Visualizing Latent Structures in Grade Correspondence Cluster Analysis and Generalized Association Plots

Wiesław Szczesny
Department of Econometrics and Informatics, Warsaw Agricultural University
Marek Wiech
Institute of Computer Science, Polish Academy of Sciences

(study partially sponsored from a grant no. 3 T11C 053 28, awarded by the MNiI)

Contents

Introduction

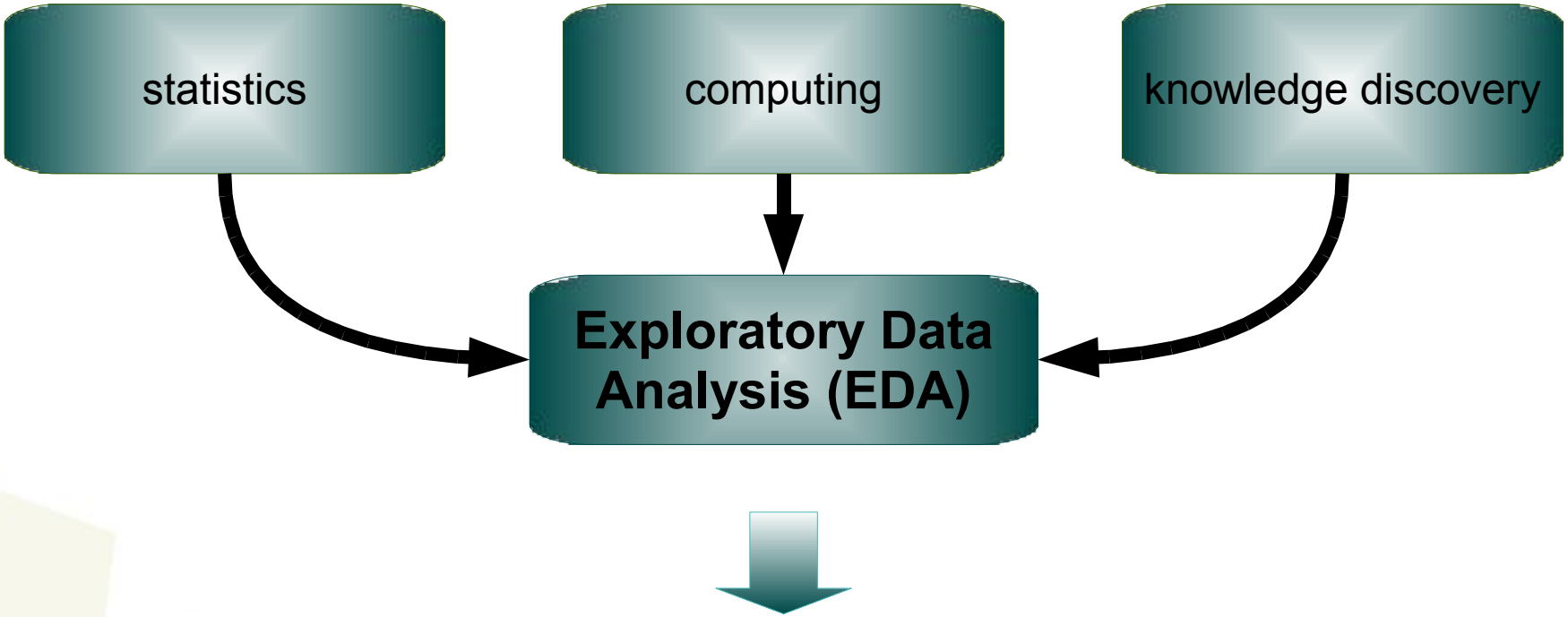
The main goal of the article is to compare two recent exploratory methods:

- Generalized Association Plots (GAP)
- Grade Correspondence Cluster Analysis (GCCA)

The comparison is made on:

- two types of highly regular artificial data sets of the same size (150x10) as the empirical data set
- empirical psychological data set, concerning belief in superstitions and some temperamental traits

The main aim of EDA



Revealing and visualizing
the latent structure of multivariate data sets

Short introduction to both methods

Here two methods of Exploratory Data Analysis (EDA) are going to be compared:

- Generalized Association Plots
- Institute of Statistical Science, Academia Sinica (Taiwan)
- software: GAP (not available during article preparation)
- http://gap.stat.sinica.edu.tw

- Grade Cluster Correspondence Analysis
- Institute for Computer Sciences, Polish Academy of Sciences
- software: GradeStat (available since 2004)
- http://gradestat.ipipan.waw.pl

Contents

Data description

Two theoretical data sets

- artificial data
- highly regular
- to make GAP and grade data analysis familiar

Empirical (experimental) data set

- real experimental data
- not so regular
- to reveal real hidden structures in data

Theoretical data sets

First Artificial Data:

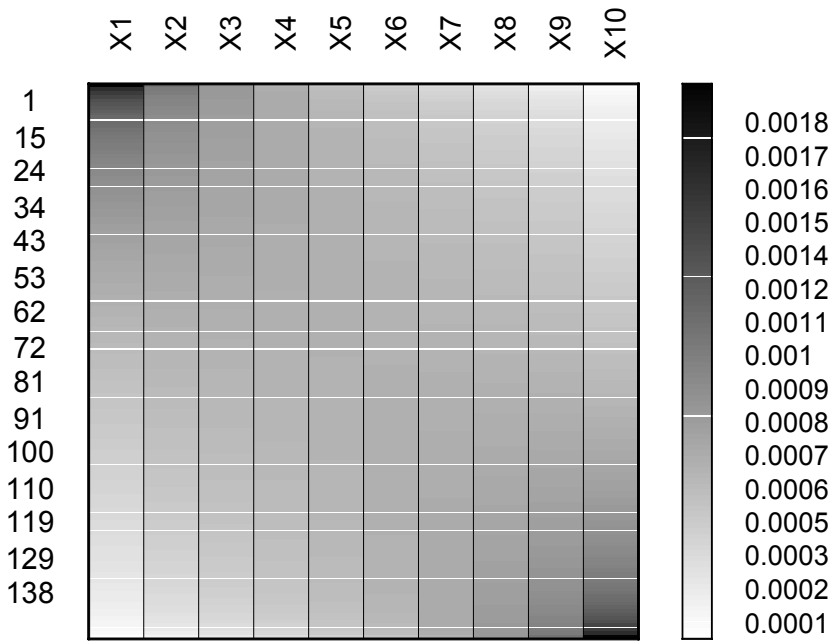
- table 150 rows x 10 columns
- by discretization and aggregation of the distribution of ($\Phi(X)$, $\Phi(Y)$), where:

Φ = cdf of normal distribution N(0,1)
(X, Y) = standard binormal pair: zero means, unit variances, correlation coefficient = 0.26 (as in psychological data)

- posterior random reordering of rows and columns, to „hide” latent structure of the data set

Theoretical data sets

First Artificial Data (before random reordering):



Such data:

- has highly regular positive dependence
- is only slightly disturbed by discretization

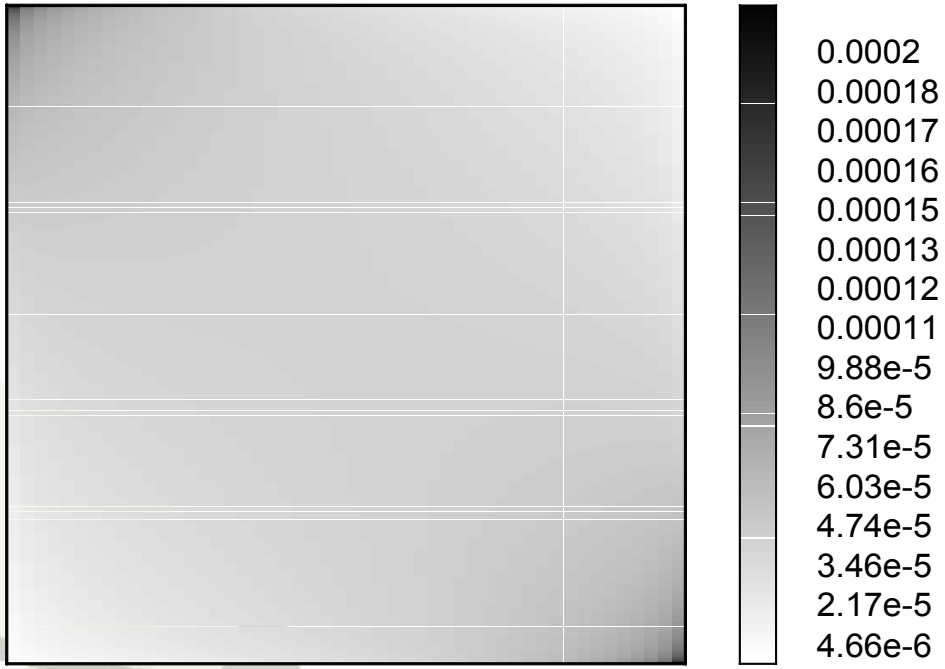
Theoretical data sets

Second Artificial Data:

