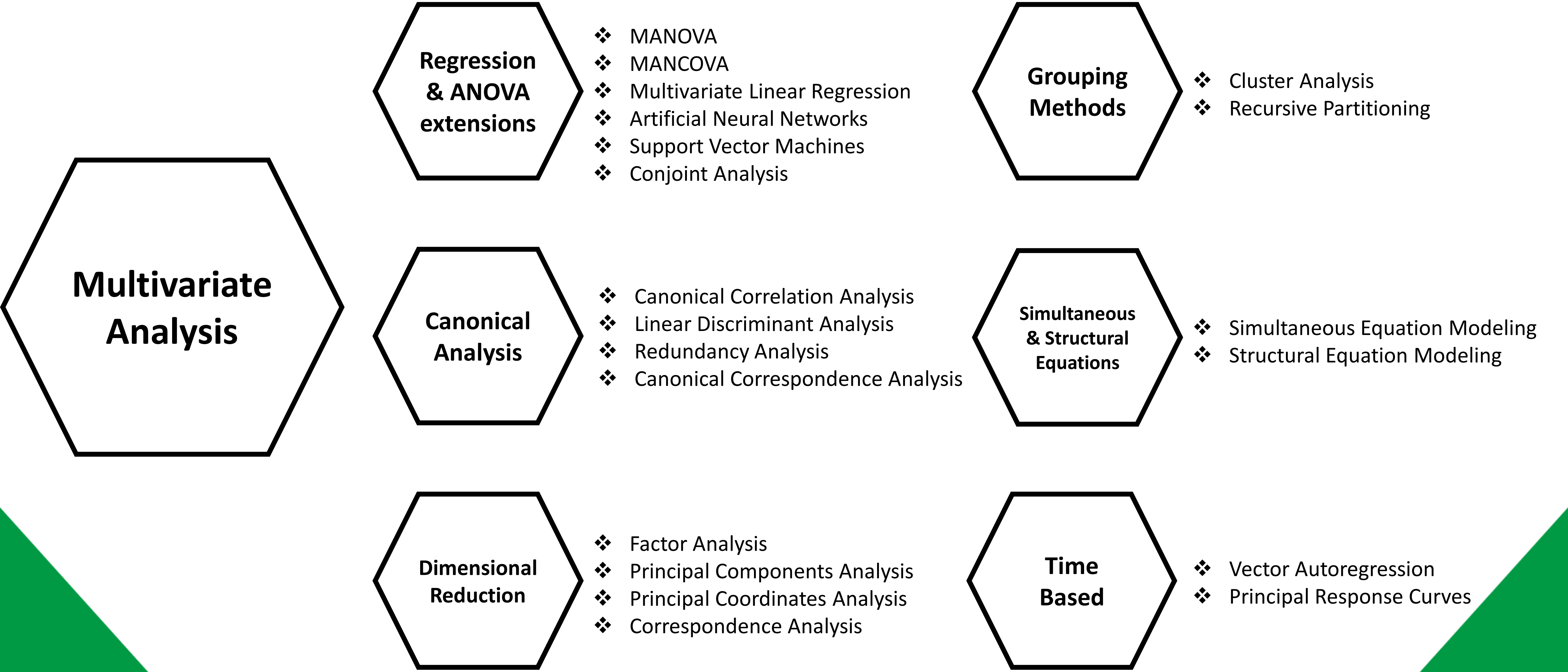
Dimensional Reduction

Grouping Methods

Simultaneous & Structural Equations

Time Based

Landscape



Multivariate Analysis

Regression & ANOVA extensions

- MANOVA
- MANCOVA
- Multivariate Linear Regression
- Artificial Neural Networks
- Support Vector Machines
- Conjoint Analysis

Grouping Methods

- Cluster Analysis
- Recursive Partitioning

Canonical Analysis

- Canonical Correlation Analysis
- Linear Discriminant Analysis
- Redundancy Analysis
- Canonical Correspondence Analysis

Simultaneous & Structural Equations

- Simultaneous Equation Modeling
- Structural Equation Modeling

Dimensional Reduction

- Factor Analysis
- Principal Components Analysis
- Principal Coordinates Analysis
- Correspondence Analysis

Time Based

- Vector Autoregression
- Principal Response Curves

Structures and Uses

Regression & ANOVA extensions

MANOVA:

- ◊ Multivariate Analysis of Variance [10]
- ◊ Extension of ANOVA with multiple Y variables
- ◊ Tests for the difference in two or more vectors

MANCOVA:

- ◊ Multivariate Analysis of Co-Variance [10]
- ◊ Extension of ANCOVA with multiple Y variables
- ◊ Covariate accounted for to reduce noise

Multivariate Linear Regression:

- ◊ Extension of Linear Regression with multiple Y variables

Support Vector Machines:

- ◊ Extends regression into non-linearity [12]
- ◊ Machine learning algorithms
- ◊ Supervised learning model
- ◊ Fast and less requirements to train

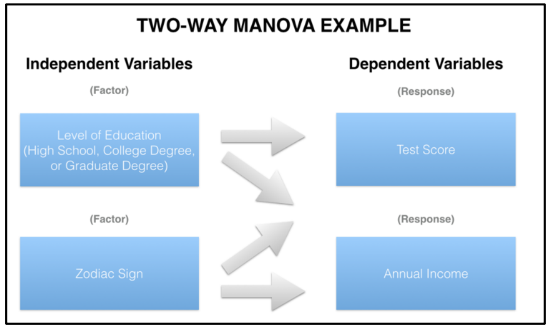
Artificial Neural Networks:

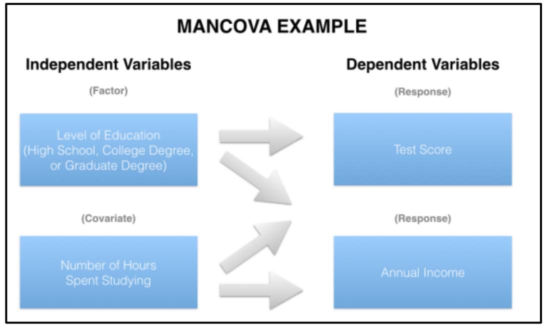
- ◊ Same base characteristics as SVM
- ◊ Simulates how brains process (collection of connected nodes)

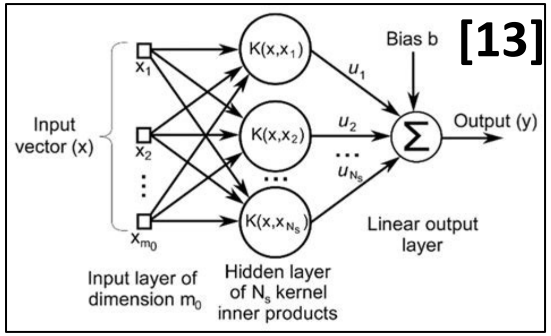
Conjoint Analysis:

- ◊ Extension of dummy variable regression [15]
- ◊ Survey-based analysis method
- ◊ Helps extract consumer preferences
- ◊ Sub-techniques include Choice-based conjoint, Adaptive conjoint, MaxDiff conjoint, etc.

