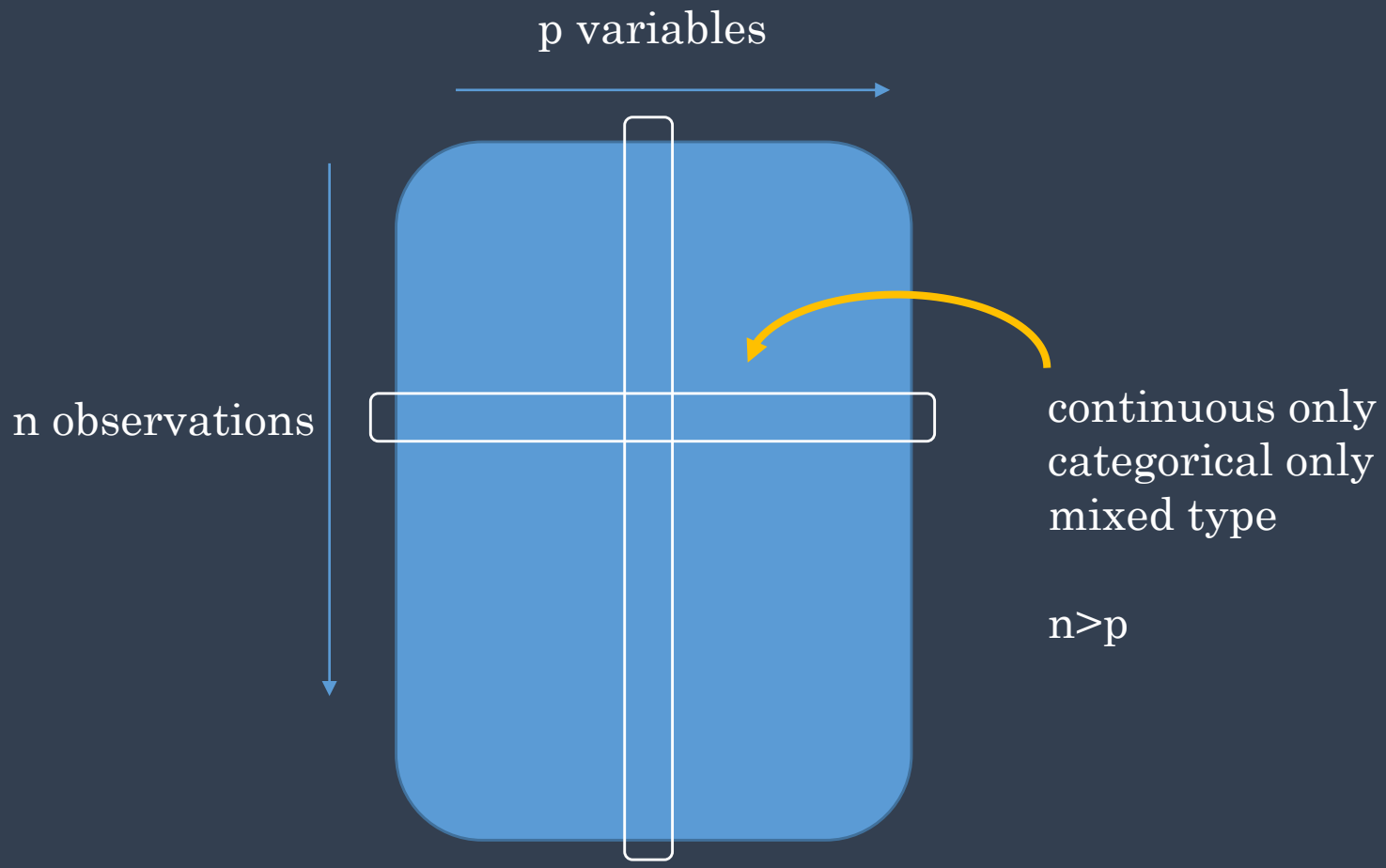
Multivariate stats (1)
1 table

Jérôme Mathieu

Sorbonne University, France

Club iEES, 2021

Type of data


p variables
n observations
continuous only
categorical only
mixed type
n>p

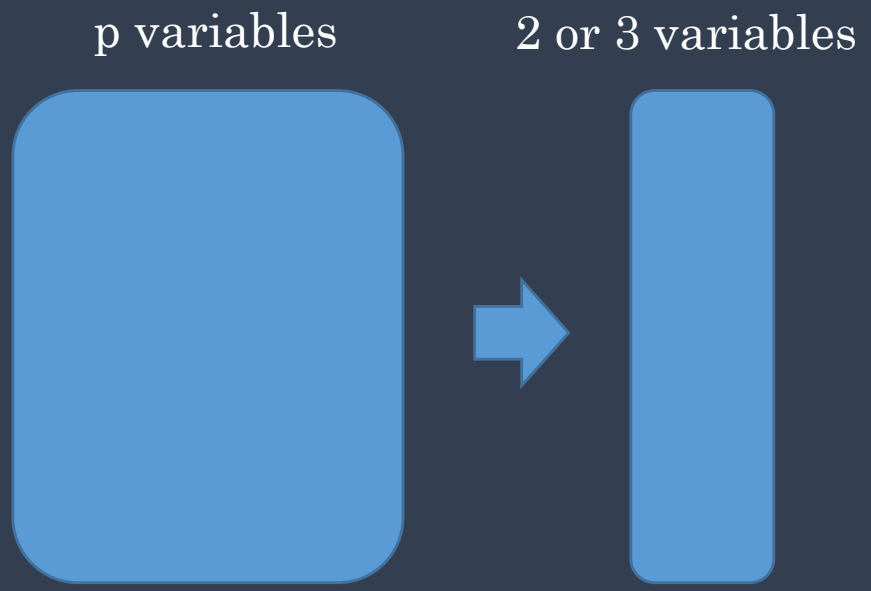
Typical questions

- Understand/vizualize the data
- Estimate the complexity of data
- Main structure of the observations: natural groups? outlyers? Gradient?
- Main relationships between variables: ~ correlation, group of variables?
- Link between the structure of the observations and the variables: which variables are associated to specific observations?
- ! Not a regression approach : no" y~x" (for this: see next presentation!)