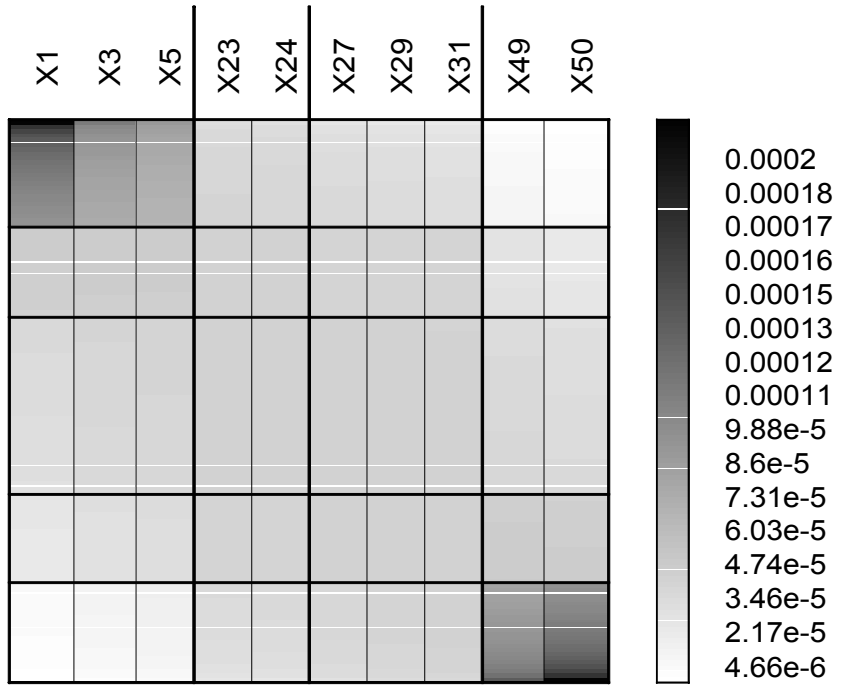
- uniform discretization of ($\Phi(X)$, $\Phi(Y)$)
- thus obtaining the data table 500 x 50
- cutting out the subtable 150 x 10, with rows forming 5 clear clusters and columns selected irregularly, however forming 3 clusters
- rows and columns of that subtable reordered randomly, analogically as in the First Artificial Data set

Theoretical data sets

Second Artificial Data (500x50, before cutting off sub-table):



Second Artificial Data (150x10, after cutting, but before reordering):



Psychological data set

Creation:

- experimental data concerning superstition
- 150 observations (persons)
- 10 variables – questionnaires' results:
  - *kop20* (belief in superstitions scale)
  - *dyrekt15* (directiveness)
  - *żw*, *pe*, *ws*, *re*, *wt*, *ak* (temperamental traits)
  - *staix2* (anxiety trait)
  - *rwa* (the right-wing authoritarianism scale)
- Final data table: 150 rows and 10 columns

Psychological data set

Properties:

- each person gained individual score on each scale
- result of the scale is the sum of points
- results measured on ordinal scales
- each result was normalized to a number in the unit interval
- primary data was ordered as inserted into data set, in effect, most probably - random

Contents

GAP and GCCA comparison

Aims:

- to restore the original order of rows and columns
- to indicate existing clusters

Assumptions and questions for:

First Artificial Data set – whether for any chosen number of rows and columns clusters of the resulting aggregated table would be roughly the same as the direct uniform discretization of the distribution ($\Phi(X)$, $\Phi(Y)$)

Second Artificial Data set – whether the clustering would lead to 5 clusters of rows and 4 of columns

Generalized Association Plots (GAP):

- Iterated sequences of correlation matrices are studied, starting from the initial proximity matrix
- Every matrix is projected onto the plane spanned by the first two eigenvectors
- Thus clear elliptical clusters begin to form at one step; the final number of clusters is chosen by an analyst
- Raw data should be measured on the same scale, and proximity matrices should be transformed to [-1,1]

Chen (2002) - example of GAP

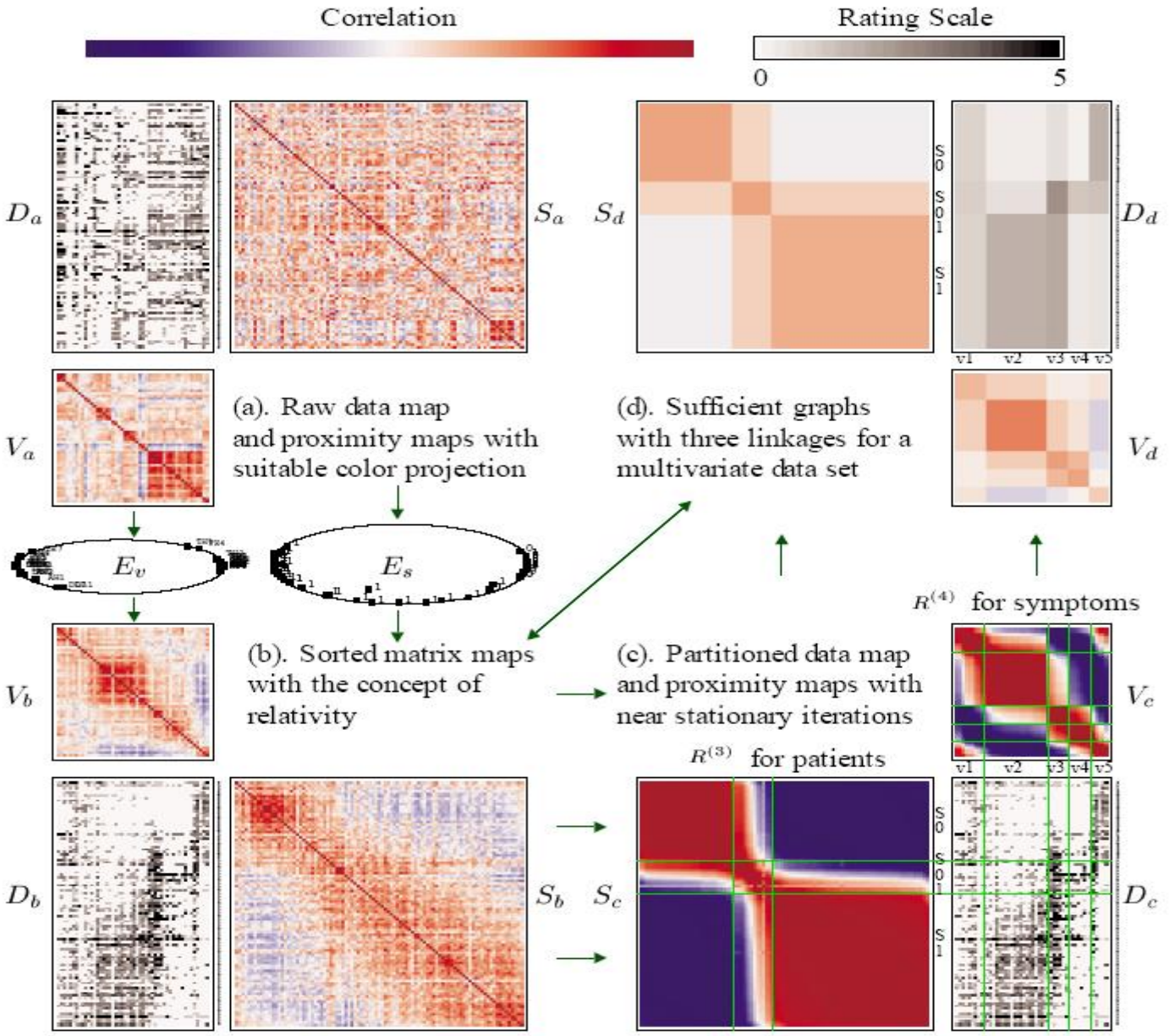


Figure 10. Complete GAP procedure for the psychosis disorder data set with ninety-five patients and fifty symptoms.