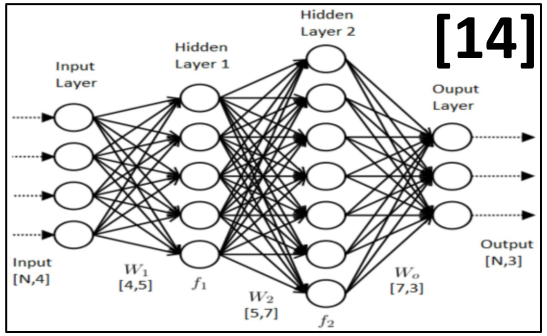
[11]



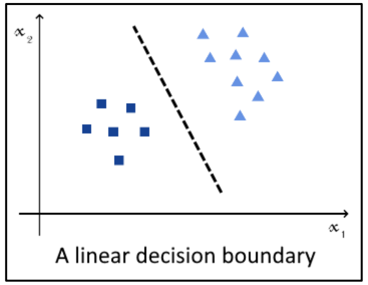




[13]

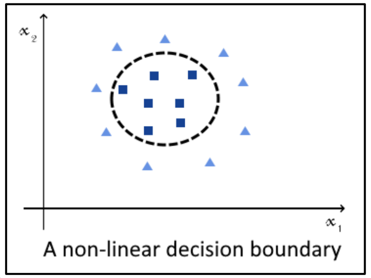


[14]

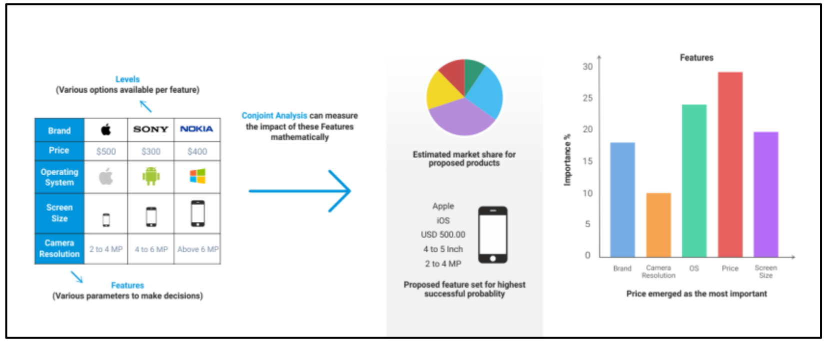


A linear decision boundary

[12]



A non-linear decision boundary


Levels
(Various options available per feature)
Brand
SONY
NOKIA
Price
$500
$300
$400
Operating System
Screen Size
Camera Resolution
2 to 4 MP
4 to 6 MP
Above 6 MP
Features
(Various parameters to make decisions)
Conjoint Analysis can measure the impact of these Features mathematically
Estimated market share for proposed products
Apple
iOS
USD 500.00
4 to 5 Inch
2 to 4 MP
Proposed feature set for highest successful probablity
Features
Importance %
0
5
10
15
20
25
30
Brand
Camera Resolution
OS
Price
Screen Size
Price emerged as the most important

[15]

Structures and Uses

Canonical Analysis

Canonical Correlation Analysis:

- Symmetrical [16] and linear [17]
- Method where two sets of variables have maximized correlation

Linear Discriminant Analysis:

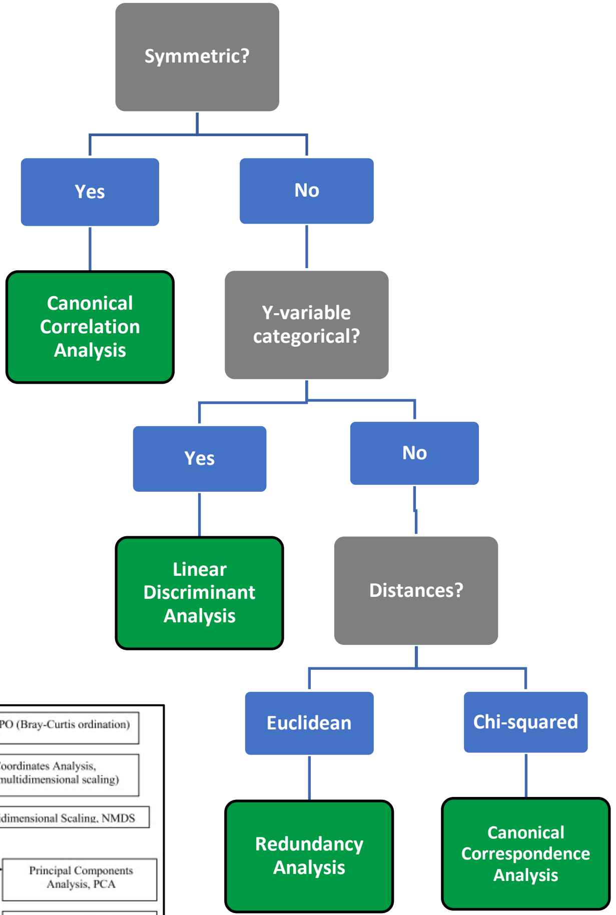
- Non-symmetric, Y is class of objects
- Essentially an extension of multinomial regression

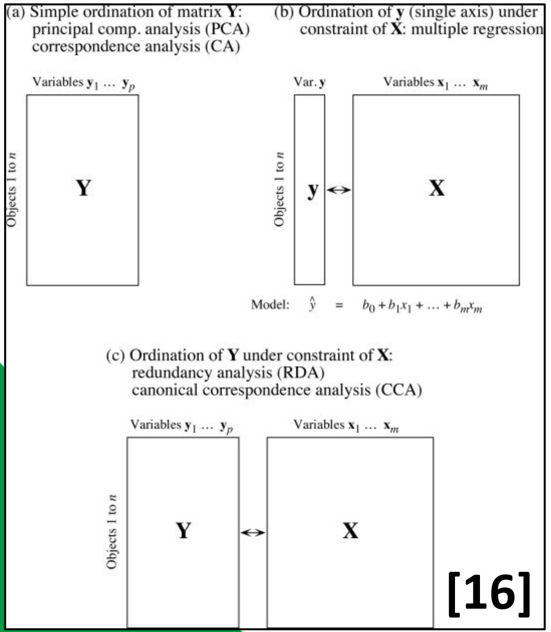
Redundancy Analysis:

- Non-symmetric, linear model, Euclidean distances
- Related to multiple linear regression, ordination added on [20]

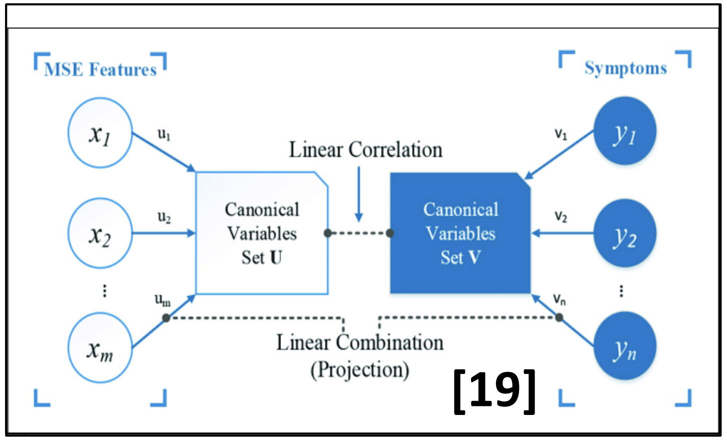
Canonical Correspondence Analysis:

- Non-symmetric, unimodal model, Chi-squared distances
- Related to multiple linear regression, ordination added on

