Statistical learning - Unsupervised learning

Rita Almeida

PhD course 3045, VT 2018

Unsupervised learning

Unsupervised learning is a set of tools to explore possible structure in a set of n observation of p variables.

- There is no outcome variable - no predictions.
- There is no way to check how well the algorithm works.
- Often part of an exploratory data analysis.
  - Aiming at visualizing the data in informing ways - PCA.
  - Aiming at finding subgroups in the data - clustering.

Principal component analysis

Problem: One has a set of n observation of p variables $X_1, X_2, \dots, X_p$. How to visualize / understand the data?

- Plot all combinations of 2 or 3 dimensional plots. Unfeasible for large p!
- Find a low-representation of the data that captures as much as the information as possible.

Principal component analysis (PCA):

- Finds a set of dimensions that are as interesting as possible.
- Interesting: a large variation of the data is on that dimension.

Principal components

Example in 2 dimensions:

- 1st principal component - direction through the data that explains the most variance
- 2nd principal component - direction orthogonal to that of the 1st component, explaining the next greatest amount of variance

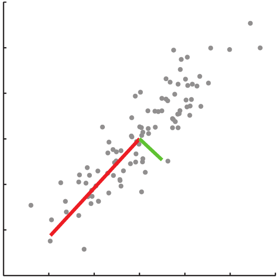


Figure adapted from Poldrack et al. 2011

Each component is a linear combination of the original variables.
First principal component:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p, \quad \text{with} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

ϕs are the loadings of the first principal component.

Visualizing the data

- 50 observations of 4 variables
- Blue dots: Scores of the observations in the first two components.
- Orange arrows: The first two principal components vectors of loadings.

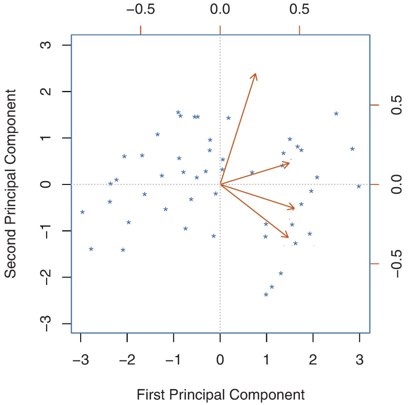


Figure adapted from Jones et al. 2013

Further remarks

- Variables should be centered and scaled before PCA.
- How many principal components shall one retain?
  - Proportion of explained variance by the first components
  - Scree plots

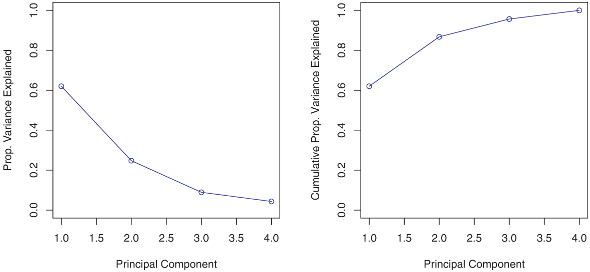


Figure adapted from Jones et al. 2013

- One chooses the smallest amount of components to retain a large amount of the total variance.

Neuropsychologia 49 (2011) 1518–1526

Contents lists available at ScienceDirect

Neuropsychologia

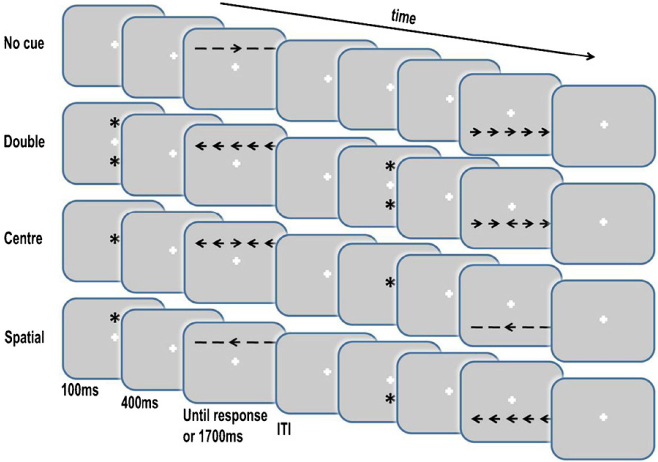
journal homepage: www.elsevier.com/locate/neuropsychologia

Principal component analysis of behavioural individual differences suggests that particular aspects of visual working memory may relate to specific aspects of attention

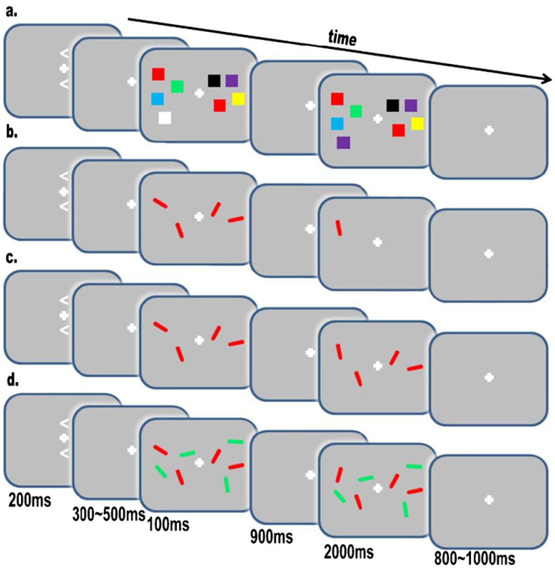
Maro G. Machizawa [a,b,*], Jon Driver [a,b,c]