Chen (2002) - example of GAP

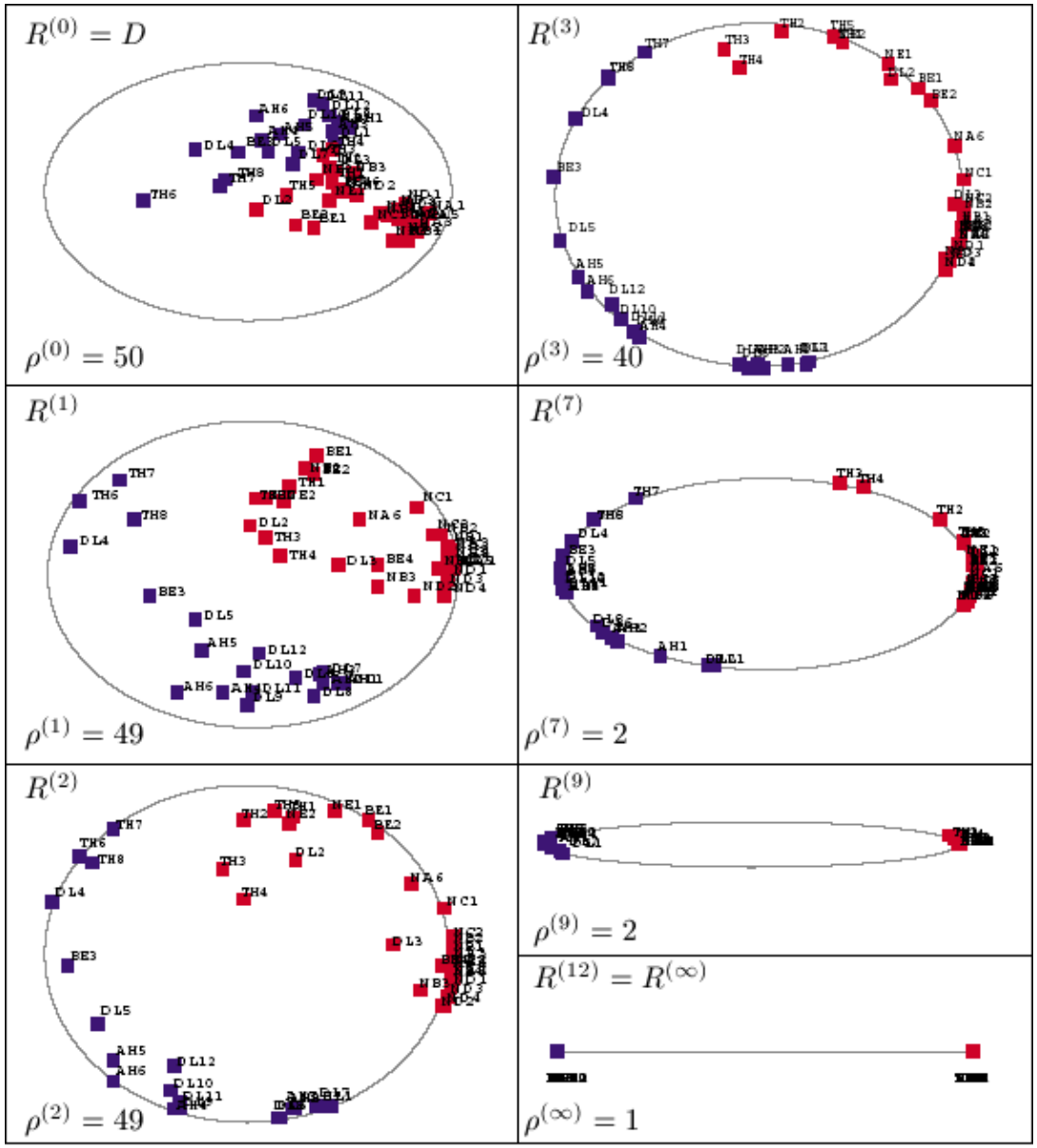
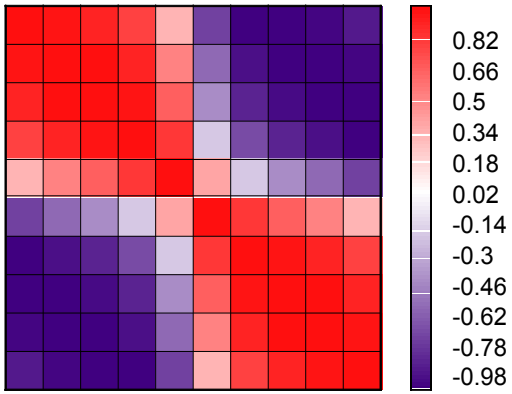
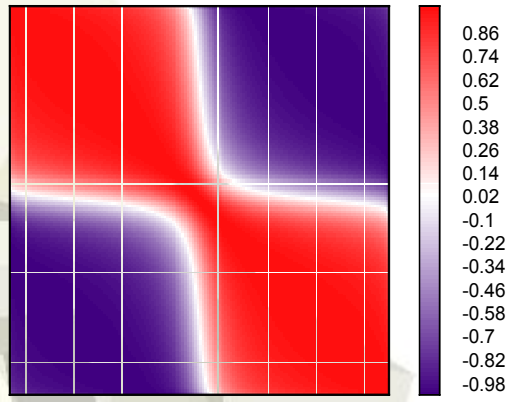


Figure 1. Plots for first two eigenvectors for selected correlation matrices in the converging sequence. ($\rho(n)$ is the rank of $R^{(n)}$).

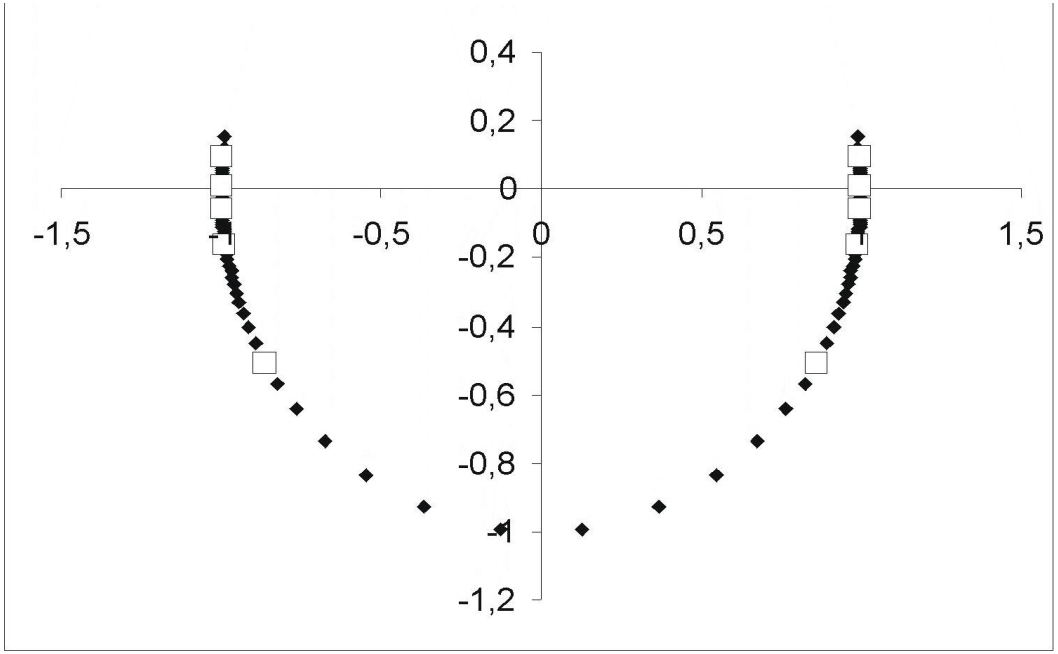
First Artificial Data - GAP

- Maps of Pearson correlations - for columns (upper map) and rows (bottom map)





- Plots for the first two eigenvectors in case of rows – black points, and columns – white points; clear ellipses formed at the first step, correct ordering



Grade Correspondence Cluster Analysis consists of two procedures:

1. GCA, simultaneously reordering rows and columns to achieve *the best* ordering;

2. Posterior clustering of adjacent rows and adjacent clusters into disjoint clusters; number of clusters is chosen by an analyst