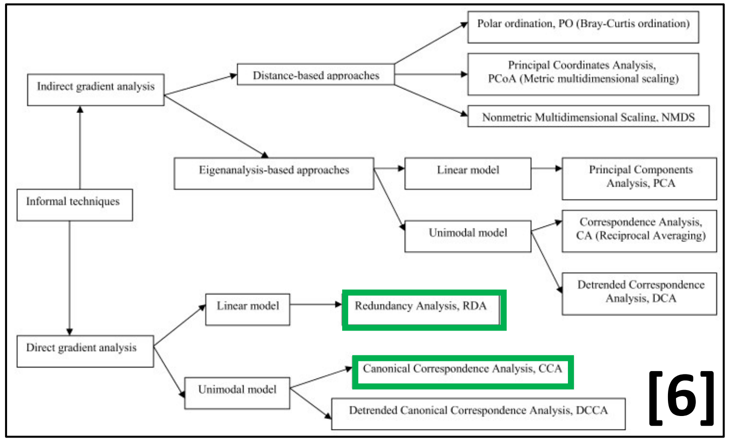



(a) Simple ordination of matrix Y: principal comp. analysis (PCA) correspondence analysis (CA)
Variables $\mathbf{y}_1 \ldots \mathbf{y}_p$
Objects 1 to n
Y
(b) Ordination of y (single axis) under constraint of X: multiple regression
Var. y
Variables $\mathbf{x}_1 \ldots \mathbf{x}_m$
Objects 1 to n
y
X
Model: $\hat{y} = b_0 + b_1x_1 + \ldots + b_mx_m$
(c) Ordination of Y under constraint of X: redundancy analysis (RDA) canonical correspondence analysis (CCA)
Variables $\mathbf{y}_1 \ldots \mathbf{y}_p$
Variables $\mathbf{x}_1 \ldots \mathbf{x}_m$
Objects 1 to n
Y
X

[16]



[19]


Indirect gradient analysis
Distance-based approaches
Polar ordination, PO (Bray-Curtis ordination)
Principal Coordinates Analysis, PCoA (Metric multidimensional scaling)
Nonmetric Multidimensional Scaling, NMDS
Eigenanalysis-based approaches
Linear model
Principal Components Analysis, PCA
Unimodal model
Correspondence Analysis, CA (Reciprocal Averaging)
Detrended Correspondence Analysis, DCA
Informal techniques
Direct gradient analysis
Linear model
Redundancy Analysis, RDA
Unimodal model
Canonical Correspondence Analysis, CCA
Detrended Canonical Correspondence Analysis, DCCA

[6]

Structures and Uses

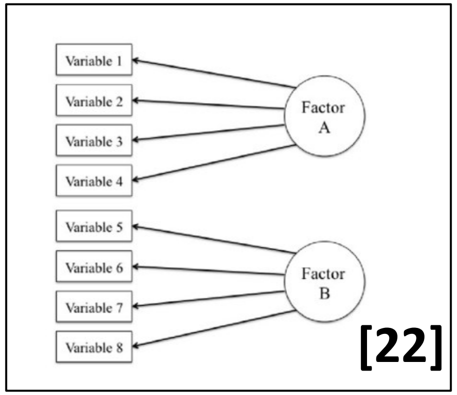
Dimensional Reduction

Factor Analysis:

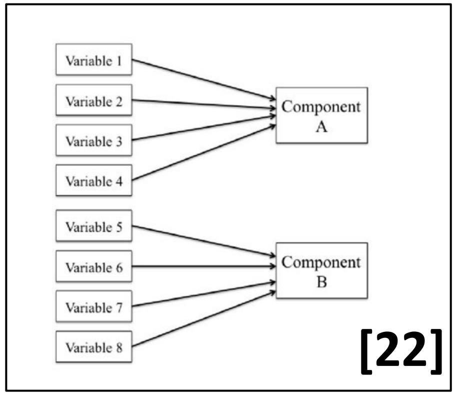
- ◊ Interdependence method that used ordination [21]
- ◊ Reducing observable variables into latent factors
- ◊ Both exploratory and confirmatory

Principal Components Analysis (PCA):

- ◊ Ordination technique
- ◊ Arranges data along gradients to simply
- ◊ Reduces dimensionality and still preserves most info
- ◊ Only highly correlated variables are together



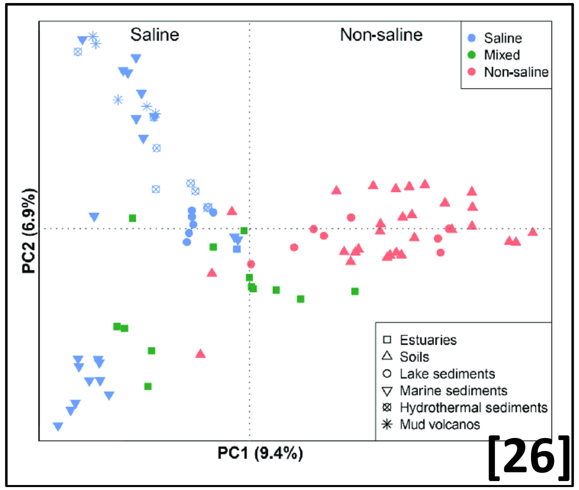
[22]



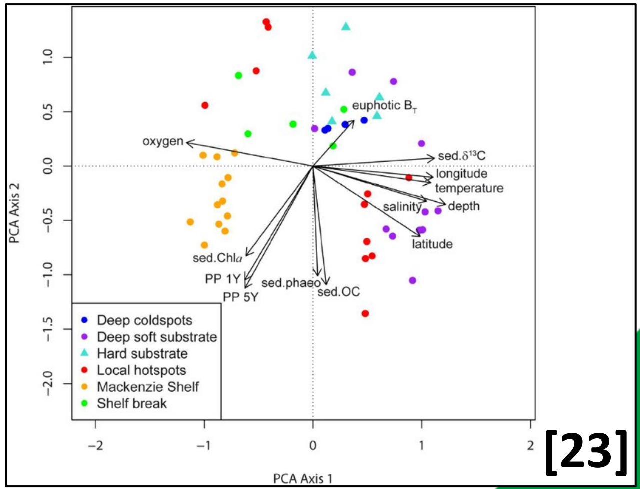
[22]

Principal Coordinates Analysis (PCoA):

- ◊ Ordination technique
- ◊ Arranges data along gradients to simply
- ◊ Represents distances between samples in low dimensional space [24]
- ◊ Better for missing data or fewer individuals than characteristics [25]



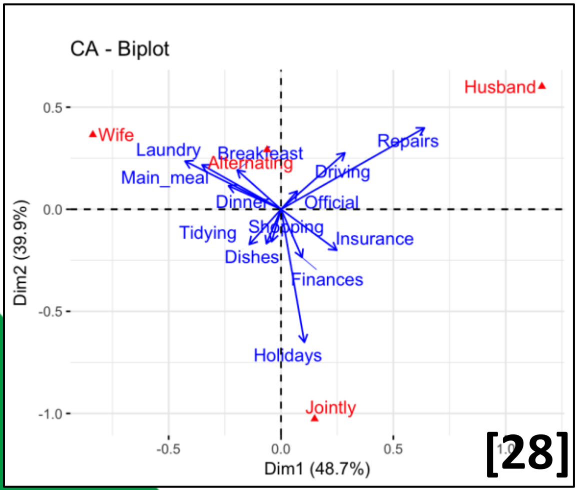
[26]



[23]

Correspondence Analysis:

- ◊ Like PCA but applies categorical rather than continuous data
- ◊ Applies contingency tables [27]



[28]

