Old wine in new bottles : Factorial analyses in the age of multi-omics

Denis Laloë, Florence Jaffrezic, Tatiana Zerjal, Andrea Rau

Old wine in new bottles : Factorial analyses in the age of multi-omics

12-16 june 2023

Denis Laloë, F Jaffrézic, T Zerjal, A Rau
GABI, AgroParisTech, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France

ASPA Meeting, Monopoli, june 2023

25th Congress
ASPA2023
Monopoli (Bari, Italy), June 13-16, 2023

INRAE
la science pour la vie, l'humain, la terre

Introduction

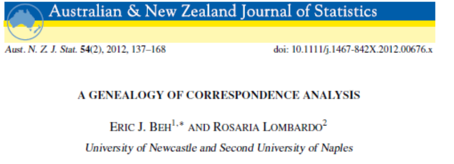
A bit of history : Back to the sixties

And even before (Pearson (1901), Hotelling (1933), Fisher (1940))

- 1962. Tukey : The future of analysis.
- 60's. Benzecri : Multivariate descriptive analysis tools.

Introduction

Further reading



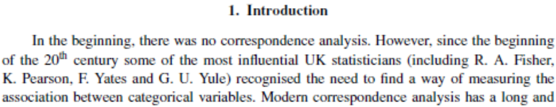
Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **54**(2), 2012, 137–168 doi: 10.1111/j.1467-842X.2012.00676.x

A GENEALOGY OF CORRESPONDENCE ANALYSIS

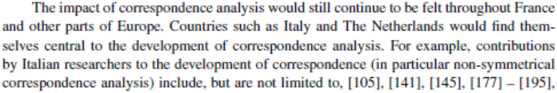
ERIC J. BEH[1,*] AND ROSARIA LOMBARDO[2]

University of Newcastle and Second University of Naples



1. Introduction

In the beginning, there was no correspondence analysis. However, since the beginning of the 20th century some of the most influential UK statisticians (including R. A. Fisher, K. Pearson, F. Yates and G. U. Yule) recognised the need to find a way of measuring the association between categorical variables. Modern correspondence analysis has a long and



The impact of correspondence analysis would still continue to be felt throughout France and other parts of Europe. Countries such as Italy and The Netherlands would find themselves central to the development of correspondence analysis. For example, contributions by Italian researchers to the development of correspondence (in particular non-symmetrical correspondence analysis) include, but are not limited to, [105], [141], [145], [177] – [195],

Data

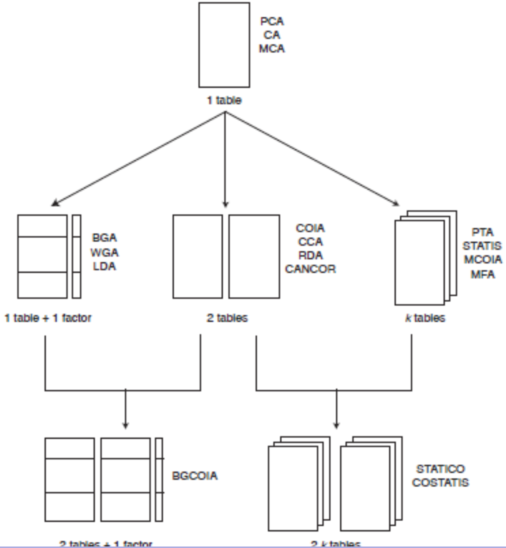
- *The model must be data-driven, not the opposite, J P Benzécri*
- *Let the data speak for themselves, J W Tukey*
- Emphasis on visualization

Factorial Analyses
Geometrical approach

- Data are represented by clouds of points in a multidimensional space
- Synthesis by dimensionality reduction
- Visualization by a factorial map / scatterplot
- One table
  - PCA, CA, MCA, MDS...

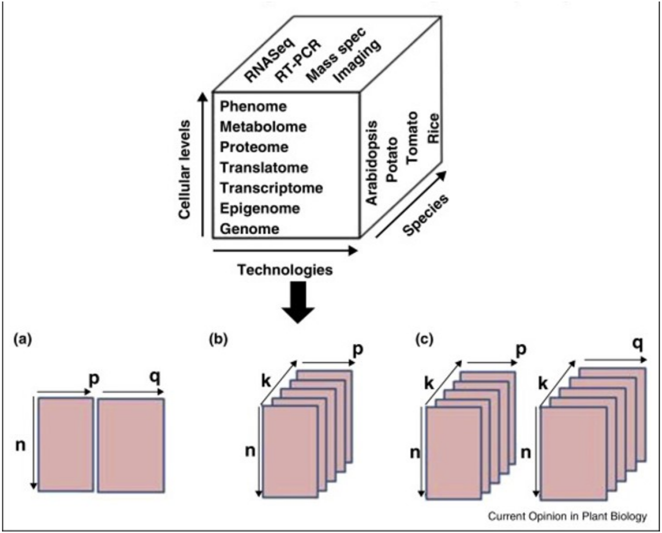
The extensions

Thioulouse, 2011. Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. Ann. Appl. Stat. 5 (4) 2300 - 2325



Big Data, x-omics

D Rajusundaram and J Selbig, 2016. More effort, more results: recent advances in integrative omics data analysis.