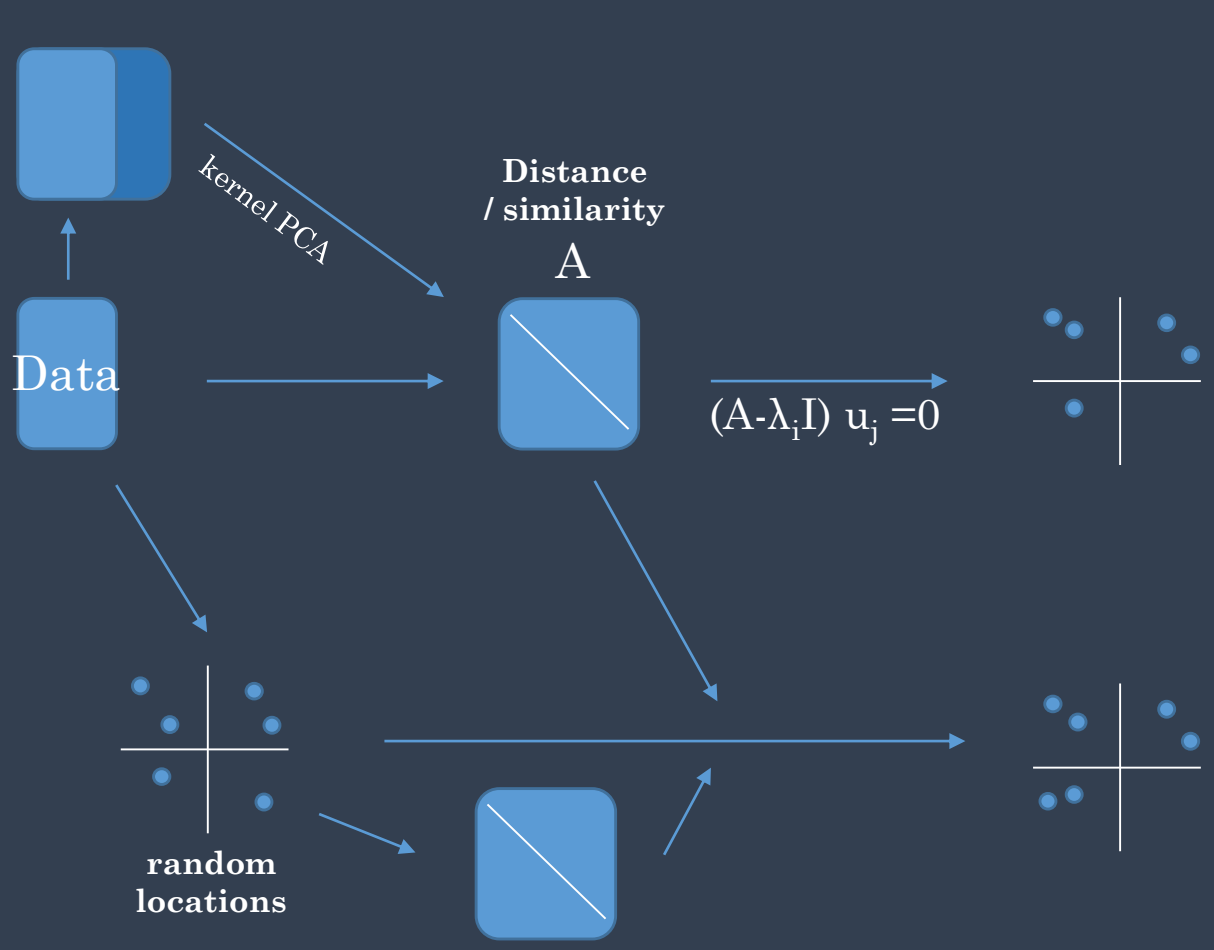
Questions

General principle : dimension reduction

- if p> 3 ➔ too complex for our brain! (except for Jacques G. ;-))
- n x p ➔ n x 2 or n x 3 (2D or 3D)
- keep maximum info



Typology of approaches



PCA
CA
pCoA
MCA
HS
MFDA
kernel PCA

Algebric
maximize variance
axes = combinations of p

NMDS (ranks)
t-SNE

Iterative
Maximize similarities
axes = arbitrary space

Benchmark

Analysis	data	distance	typical use	pitfalls	solution	function	package
PCA	any quantitative	euclidian	most used, short gradient	double 0, variable heterogeneity	transformation, scaling	dudi.pca	ade4
CA	positive number, homogenous	chi 2	abundance, long gradient	gutman effect, sensitive to rare species	detrending?	dudi.coa	ade4
pCoA	any quantitative	metric	ecological communities	non metric distances	transformation	dudi.pco	ade4
NMDS	any quantitative	non metric	ecological communities	right metric, instability	repeat	metaMDS	vegan
kernel PCA	any quantitative	euclidian	complex data			kpca	kernlab
t-sne	any quantitative	euclidian	complex data			Rtsne	Rtsne
MCA	qualitative	euclidian		insufficent data in a category		dudi.acm	ade4
HS	quantitative + qualitative	euclidian	mixed data			dudi.hillsmith	ade4
MFDA	quantitative + qualitative	euclidian	mixed data			dudi.mix	ade4

! in PCA, CA, pCoA, close points ≠ similar points