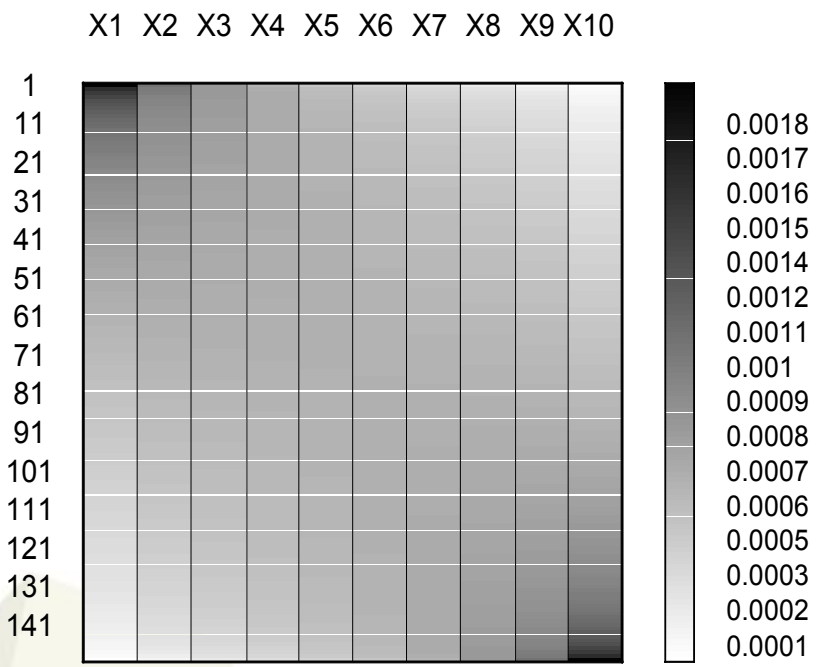
- The best ordering is where the grade correlation ρ^* (between the so called latent column variable and the latent row variable) is maximal

ρ^* (a sum of concentration indices for all pairs of rows/columns weighted by the distance between rows/columns) is a measure of dissimilarity, VERY sensitive on the orderings

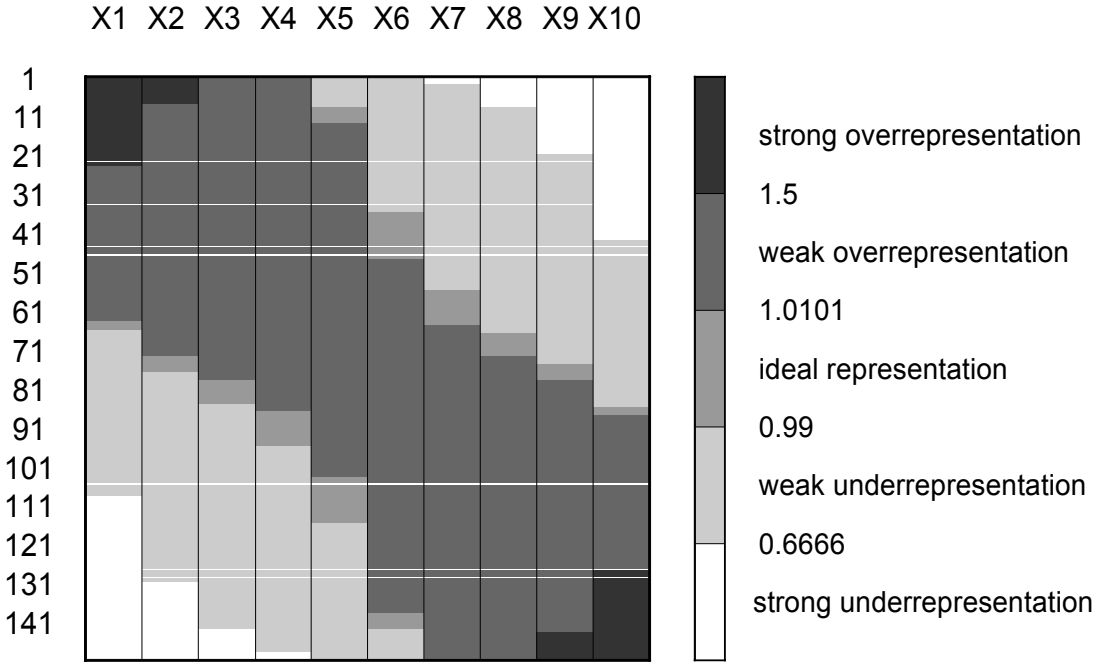
- The post-GCA clustering maximizes the grade correlation for aggregated table, for any chosen number of row (or column) clusters

overrepresentation - GCCA

Raw data map

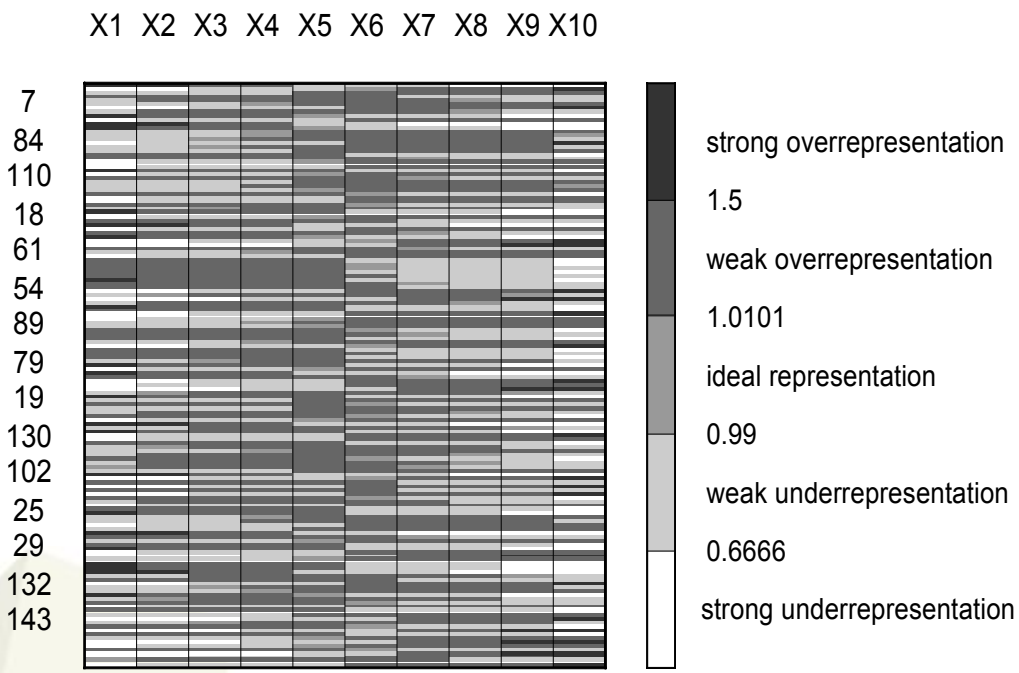


Overrepresentation map



An overrepresentation map **is** a raw data map
but applied to transformed data: the number at the intersection of a row and a column is divided by the product of the row and column totals

overrepresentation - GCCA

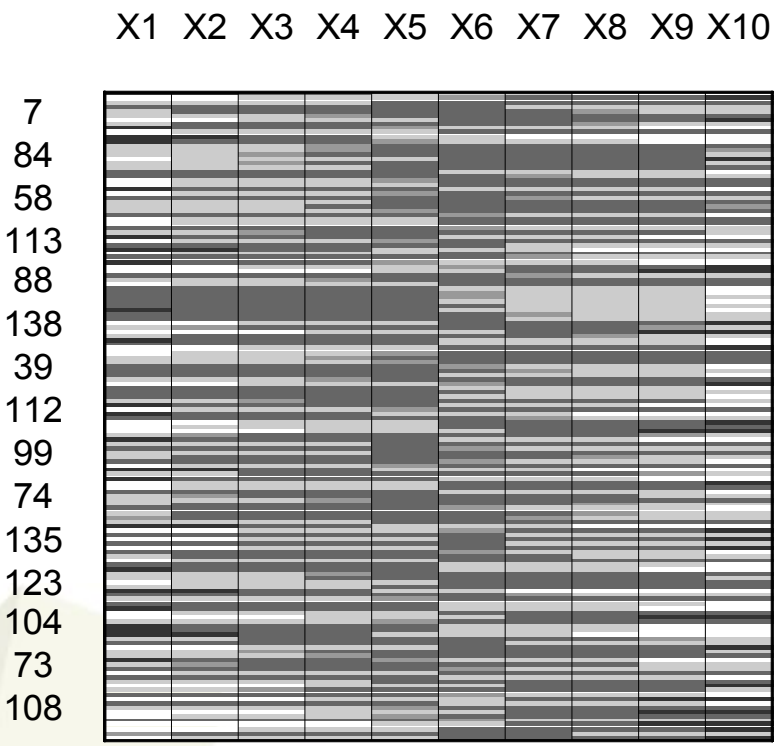


- Index above 1 (dark grey to black) – result is higher than „fair" representation – overrepresentation
- Index close to 1(grey) – result roughly as expected for this row and column - „fair" representation
- Index below 1 (white to light grey) – result lower than expected – underrepresentation