Structures and Uses

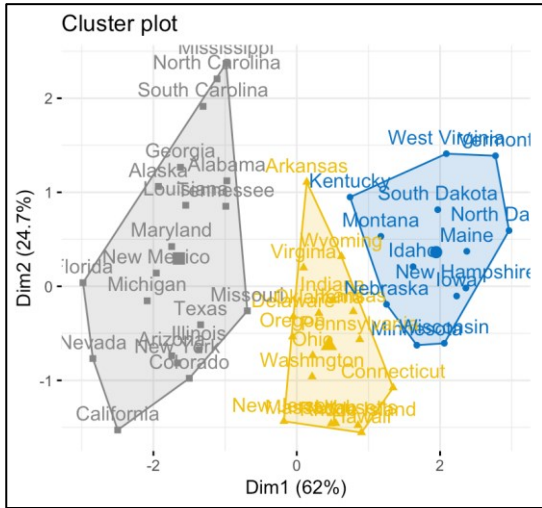
Grouping Methods

Cluster Analysis:

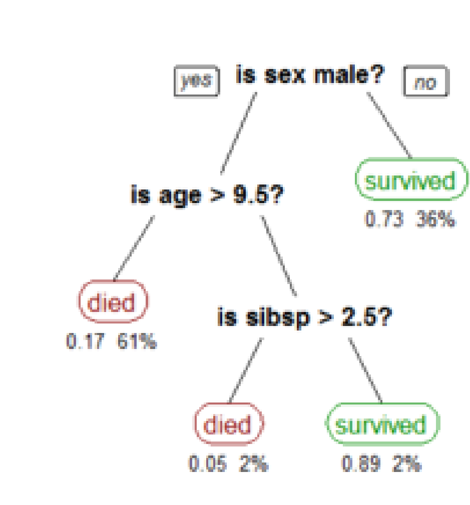
- Grouping set of objects so that each group is more similar to members than non-members [29]
- Difficult computational problem
- Specific types include hierarchical, k-means, distribution-based, density-based, etc. [30]

Recursive Partitioning:

- Inverse of clustering [32]
- Creates a decision tree to attempt to correctly classify members of the population based on several dichotomous dependent variables (yes/no)
- Part of the more general technique of decision trees
- Intuitive models that can be tweaked for sensitivity or specificity but can overfit data [33]



[31]



[33]

Structures and Uses

Simultaneous & Structural Equations

Simultaneous Equation Modeling:

◊ System of two or more linear simultaneous equations [34]

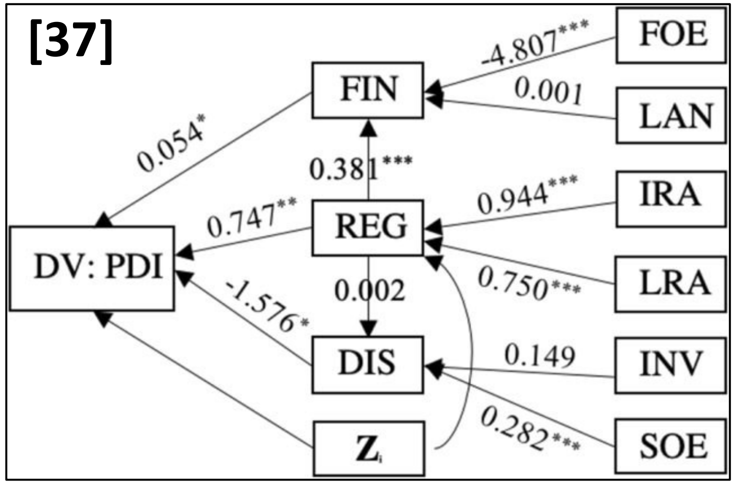
◊ Variables used are classified as endogenous (jointly determined, dependent) and exogenous (predetermined, independent) [35]

◊ Endogenous influenced by exogenous by not the other way round

◊ Complete model is when the number of endogenous variables equals the number of equations and is a Structural Equation Model

◊ Solution is determined by equilibrium among opposing forces [36]

◊ Structural form includes multiple endogenous variables; reduced form only has one; reduced can be estimated by least squares



[37]

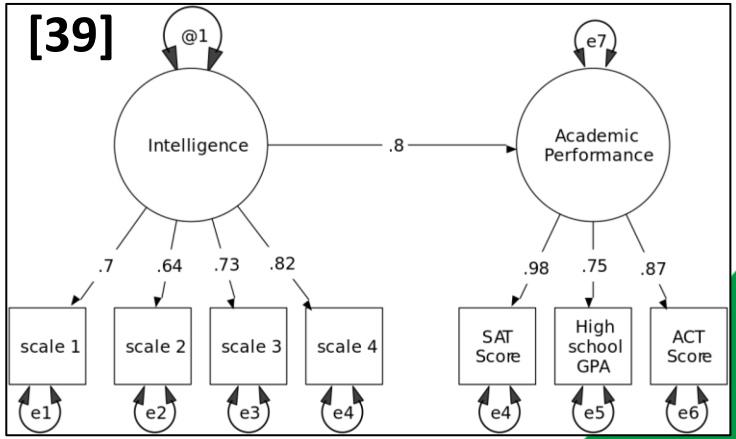
Structural Equation Modeling:

◊ Combination of factor analysis and regression [38]

◊ Interest is usually on latent factors that underlie observable variables

◊ Can be used to impute relationship between those latent factors from the observable variables

◊ Includes confirmatory factor analysis, confirmatory composite analysis, path analysis, partial least squares modeling, and latent growth modeling [29]



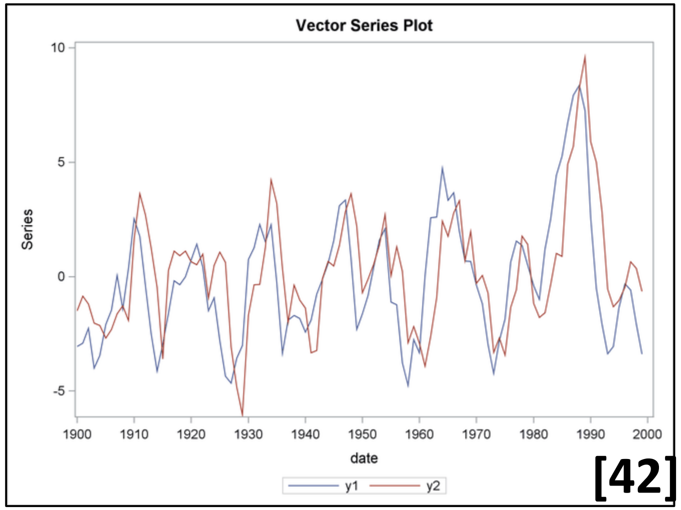
[39]

Structures and Uses

Time Based

Vector Autoregression:

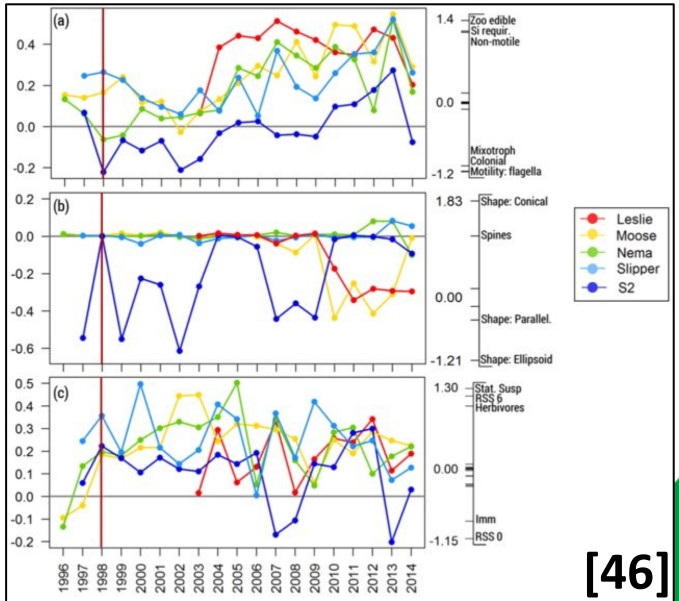
- ◊ Extension of univariate autoregressive model
- ◊ Stochastic process model that tries to understand the change in multiple quantities over time [40]
- ◊ Used when two or more time series influence each other [41]
- ◊ Each variable has an equation modeling its change over time, including past (lagged) values



