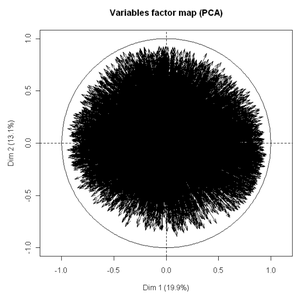
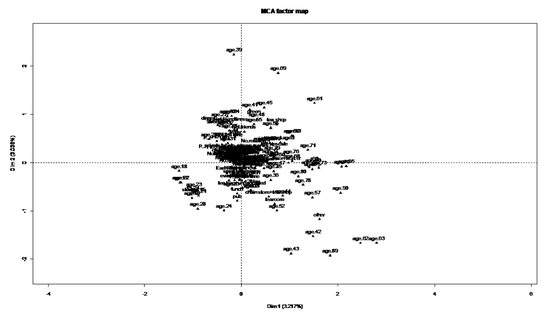
Visualization of High-Dimensional Data in an Exploratory Multivariate Framework

BUCHE Marianne[1], CARRE Yohann[1], LE Sébastien[1]
[1] AGROCAMPUS OUEST – Laboratoire de Mathématiques Appliquées, Rennes
Contact author : sebastien.le@agrocampus-ouest.fr

- Data collected nowadays: higher and higher dimensions (biotechnology, picture or spectrum analysis, surveys ...)
- We will focus on the visualization of quantitative and qualitative variables.
- In this exploratory context, the reference methods are PCA and MCA.
- 2 main problems for the representation of objects in high dimension :
  - Display of numerous objects
  - Selection of information to be displayed

PCA of an example of dataset for the quantitative case.

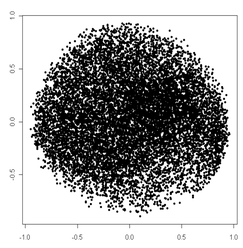




MCA of an example of dataset for the qualitative case.

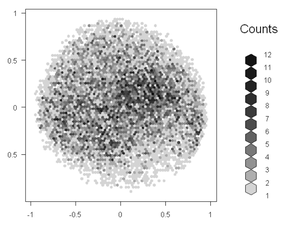
High-dimensional data : How to represent variables ?

Points approach



Points instead of arrows
Plot {graphics}

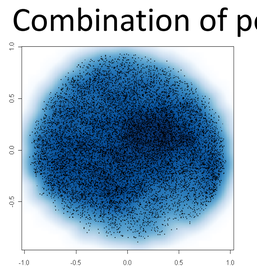
Density approach



Density hexagons
hexbin {hexbin}

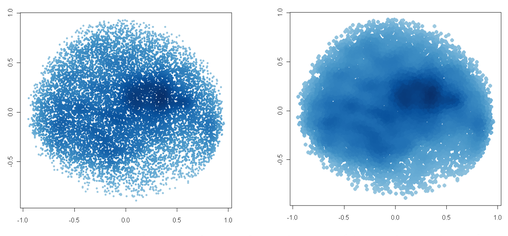
Advantages and drawbacks:
- Points: conservation of real positions of variables
- Density: practical, but the hexagons only represent space divisions

Combination of points and density approaches:

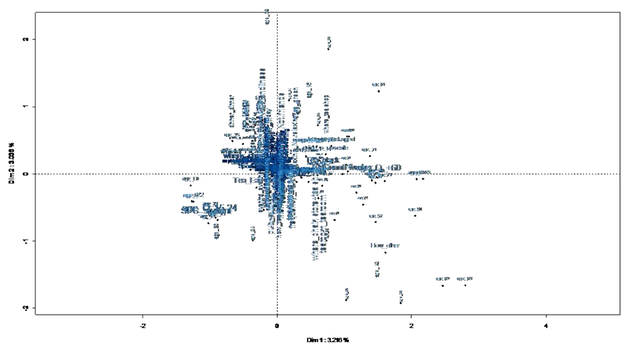


Density map with points
smoothScatter {RcolorBrewer}

Points with density colors:
- The points correspond to the real position of the variables on the graph
- Color and transparency give a faithful idea of the variable distribution