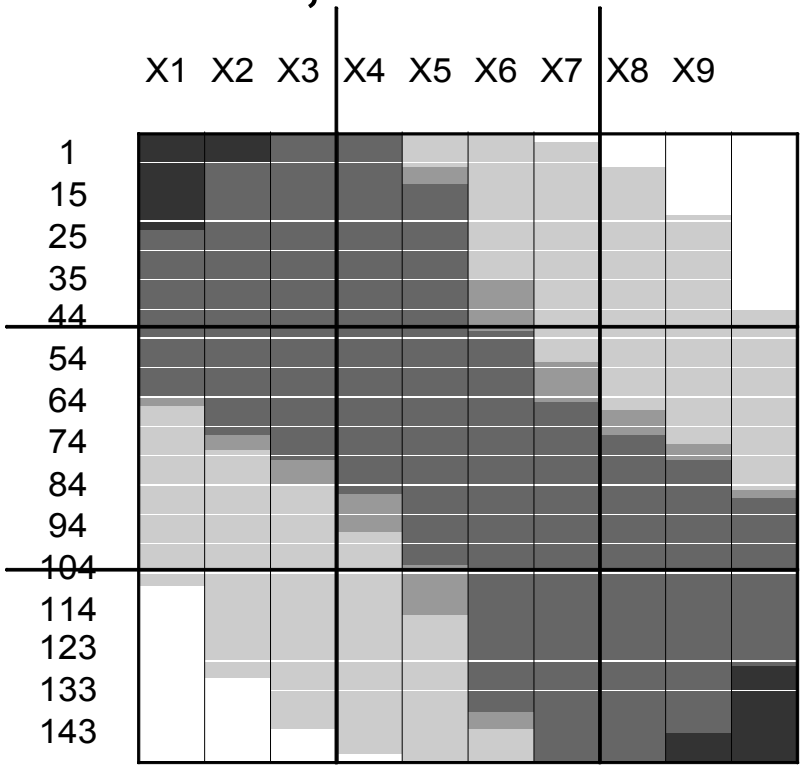
The matrix is not ordered. We cannot see any clear data structure. It is chaotic. **The main aim:** to reorder data to discover the hidden structure

First Artificial Data - GCCA

Original ordering

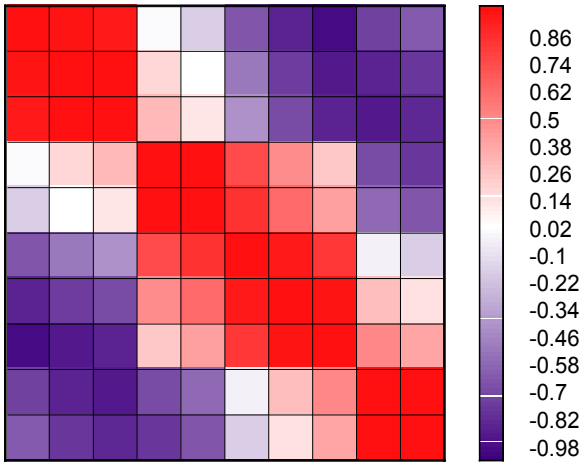


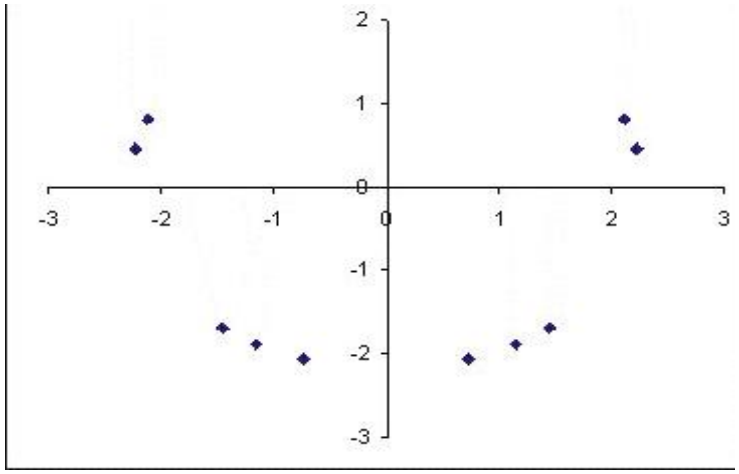
Reordered by GCA: 3 clusters for columns and 3 for rows, correct ordering



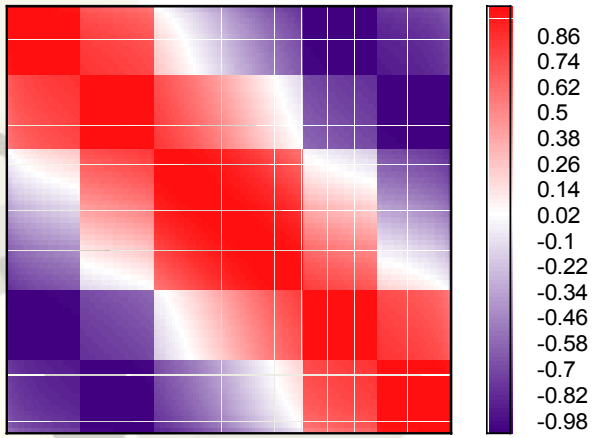
Second Artificial Data - GAP

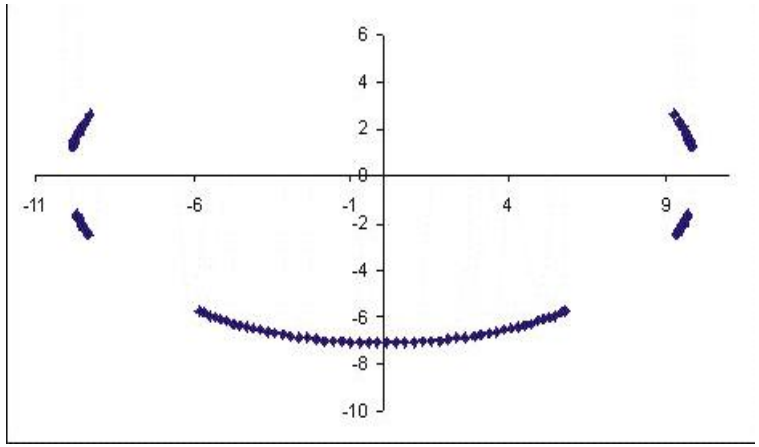
- *Columns*: map of Pearson correlations (left) and plots for the first two eigenvectors (right); correct ordering