[42]

Principal Response Curves:

- ◊ Used to analysis of treatment effects in repeated measures [43]
- ◊ Corrects for temporal trends in control treatments
- ◊ Special kind of redundancy analysis [44]
- ◊ Allows for summarizing and plotting of the results that is much more interpretable than a bi-plot [45]

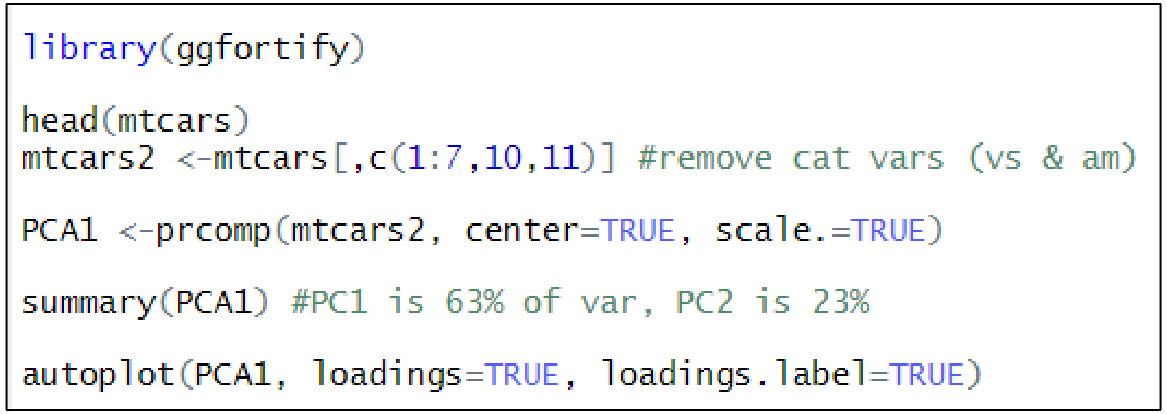


[46]

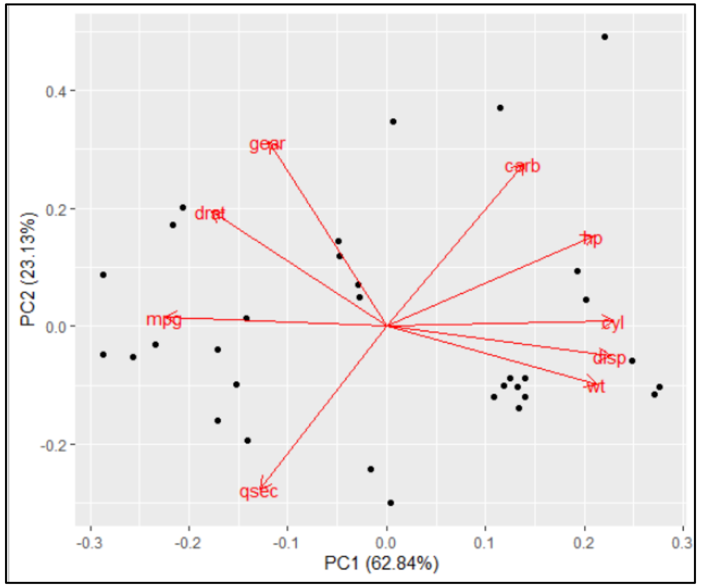
Examples

R

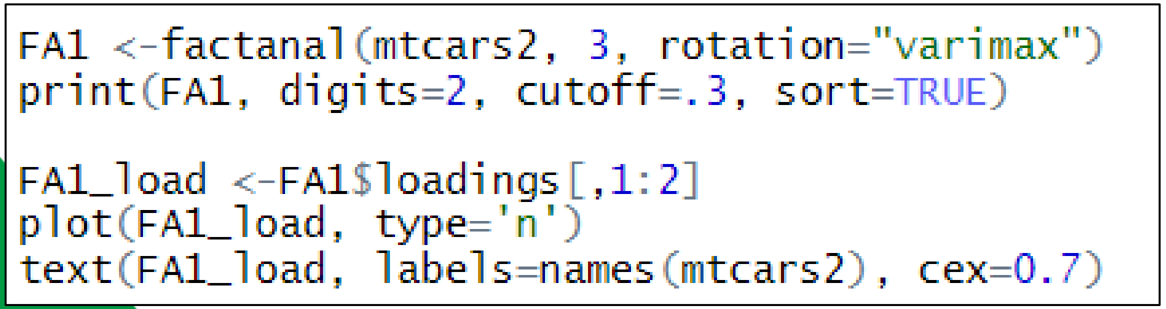
[47,49]



```
library(ggfortify)

head(mtcars)
mtcars2 <-mtcars[,c(1:7,10,11)] #remove cat vars (vs & am)

PCA1 <-prcomp(mtcars2, center=TRUE, scale.=TRUE)

summary(PCA1) #PC1 is 63% of var, PC2 is 23%

autoplot(PCA1, loadings=TRUE, loadings.label=TRUE)
```

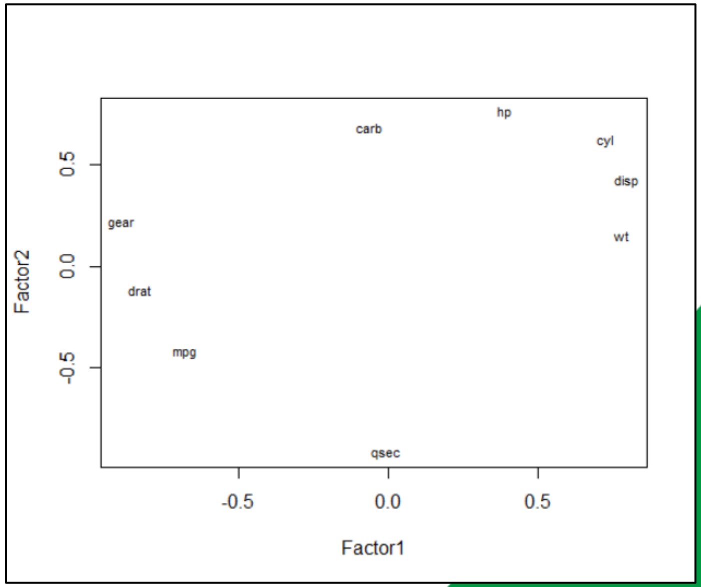


```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7    PC8     PC9
Standard deviation     2.3782 1.4429 0.71008 0.51481 0.42797 0.35184 0.32413 0.2419 0.14896
Proportion of Variance 0.6284 0.2313 0.05602 0.02945 0.02035 0.01375 0.01167 0.0065 0.00247
Cumulative Proportion  0.6284 0.8598 0.91581 0.94525 0.96560 0.97936 0.99103 0.9975 1.00000
```