Density-colored points (2 point sizes)
densCol {grDevices}

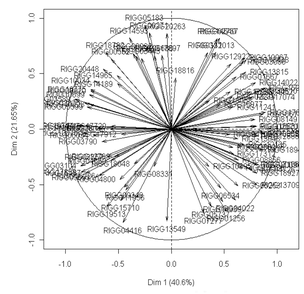


Idea: add information on the MCA graph
- The categories are shown darker according to the density of category around, which enable to identify quickly the proximity between a lot of categories.
- The bigger the label is displayed, the more the test value for the category is important.
- Categories are oriented according to the axis to which they contribute the most.

Even after an improvement of the graphical representation, it is still very difficult to analyze the results: Hence the idea of methods to select the most relevant variables

Hierarchical clustering and representation of the best projected variables

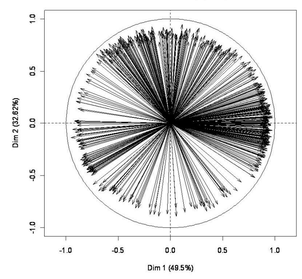
Method:
- Clustering on the variables using (1-r) as a distance criterion
- PCA with the best projected variables of each of the *n* groups
- Conservation of the initial structure: removal of variables correlated to the 1st or 2nd dimension to keep the initial λ_1/λ_2 ratio



- Efficiently summarize the information of the whole dataset
- Expensive calculation: the correlation matrix

Comparison with a randomly-generated dataset and selection of the best informative variables

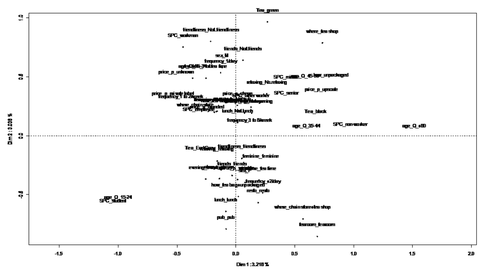
Method:
- Projection threshold based on simulations of random data: conservation of the real variables over this confidence threshold of information
- PCA with the selected variables
- Conservation of the initial structure: λ_1/λ_2 ratio (see previous method)



- Criterion to detect informative versus non-informative variables
- Number of selected variables: not chosen

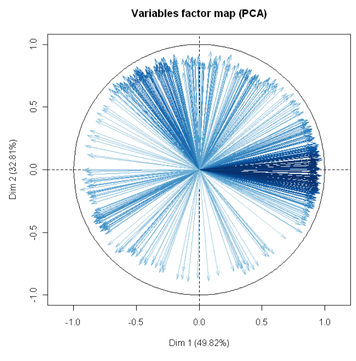
Selection of variables with a test on the categories

Method:
Only the variables which have at least one significant category are represented.

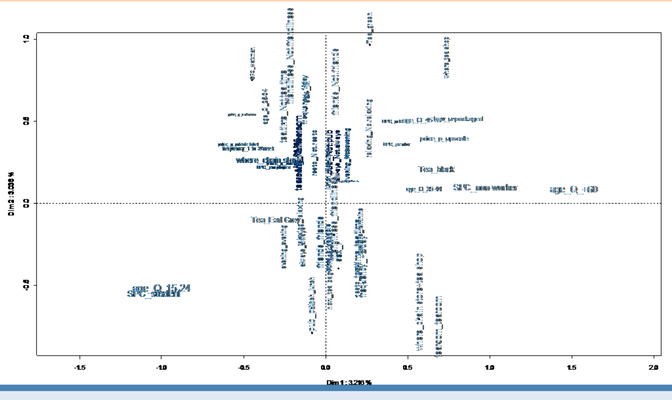


- Suppression of variables with no significant category

Representation of the selection of variables summarizing the information



Representation of the selected variables only. Colors according to the density of variables.



Conclusion:

- Efficient methods to select and graphically visualize quantitative and qualitative high-dimensional variables.
- High-dimensional spaces have properties making it difficult to select informative variables : there is a hardly separable noise/information mingling.