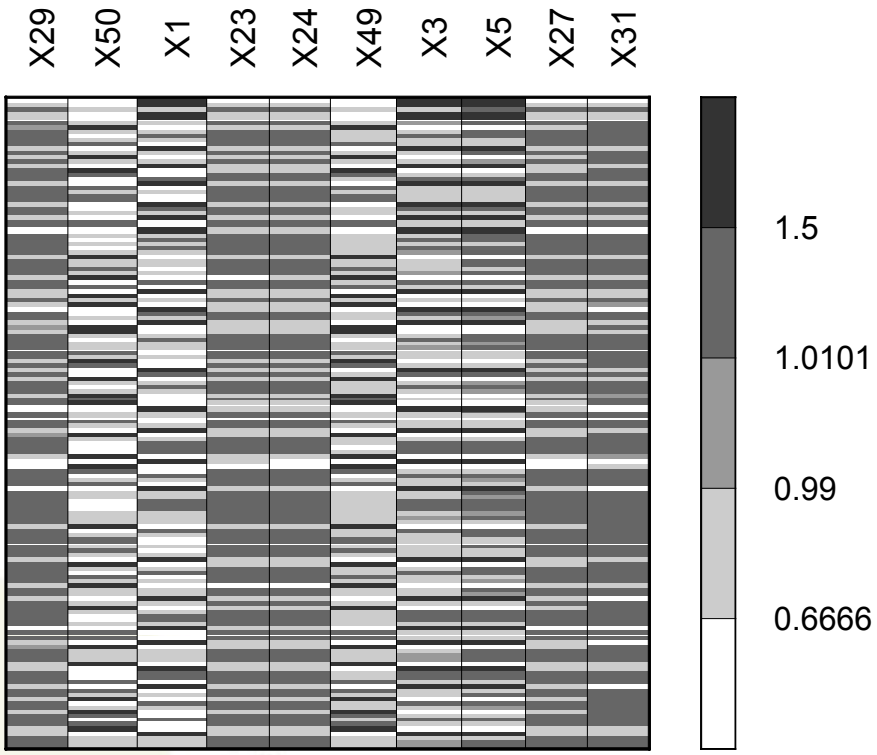
- *Rows*: map of Pearson correlations (left) and plots for the first two eigenvectors (right); correct ordering



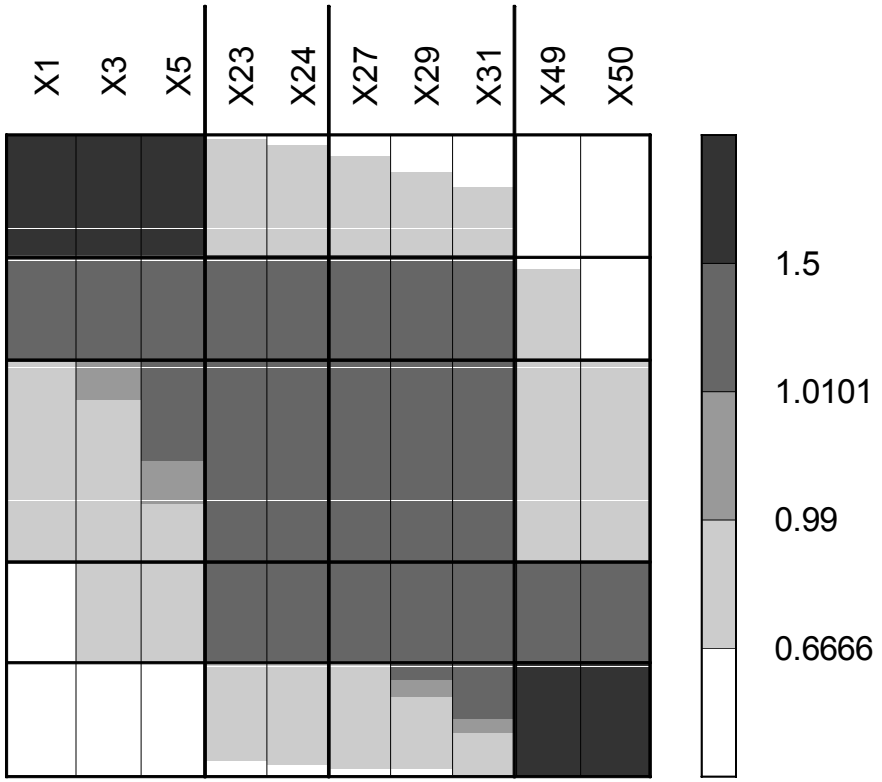


Second Artificial Data - GCCA

Original ordering



Reordered by GCA: 4 clusters for columns and 5 for rows; correct ordering

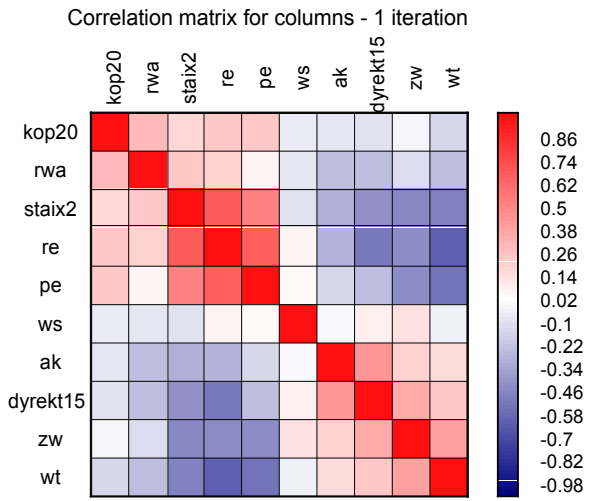


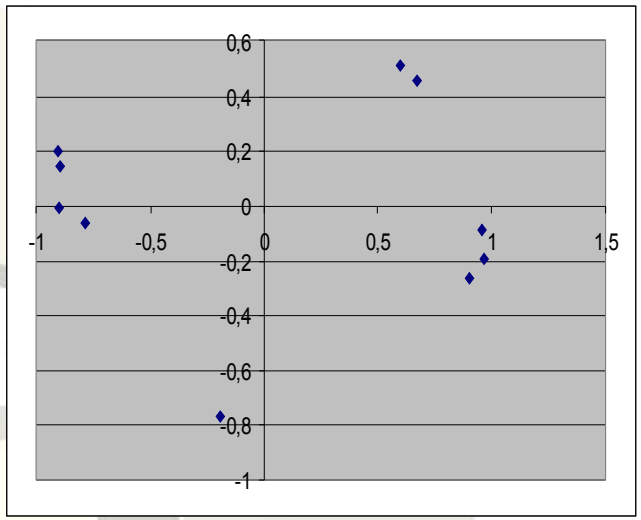
Contents

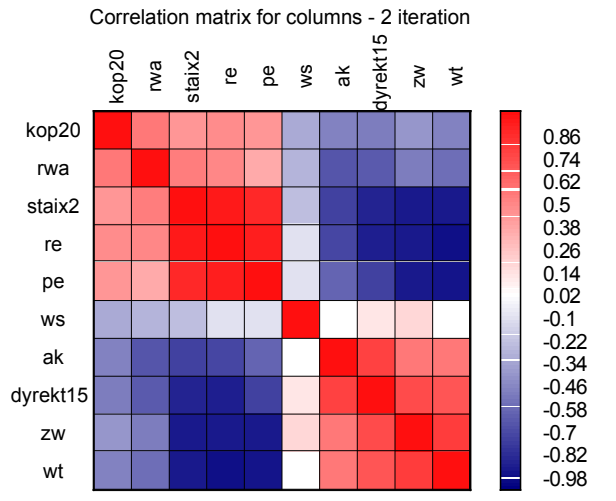
1. Introduction
2. Data description
3. Analysis of two artificial symmetrical data sets
4. Analysis of psychological data
5. Conclusions

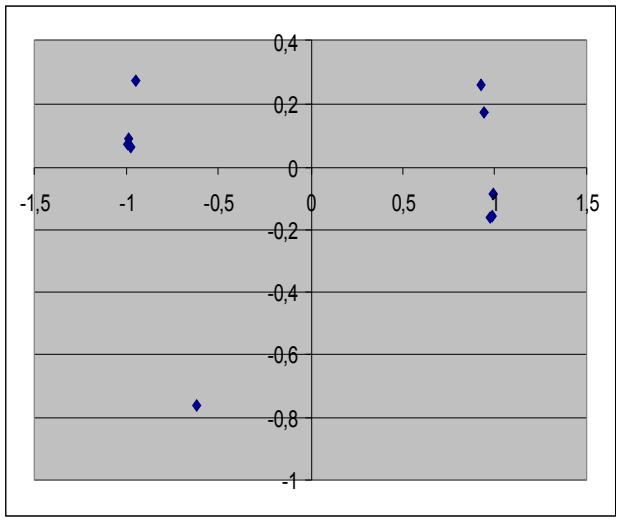
Superstition Data - GAP

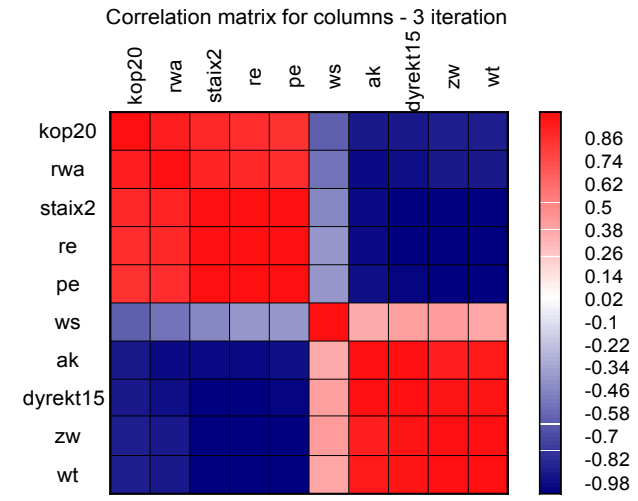
- *Columns*: Pearson correlations (upper maps) and plots for the first two eigenvectors (bottom maps), 3 iterations