[a] UCL Institute of Cognitive Neuroscience, University College London, United Kingdom
[b] UCL Institute of Neurology, University College London, United Kingdom
[c] Wellcome Trust Centre for Neuroimaging at UCL, University College London, United Kingdom

Example - attention tasks



Example - working-memory tasks



Example - results

- 50 subjects
- 3 measures of attention: alerting score, orienting score, conflict score
- 3 measures of visual working-memory: capacity, precision, filtering efficiency

	Components		
	1	2	3
WM capacity	**.76**	.08	.17
ANT alerting	**.70**	−.06	−.10
WM precision	.34	**.80**	.17
ANT orienting	.31	−**.82**	.16
WM filtering	.20	−.04	**.74**
ANT executive	−.14	.04	**.82**

Clustering - K-means clustering

Aim: separating the observation into K clusters.

- Each observation will be assigned to one and only one cluster - no overlapping.
- Clusters are determined so that the within cluster variation is as small as possible.
- Variation is defined using a distance - usually euclidean distance.
- The number K of clusters is pre-specified.

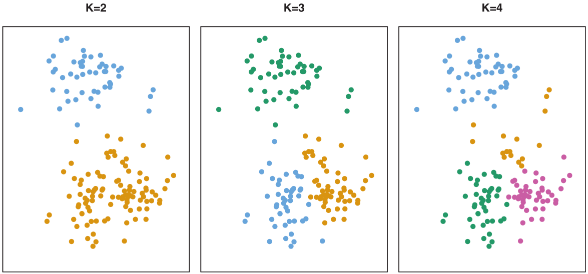
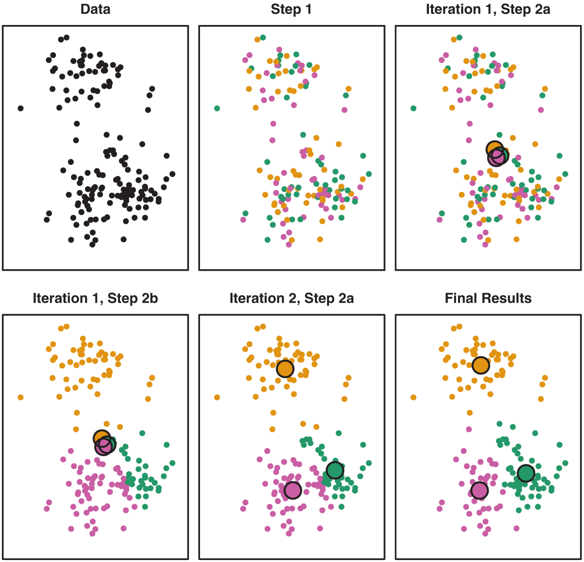


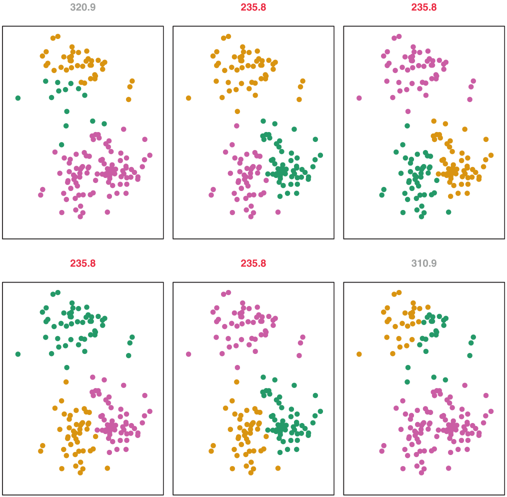
Figure adapted from Jones et al. 2013

K-means clustering - one algorithm illustration



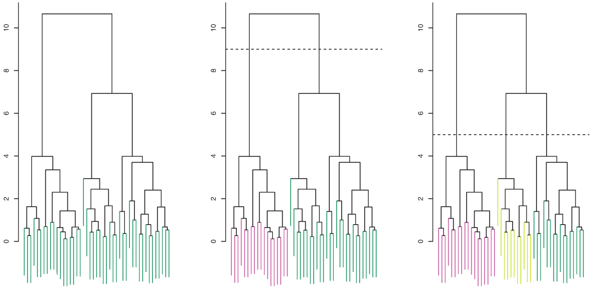
Local optimum

- The final solution is a local and not a global optimum.
- One runs several times and selects the best solution.



Hierarchical clustering

- Agglomerative clustering: clusters are fused, a pair at a time, based on similarity.
- The resulting dendogram can be cut at different levels.
- A measure of distance is defined - usually euclidean distance.
- A measure of cluster similarity is defined. For example: complete linkage or average linkage.



Hierarchical clustering