Data integration

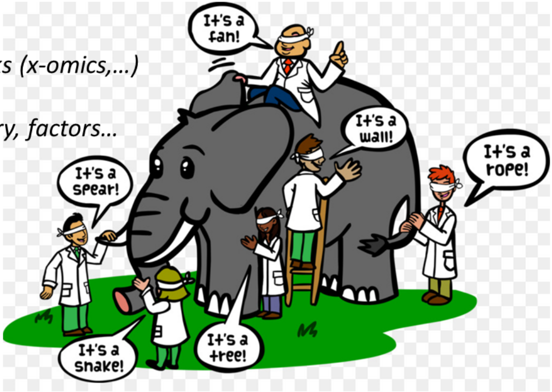
Different

- *variables in blocks (x-omics,...)*

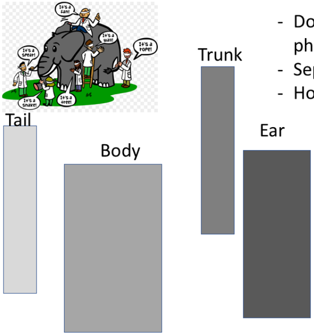
Heterogeneous

- *continuous, binary, factors...*

...



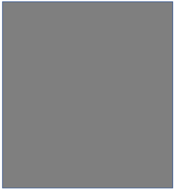
Data integration
No structure, different data



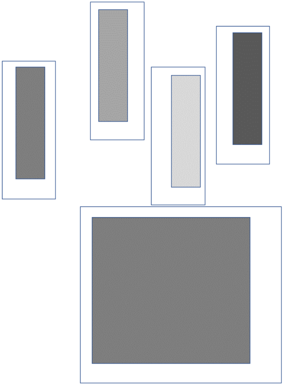
- Don't know it's an elephant (same phenomenon);
- Separate analyses-> different conclusions;
- How to combine conclusions?

Data integration
No structure, Everything is mixed

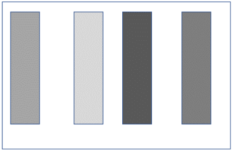
- Don’t know it’s an elephant (same phenomenon)
- But divided in different groups of data;
- Same analysis ? Lost specificities



Data integration
Structure - multitables

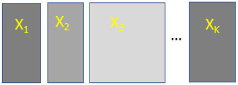


From k different analyses to a common analysis, that accounts for block specificities



K-tables analysis at a glance: Three steps

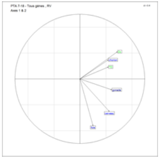
K tables



$RV_{(K,K)}$

$RV_{(K,K)}$: Similarities between tables

Interstructure

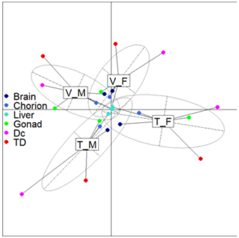


Compromise / Consensus table

$X=f(X_k)$

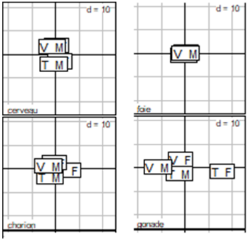
Intrastructure

Superposition of Compromise and partial plots

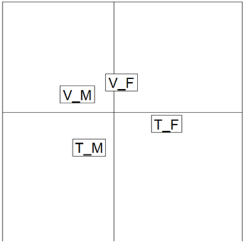


Projection of each X_K

Partial plots

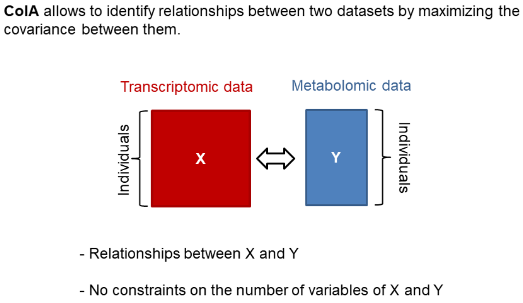


Compromise plot



Two Tables:
Coinertia and Redundancy Analysis

- Assymmetry: X explains Y: Redundancy Analysis (RDA)
- Symmetry: same role for X and Y : Coinertia Analysis (COIA)
- CoiA is to covariance what PCA is to variance



CoIA allows to identify relationships between two datasets by maximizing the covariance between them.