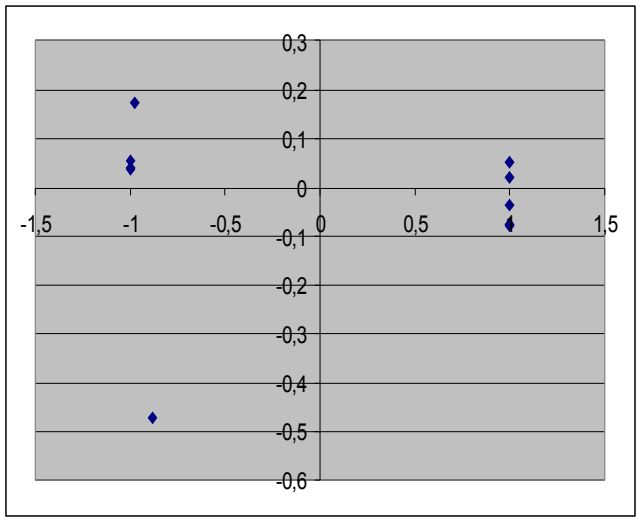






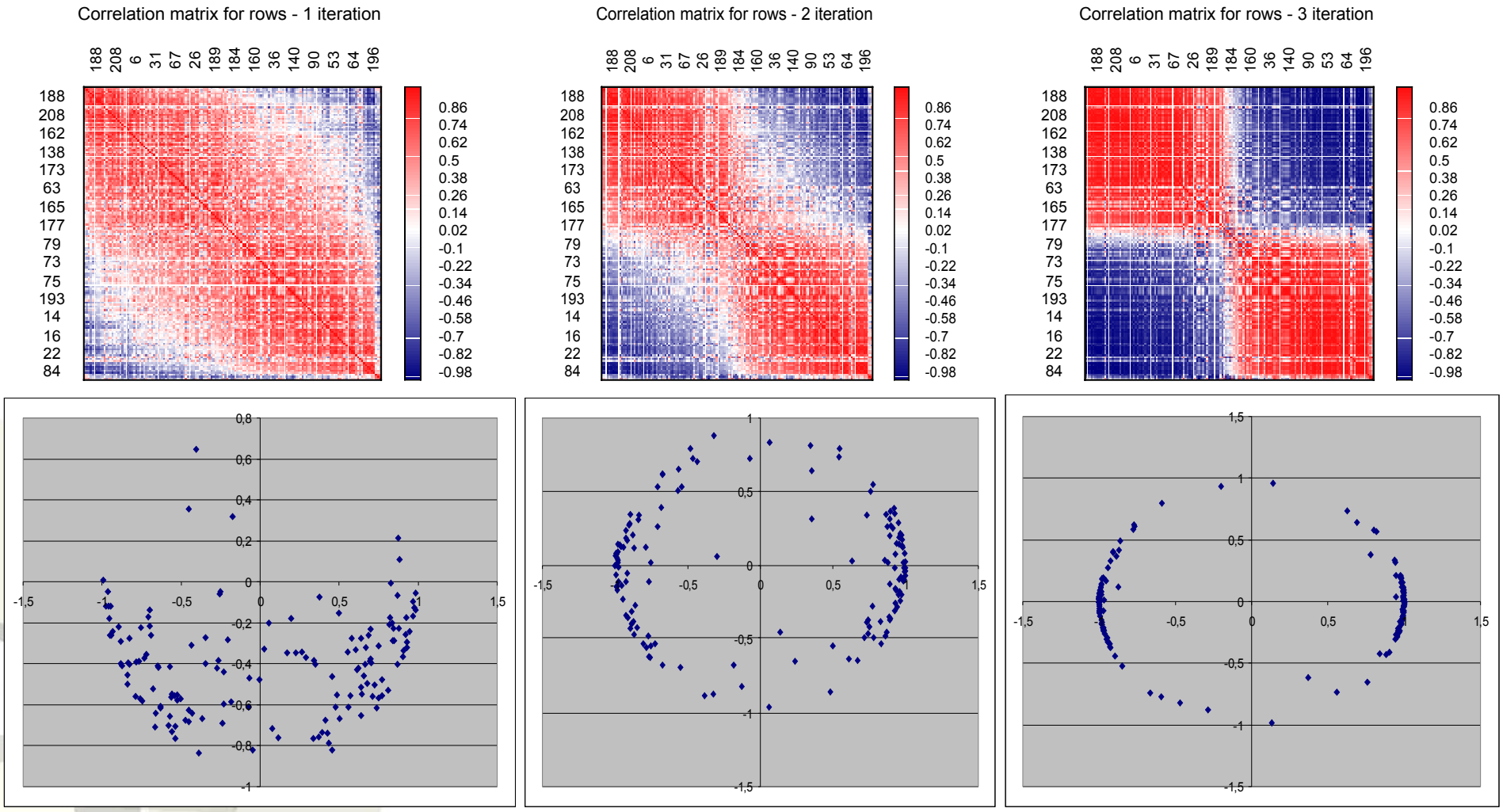






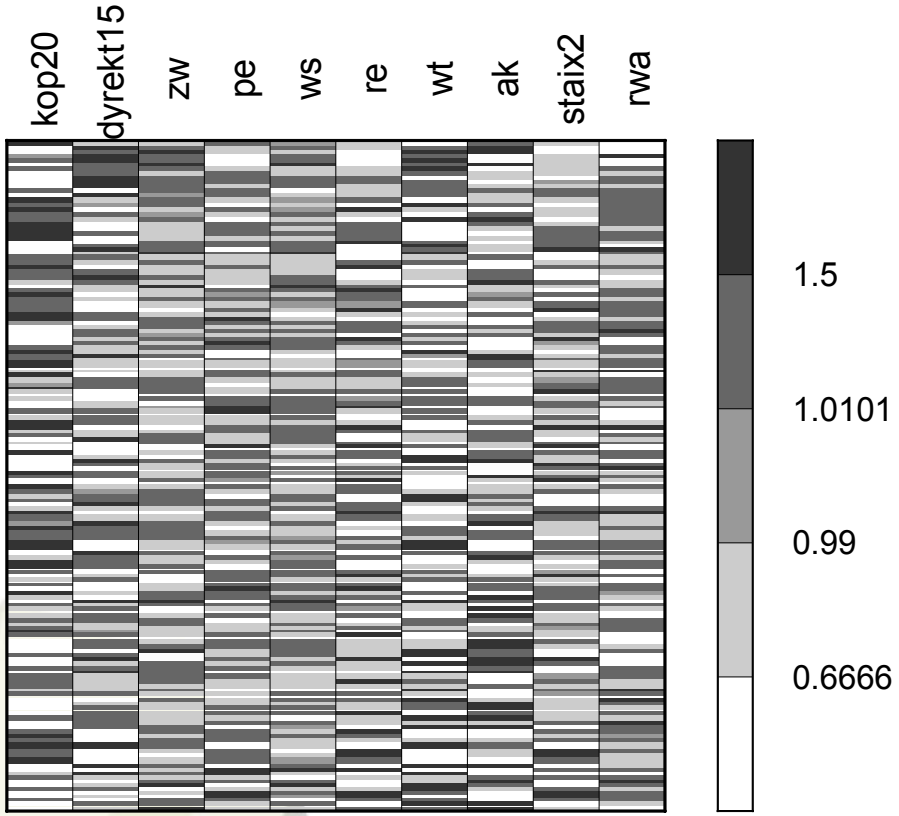
Superstition Data - GAP

- *Rows*: Pearson correlations (upper maps) and plots for the first two eigenvectors (bottom maps), 3 iterations

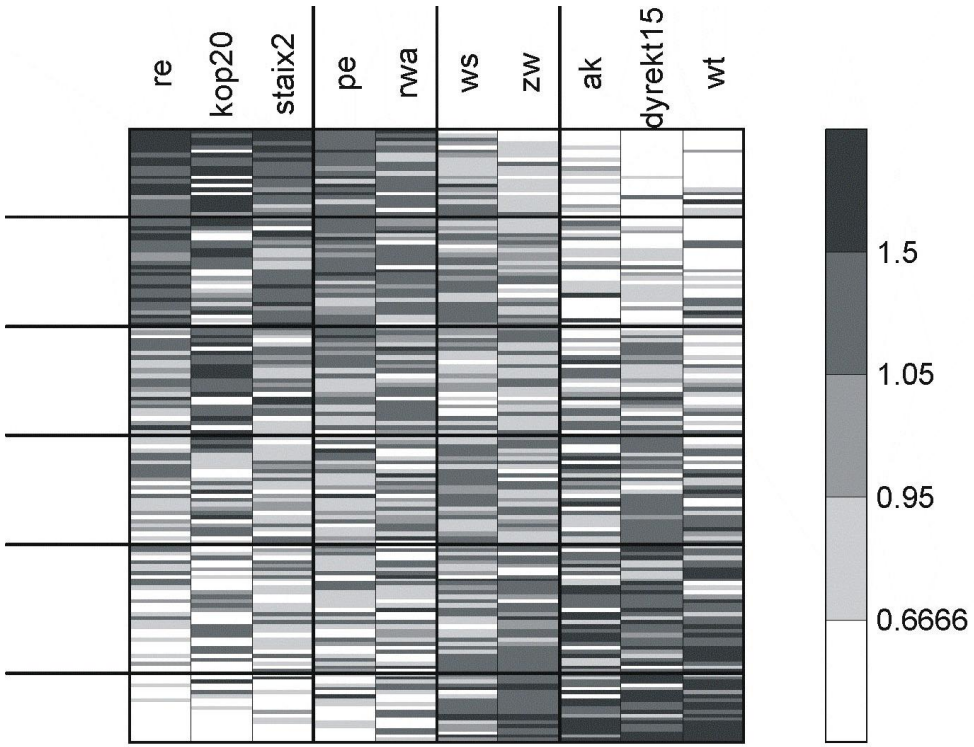


Superstition Data - GCCA

Original ordering



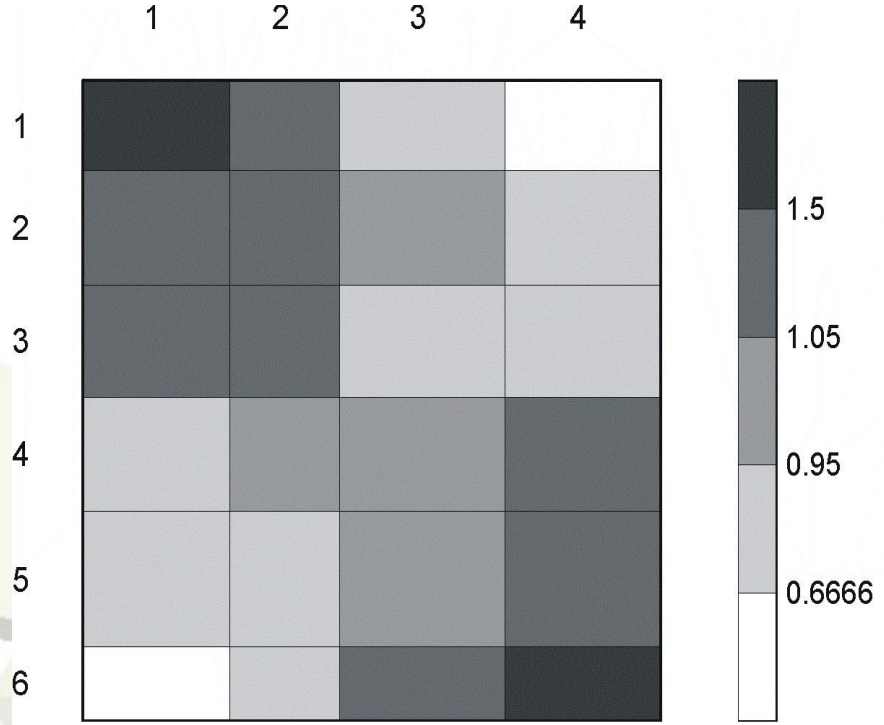
Reordered by GCA: 4 clusters for columns and 6 for rows



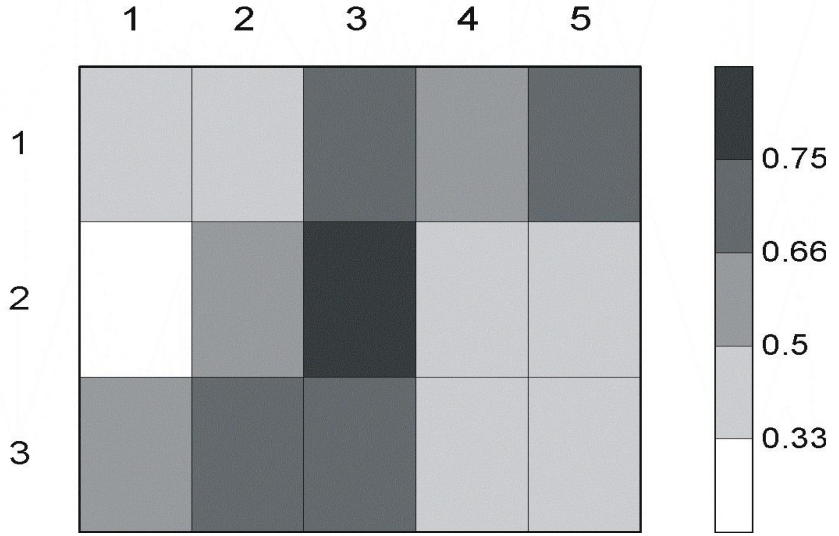
Superstition data - aggregations

- Overrepresentation maps for aggregated columns and rows

GCCA



GAP



Superstition data - aggregations

- Comparison of averages in aggregated clusters

GCCA

	re	kop20	staix2	pe	rwa	ws	zw	ak	dyrekt15	wt	ITEMS
1	0.85	0.63	0.55	0.78	0.45	0.75	0.55	0.22	0.2	0.28	21
2	0.73	0.44	0.44	0.7	0.43	0.81	0.71	0.31	0.35	0.32	27
3	0.56	0.65	0.41	0.66	0.46	0.7	0.74	0.42	0.54	0.4	25
4	0.47	0.48	0.34	0.59	0.39	0.82	0.73	0.49	0.61	0.51	26
5	0.39	0.29	0.3	0.52	0.35	0.78	0.79	0.53	0.64	0.6	32
6	0.11	0.16	0.18	0.32	0.3	0.75	0.92	0.6	0.68	0.76	19
Avg.	0.52	0.44	0.37	0.6	0.4	0.77	0.74	0.43	0.51	0.48	150

GAP

	kop20	rwa	staix2	re	pe	ws	ak	dyrekt15	zw	wt	ITEMS
1	0.35	0.36	0.27	0.31	0.45	0.78	0.53	0.66	0.84	0.62	68
2	0.18	0.38	0.42	0.67	0.72	0.81	0.45	0.52	0.6	0.38	19
3	0.62	0.45	0.46	0.7	0.72	0.75	0.32	0.35	0.68	0.35	63
Avg.	0.44	0.4	0.37	0.52	0.6	0.77	0.43	0.51	0.74	0.48	150

Superstition data - summary

- GAP's variable ordering separated two groups positively inter-correlated, they are also negatively correlated between them; single variable *ws* forms „the third” group, neutral;
- GCCA divided persons into clusters easier to interpret for researcher; clusters 1 and 3 consist of equally superstitious persons, but with different profiles of psychological traits

Contents

Conclusions

- GAP works on data matrix only by referring to chosen proximity matrices for rows and for columns; GCCA transforms data matrix into matrix of overrepresentation indices, and basing on it seeks for structures with the highest interdependence between ordered clusters of rows and of columns
- In case of regular data both methods perform well
- Lack of agreement between results of GAP and GCCA can be a good indicator that the initial data are not sufficiently regular
- For desired experimental data GCCA gave results a little easier to interpret for the researcher

Thank you!

Please visit us at:
http://gradestat.ipipan.waw.pl