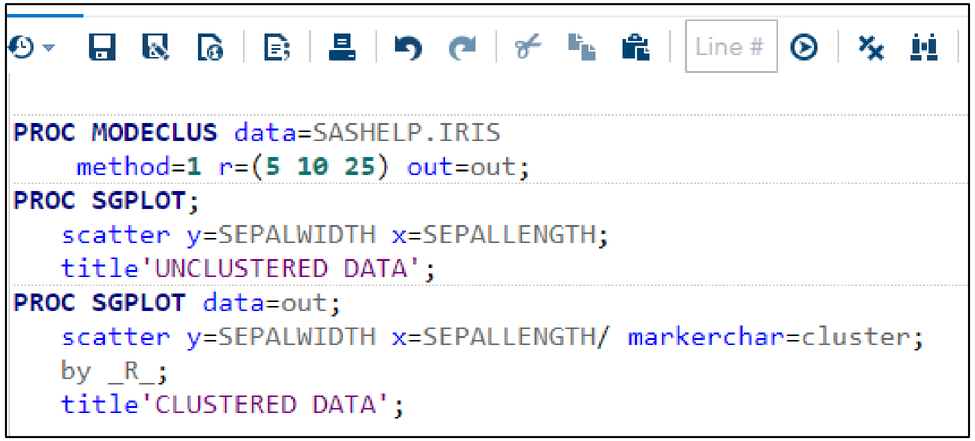
```
FA1 <-factanal(mtcars2, 3, rotation="varimax")
print(FA1, digits=2, cutoff=.3, sort=TRUE)

FA1_load <-FA1$loadings[,1:2]
plot(FA1_load, type='n')
text(FA1_load, labels=names(mtcars2), cex=0.7)
```

```
               Factor1 Factor2 Factor3
SS loadings       3.85    2.74    1.28
Proportion Var    0.43    0.30    0.14
Cumulative Var    0.43    0.73    0.87
```

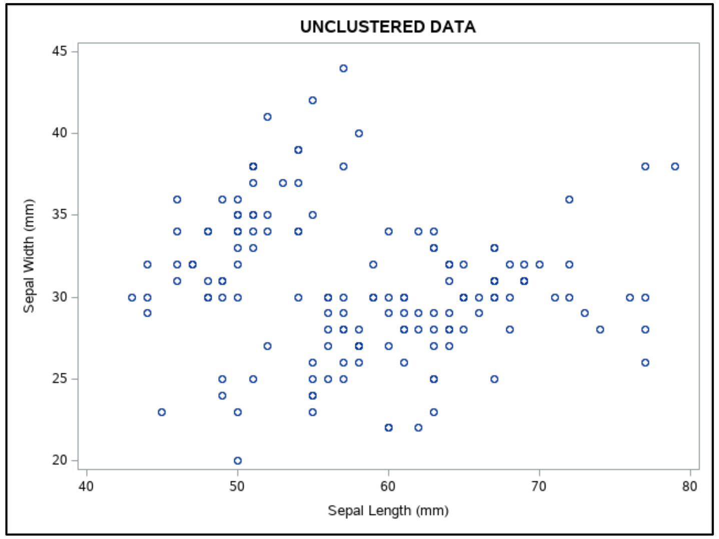


Examples

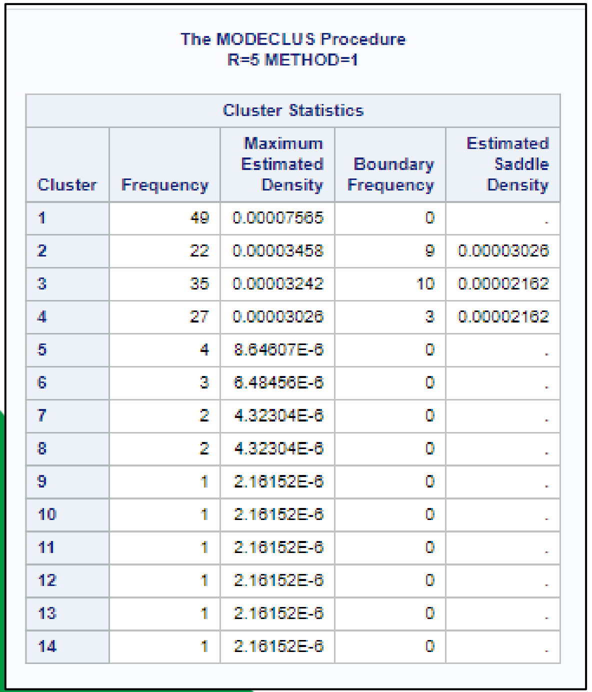
SAS [50]



```
PROC MODECLUS data=SASHELP.IRIS
    method=1 r=(5 10 25) out=out;
PROC SGPLOT;
   scatter y=SEPALWIDTH x=SEPALLENGTH;
   title'UNCLUSTERED DATA';
PROC SGPLOT data=out;
   scatter y=SEPALWIDTH x=SEPALLENGTH/ markerchar=cluster;
   by _R_;
   title'CLUSTERED DATA';
```



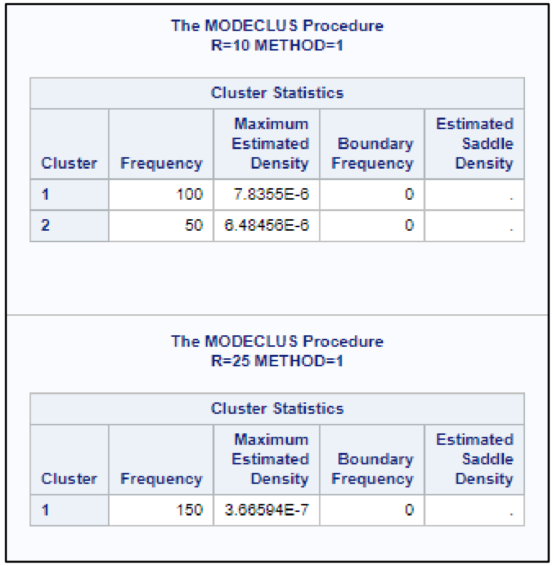
UNCLUSTERED DATA



The MODECLUS Procedure
R=5 METHOD=1

Cluster Statistics

Cluster	Frequency	Maximum Estimated Density	Boundary Frequency	Estimated Saddle Density
1	49	0.00007565	0	.
2	22	0.00003458	9	0.00003026
3	35	0.00003242	10	0.00002162
4	27	0.00003026	3	0.00002162
5	4	8.64607E-6	0	.
6	3	6.48456E-6	0	.
7	2	4.32304E-6	0	.
8	2	4.32304E-6	0	.
9	1	2.16152E-6	0	.
10	1	2.16152E-6	0	.
11	1	2.16152E-6	0	.
12	1	2.16152E-6	0	.
13	1	2.16152E-6	0	.
14	1	2.16152E-6	0	.



The MODECLUS Procedure
R=10 METHOD=1

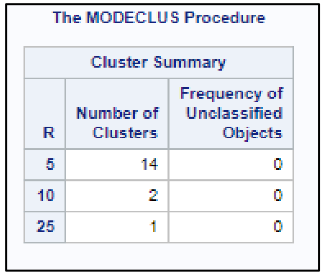
Cluster Statistics

Cluster	Frequency	Maximum Estimated Density	Boundary Frequency	Estimated Saddle Density
1	100	7.8355E-6	0	.
2	50	6.48456E-6	0	.