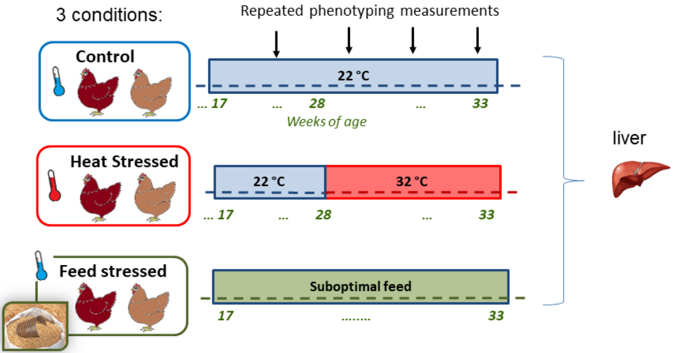
- Relationships between X and Y

- No constraints on the number of variables of X and Y

Coinertia : an example

An integrated metabolomic and transcriptomic analysis evaluating heat and feed stress in layer chickens (Zerjal and Laloë, 2019)

Chicken Model: 2 Rhode Island Red lines divergently selected for low (R-) and high (R+) residual feed intake (50 animals per line and condition).



Coinertia
Transcriptomic and metabolomics analysis

8 animals per line and condition

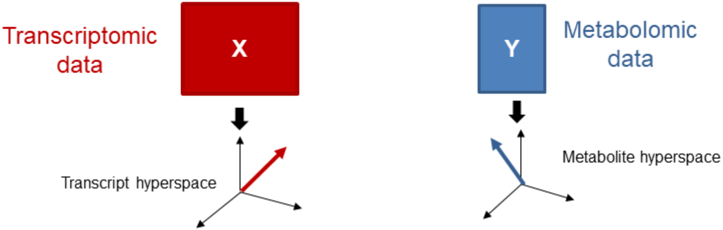
Total RNA

12873 expressed genes

Metabolome

142 NMR buckets

Results: A simultaneous representation



Transcriptomic data

X

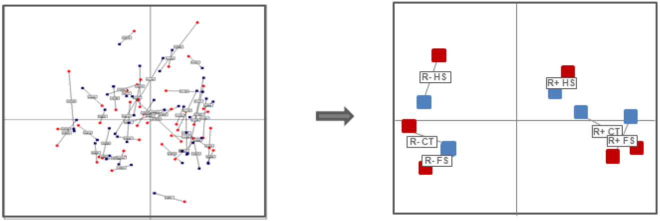
Y

Metabolomic data

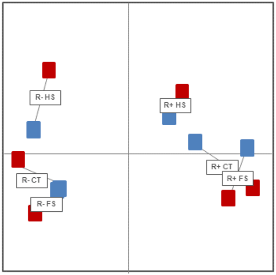
Transcript hyperspace

Metabolite hyperspace

Summarize the common structure existing in the two datasets



Results: A simultaneous representation



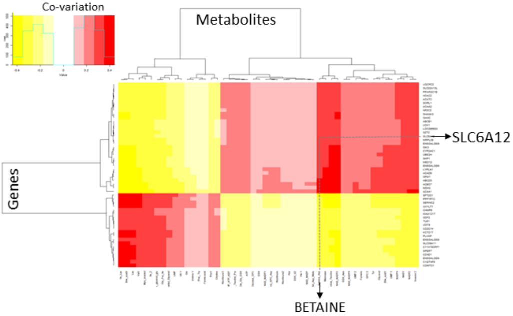
- The transcriptomics and the metabolomics data are impacted by both “line” and “stress” but the impact of the stresses is not of the same amplitude

- The heat stress seems to impact more the liver transcriptome and the metabolome than the feed stress

- The transcriptome seems to be more impacted than the metabolome

Results: Detection of covariant variables

Estimation of co-variation of related biological compounds under heat stress

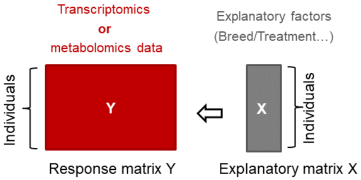


SLC6A12 encodes for a membrane transporter of betaine. In liver cells, SLC6A12 and betaine may have a role in volume regulation for cell survival under stress.

Redundancy Analysis

- Symmetry: same role for X and Y : Coinertia Analysis (COIA)
- Assymmetry: X explains Y: Redundancy Analysis (RDA)

RDA allows to summarize the variation in a set of response variables (Y) that can be explained by a set of explanatory variables (X).



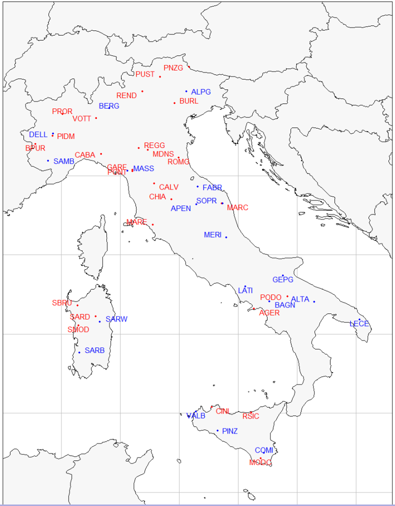
- Linear model approach applied to a multivariate context
- Constraints on the number of variables for the X matrix

Redundancy analysis

- Multivariate linear modelling
- Anova table
- Significance tests (permutation);
- Pourcentage of inertia due to
  - the model
  - the axes
  - the different factors in the model
  - Inertia % = Fst

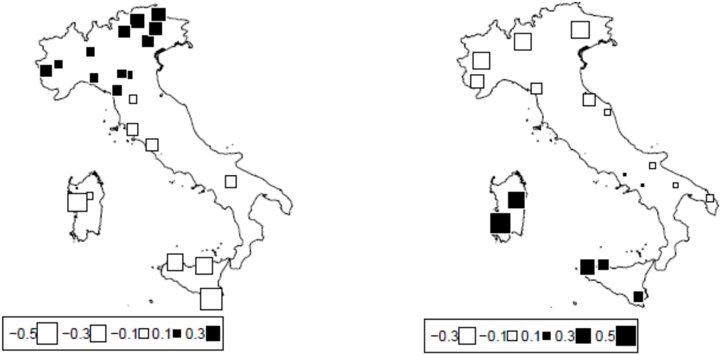
RDA and environmental genomics spatio-climatic structuration of Italian cattle and sheep

Senczuk et al, 2022. How Geography and Climate Shaped the Genomic Diversity of Italian Local Cattle and Sheep Breeds. Animals, 12(17), 2198.



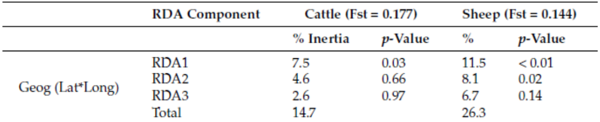
geographical structuration of Italian cattle and sheep

How cattle (left) and sheep (right) are structured according to geography (latitude+longitude)



geographical structuration of Italian cattle and sheep

Inertia components (geography)



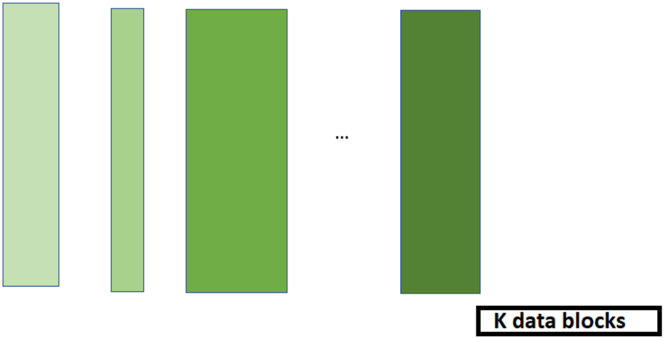
	RDA Component	Cattle (Fst = 0.177)		Sheep (Fst = 0.144)	
		% Inertia	*p*-Value	%	*p*-Value
Geog (Lat*Long)	RDA1	7.5	0.03	11.5	< 0.01
	RDA2	4.6	0.66	8.1	0.02
	RDA3	2.6	0.97	6.7	0.14
	Total	14.7		26.3	

- Cattle : 14,7 %
- Sheeps : 26,3 %

- Simplistic modelling of geography (Latitude:Longitude).
- Spatial PCA, neighbouring graphs. Laloë et al. 2010. Spatial trends of genetic variation of domestic ruminants in Europe. Diversity, 2(6), 932-945.