The MODECLUS Procedure
R=25 METHOD=1

Cluster Statistics

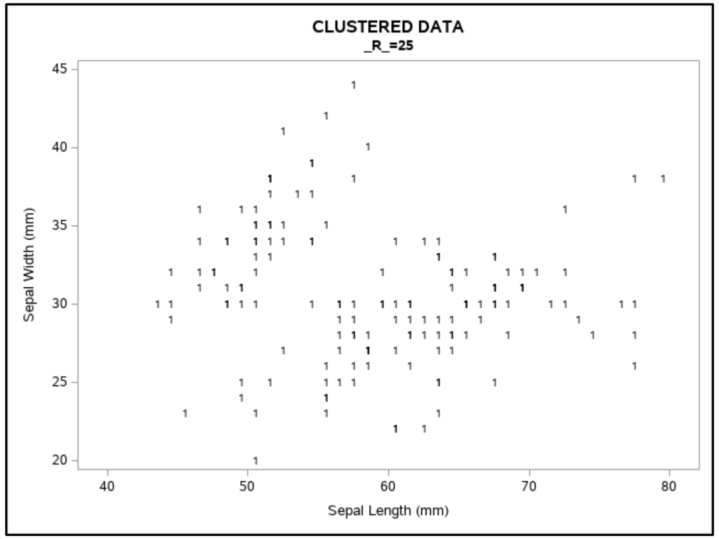
Cluster	Frequency	Maximum Estimated Density	Boundary Frequency	Estimated Saddle Density
1	150	3.66594E-7	0	.



The MODECLUS Procedure

Cluster Summary

R	Number of Clusters	Frequency of Unclassified Objects
5	14	0
10	2	0
25	1	0

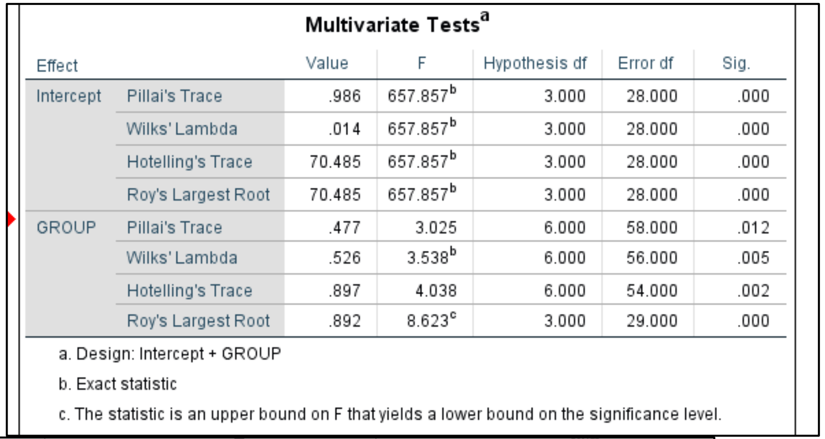


CLUSTERED DATA
_R_=25

Examples

SPSS

[51,52]



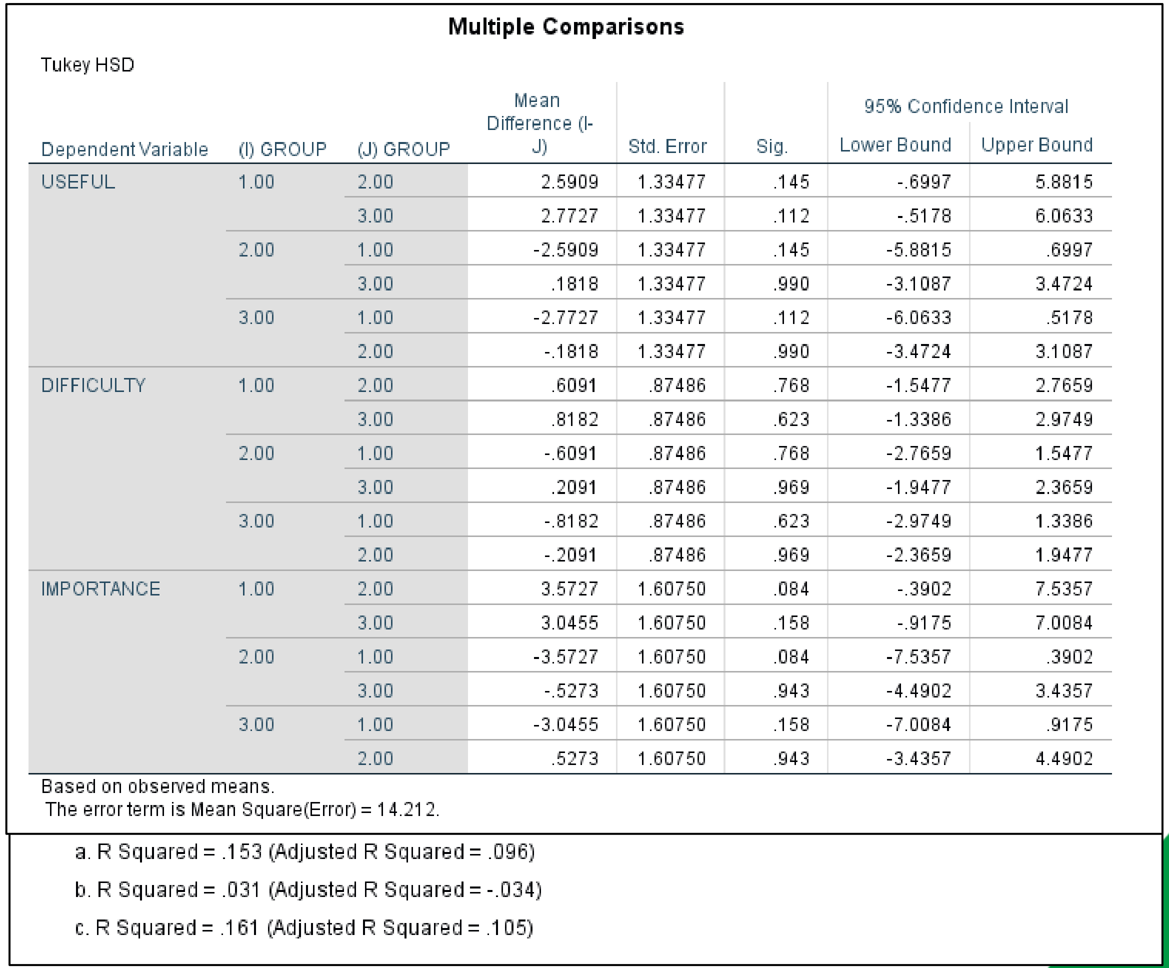
Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.986	657.857[b]	3.000	28.000	.000
	Wilks' Lambda	.014	657.857[b]	3.000	28.000	.000
	Hotelling's Trace	70.485	657.857[b]	3.000	28.000	.000
	Roy's Largest Root	70.485	657.857[b]	3.000	28.000	.000
GROUP	Pillai's Trace	.477	3.025	6.000	58.000	.012
	Wilks' Lambda	.526	3.538[b]	6.000	56.000	.005
	Hotelling's Trace	.897	4.038	6.000	54.000	.002
	Roy's Largest Root	.892	8.623[c]	3.000	29.000	.000

a. Design: Intercept + GROUP

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.