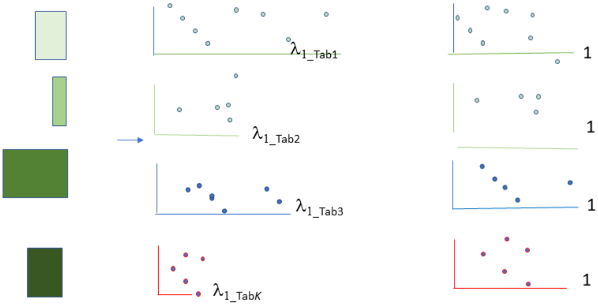
K-Tables Analysis
Multiple Factor Analysis (MFA, Escofier and Pagès, 1997)

Same observations, K data tables



MFA
a weighted PCA

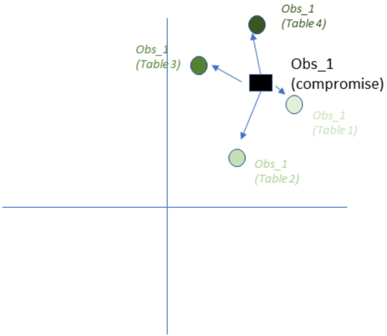
Ponderation by $1/\lambda_1$
Scale of the first axis
of each block PCA



MFA
A simultaneous representation of the compromise and the partial analyses

MFA :

- **PCA on the concatenation of weighted tables -> compromise**
- **Projection of tables on the compromise -> partial points**

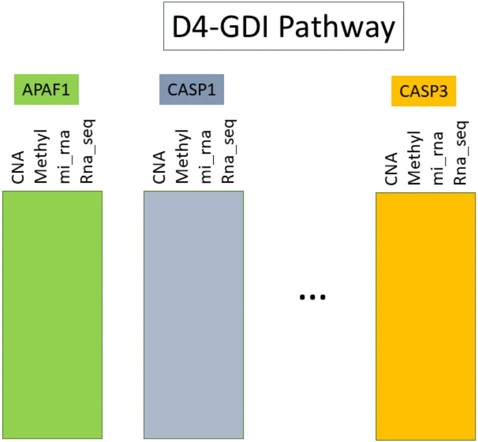


- Observations :
  - * A compromise point
  - * *K* partial points

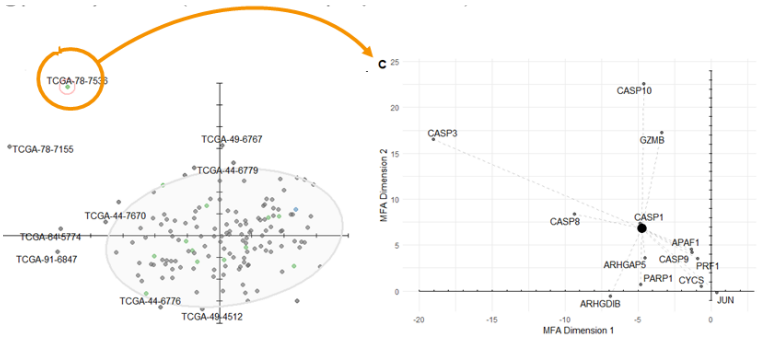
The compromise point : barycenter of the partial points cloud

Individualized multi-omic pathway deviation scores using multiple factor analysis. An example on TCGA data

Rau et al 2022. Individualized multi-omic pathway deviation scores using multiple factor analysis. Biostatistics, 23(2), 362-379.



Individualized multi-omic pathway deviation scores using multiple factor analysis. An example on TCGA data



A modern approach

- No matrix inversion step
    - Big Data / Computing
    - $p >> n$
- No (or a few) assumptions on distributions
- Focus on visualization

A modern approach

Versatility

- Mixing quantitative and qualitative
- Various designs
  - Two tables (Coinertia)
  - Sequences of tables (Multiple Factor Analysis)
  - Data cubes (Partial Triadic Analysis)
- Supervised version: modelling, prediction...
  - Two or three tables (Redundancy Analysis)
  - k- tables (Multiblock Redundancy Analysis)

References / R packages

- Beh, E. J., and Lombardo, R. (2012). A genealogy of correspondence analysis. Australian and New Zealand Journal of Statistics, 54(2), 137-168.
- Dray, S. and Dufour, A-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4).
- Lebart, L., Piron, M, Warwick, K. (1984) .Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices.*Wiley*
- Legendre P., Legendre L. (2012). Numerical ecology. *Elsevier*.

- ade4. *http://pbil.univ-lyon1.fr/ade4/*
- FactomineR *http://factominer.free.fr/*
- vegan *http://cran.r-project.org/web/packages/vegan/*