Multiple Comparisons

Tukey HSD

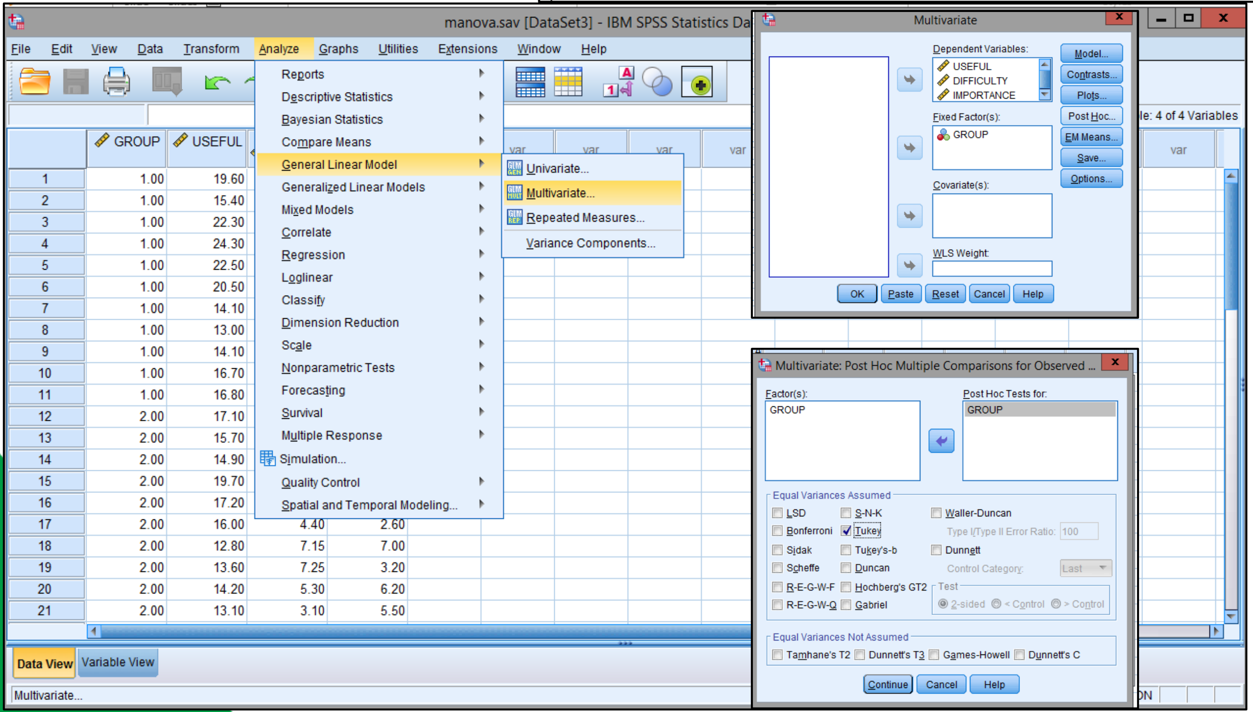
Dependent Variable	(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
USEFUL	1.00	2.00	2.5909	1.33477	.145	-.6997	5.8815
		3.00	2.7727	1.33477	.112	-.5178	6.0633
	2.00	1.00	-2.5909	1.33477	.145	-5.8815	.6997
		3.00	.1818	1.33477	.990	-3.1087	3.4724
	3.00	1.00	-2.7727	1.33477	.112	-6.0633	.5178
		2.00	-.1818	1.33477	.990	-3.4724	3.1087
DIFFICULTY	1.00	2.00	.6091	.87486	.768	-1.5477	2.7659
		3.00	.8182	.87486	.623	-1.3386	2.9749
	2.00	1.00	-.6091	.87486	.768	-2.7659	1.5477
		3.00	.2091	.87486	.969	-1.9477	2.3659
	3.00	1.00	-.8182	.87486	.623	-2.9749	1.3386
		2.00	-.2091	.87486	.969	-2.3659	1.9477
IMPORTANCE	1.00	2.00	3.5727	1.60750	.084	-.3902	7.5357
		3.00	3.0455	1.60750	.158	-.9175	7.0084
	2.00	1.00	-3.5727	1.60750	.084	-7.5357	.3902
		3.00	-.5273	1.60750	.943	-4.4902	3.4357
	3.00	1.00	-3.0455	1.60750	.158	-7.0084	.9175
		2.00	.5273	1.60750	.943	-3.4357	4.4902

Based on observed means.
The error term is Mean Square(Error) = 14.212.

a. R Squared = .153 (Adjusted R Squared = .096)

b. R Squared = .031 (Adjusted R Squared = -.034)

c. R Squared = .161 (Adjusted R Squared = .105)



	GROUP	USEFUL		
1	1.00	19.60		
2	1.00	15.40		
3	1.00	22.30		
4	1.00	24.30		
5	1.00	22.50		
6	1.00	20.50		
7	1.00	14.10		
8	1.00	13.00		
9	1.00	14.10		
10	1.00	16.70		
11	1.00	16.80		
12	2.00	17.10		
13	2.00	15.70		
14	2.00	14.90		
15	2.00	19.70		
16	2.00	17.20		
17	2.00	16.00	4.40	2.60
18	2.00	12.80	7.15	7.00
19	2.00	13.60	7.25	3.20
20	2.00	14.20	5.30	6.20
21	2.00	13.10	3.10	5.50

Multivariate

Dependent Variables: USEFUL, DIFFICULTY, IMPORTANCE

Fixed Factor(s): GROUP

Covariate(s):

WLS Weight:

Multivariate: Post Hoc Multiple Comparisons for Observed ...

Factor(s): GROUP

Post Hoc Tests for: GROUP

Equal Variances Assumed

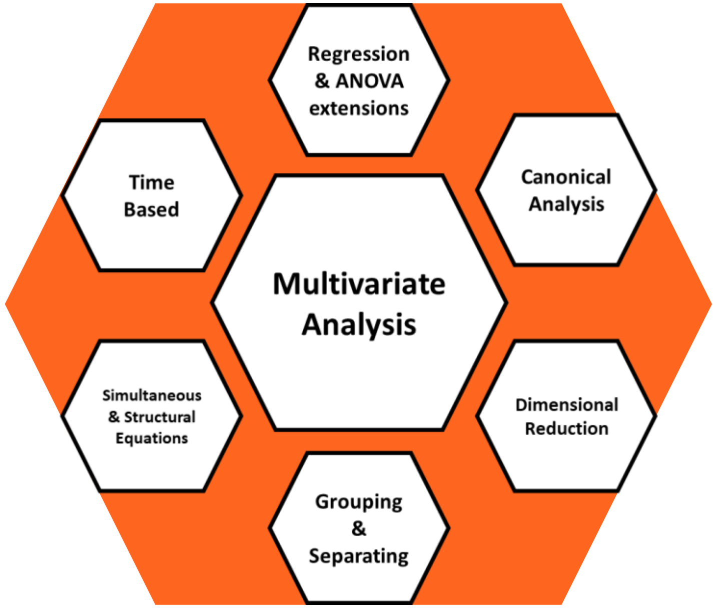
□ LSD □ S-N-K □ Waller-Duncan
□ Bonferroni ☑ Tukey Type I/Type II Error Ratio: 100
□ Sidak □ Tukey's-b □ Dunnett
□ Scheffe □ Duncan Control Category: Last
□ R-E-G-W-F □ Hochberg's GT2 Test
□ R-E-G-W-Q □ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C

Summary and Conclusion

- Multivariate analysis is a series of advanced methods that typically feature multiple predictor (Y) variables
- Features such as ordination, non-linearity, or repeated measures are also common
- Uses include data simplification, data exploration, and hypothesis testing
- Tune in next time for a more detailed look at multivariate analysis in Multivariate Analysis Module II: Leaves and Trees



Assessment & Acknowledgements

- Please take the 5-question assessment at:
  https://und.qualtrics.com/jfe/form/SV_eLhvOwQNlPAYNPE
- References cited in this presentation are available here:
  multivariate_analysis_module_1_refs
- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY