COMPSTAT 2010, Paris

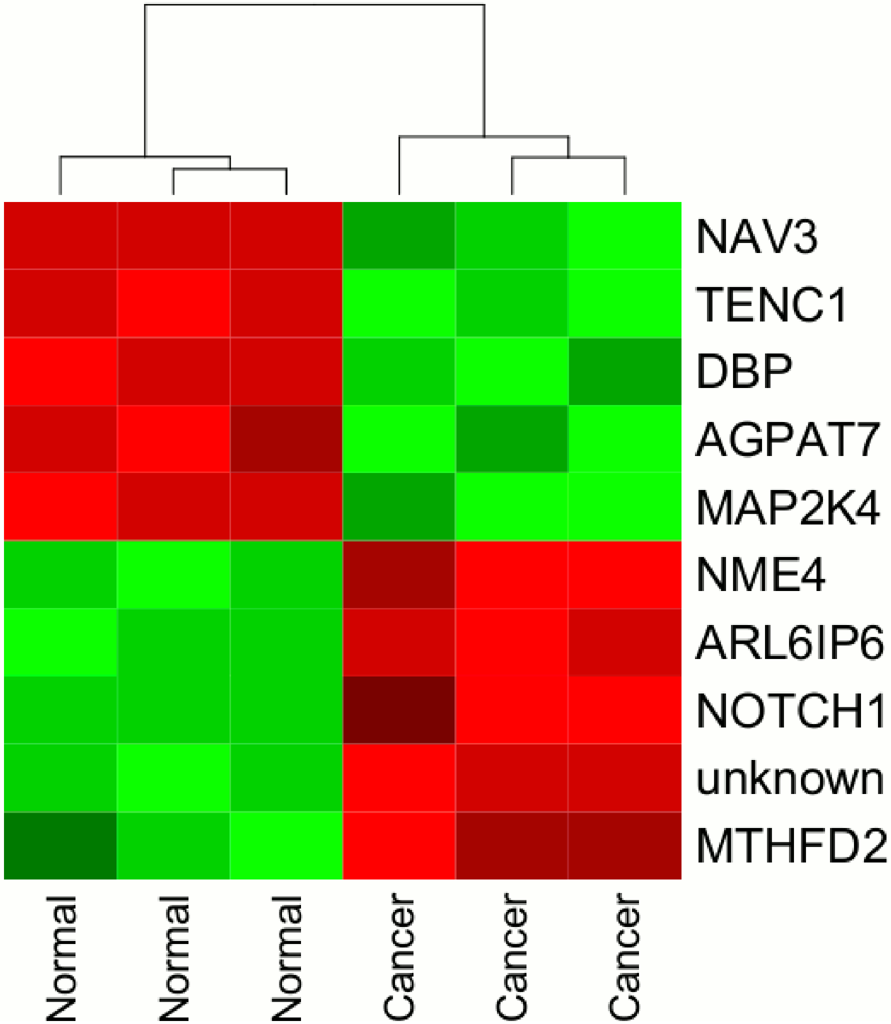
Two-way classification of a table with non-negative entries: Validation of an approach based on Correspondence Analysis and Clustering

Antonio Ciampi, Alina Dyachenko* and*
*Yves Lechevallier***
**Department of Epidemiology, Biostatistics*
and Occupational Health, McGill
University, Montreal, Qc., Canada
*** INRIA-Rocquencourt, France.*

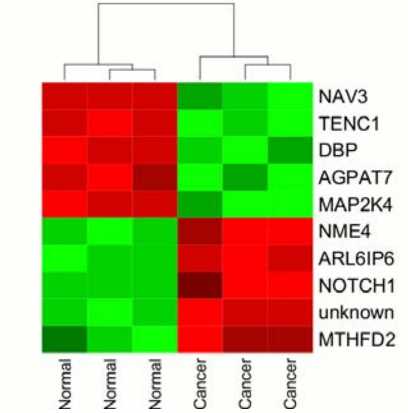
INTRODUCTION

- Exploratory techniques based on 2-way clustering and the heat map, are now used in a broad variety of fields
- Major example:
  - – Analysis of gene expression data
  - – Rectangular table, with entry n_{ij} , the expression intensity of a *gene i* in a particular *tissue j*

A heat map


NAV3
TENC1
DBP
AGPAT7
MAP2K4
NME4
ARL6IP6
NOTCH1
unknown
MTHFD2
Normal
Normal
Normal
Cancer
Cancer
Cancer

Constructing a heat map



- The heat map represents a data matrix by the following steps:
    1. Calculate a distance between rows and columns (5 options)
    2. Perform a hierarchical clustering algorithm on both rows and columns to obtain a non-unique ordering of both (5 options)
    3. Replace numbers with colors

What do we propose that is ‘new’?

- Recently:
    - A ‘natural distance’ for contingency tables
    - Reduce the dimensionality of the underlying space
    - Apply clustering algorithms in the reduced space and reorder rows and columns in consequence
- In this presentation:
    - An evaluation of the approach

Plan of presentation:

1. Background: Correspondence Analysis and 2-way clustering
2. Our general approach
3. Evaluation through limited simulations
4. An example
5. Discussion

1. Background: Correspondence Analysis and 2-way clustering

- Our approach is geometrical in nature
- The geometry is that of Correspondence Analysis
  - Quantities based on statistical inference are used in a flexible way in an exploratory mode. They are absolutely valid only under ideal circumstances
- We work with hierarchical clustering
  - we do not assume on a priori number of classes but algorithms may generate useful hints

A natural non-Euclidean distance between rows and columns of contingency tables:

– The χ^2 distance between rows i and i’:

$$d_r(i,i') = \sqrt{\left(\sum_{j=1}^{J} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{ij'}}{n_{i'+}} \right)^2 \middle/ n_{+j} \right)}$$

– and between columns j and j’:

$$d_c(j,j') = \sqrt{\left(\sum_{i=1}^{I} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{ij'}}{n_{+j'}} \right)^2 \middle/ n_{i+} \right)}$$

Properties of the χ^2 distance:

- *Distributional equivalence*:

  The distance between any two columns (rows) does not change if we agglomerate rows (columns)

- It is the Wald statistic associated to the LRS that compares:
  - H_0: 2 rows (columns) of a contingency table are from the same multinomial distribution

  with

  - H_1: they come from two different multinomial distributions.

Clustering with the χ^2 distance

- One way to obtain insight in the structure of a data matrix is to cluster rows and columns. We speak then of *two way clustering*
- The χ^2 distance can be used to develop
  - *hierarchical clustering* algorithm (Greenacres)
  - *optimal partitioning* algorithm (Govaert)
- It is interesting to compare the two types of algorithms, but this is beyond the scope of this presentation

Correspondence Analysis in a nutshell

- *Correspondence Analysis* (CA) is a Principal Component Analysis (PCA) with the χ^2 metric.
- It exploits the symmetry between rows and columns more conveniently than ordinary PCA
- It includes powerful aids to interpretation. The most important for us is an index of *quality of the representation*:
  - Given any k-dimensional subspace identified by CA, each row and column has a quality of representation index from 0 to 100

Inertia and its decomposition

- Two basic concepts in CA
  - *Weight* of a row (column): marginal total of the row (column)
  - *Inertia* of a cloud of points: sum of the square distances of each point from the cloud's center of gravity
- Properties of inertia:
  - Row-cloud inertia = Column-cloud inertia = Inertia
  - The inertia is an approximation to the Kullback-Liebler number, which is a likelihood-based measure of the information contained in a contingency table

Reducing dimensionality before clustering

- If our $r \times c$ table is large a preliminary reduction of dimension is often desirable
- CA represents *isometrically* the row and column clouds in and Euclidean space E^m , $m = \min(r, c)$
- CA induces a decomposition of the inertia into a sum of decreasing non-negative components associated to 1-dimensional subspaces of E^m
- The decomposition is obtained by spectral analysis of a transformed version of the original data table (Greenacres)

- Suppose now that the first *k* components represent a high percentage of the total inertia (information), e.g. 80%. Let us project both clouds in E^{k}, the Euclidean space spanned by the first *k*-dimensional subspaces of the decomposition
- Then this lower dimensional representation contains most of the information in the data:
  - *This may be enough to capture the essential features: by eliminating dimensions with negligible components of inertia, one may eliminate noise and highlight signal*
- *Remark:* Inertia indicates presence of clusters in the data (Caussinus & Ruiz):
  - *roughly speaking, a subspace containing most of the inertia contains most of the clustering structure present in the data*

2. Our general approach

1. Perform a CA and select $k << m$, so that the inertia contained in E^k is p% of the total (for a pre-determined p)
2. In E^k calculate the distances between the row and column clouds. Apply hierarchical clustering to both clouds. Rearrange rows and columns according to a non-unique order induced by the clustering
3. Cut the dendrograms, obtaining a partition of the row cloud into p_r classes and a partition of the column cloud into p_c classes, so as to keep essential information

Choices: dimensionality of the representation and number of classes

- To specify the algorithm, we have to choose the dimensionalitty of the representation *k*
- To cut the row and column dendrograms we have to specify the number of row classes, p_r and the number of column classes p_c
- We obtain from this $p_r \times p_c$ *blocks*

Algorithm for choosing *k*

- From the data table *T*, randomly generate *T** with the same marginal totals as T but with independent rows and columns. Repeat this construction *M* times (*M* large) to obtain a family $\{T^*(m), m = 1,\ldots, M\}$
- Draw the scree plot of *T*, and the envelope of the scree plots of $\{T^*(m), m = 1,\ldots M\}$ on the same graph.
- Choose for *k* the abscissa of the point immediately preceding the first point which falls below the simulation band

Algorithm to choose p_r and p_c : 2 variants

1. Associate a statistical model to any clustering of rows and columns. Calculate the *AIC (BIC)*
2. Variant A: Cut each dendrogram at the minimum *AIC (BIC)* level, obtaining pr and pc separately
3. Variant B: Consider all pairs of levels and calculate the *AIC (BIC)* for each such pair and determine the pair with minimum *AIC (BIC)*

3. Evaluation

- Two key choices to validate:
  - a) number of axes of the CA of the original data;
  - b) number of clusters using *AIC/BIC* (two variants)
- Evaluation by simulation experiments

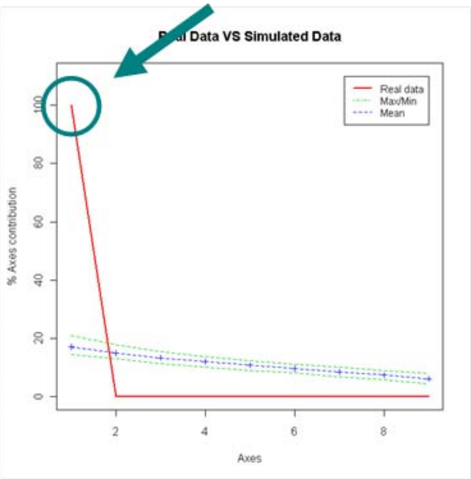
a) Choice of number of axes: simulation

- We generated one 100 x 10 contingency table of total sum 10,000, under the hypothesis of independence of rows and columns
- We applied the Singular Value Decomposition (SVD) to the matrix of standardized residuals to extract Singular Values (SV) and the matrices of left and right singular vectors of dimension 10×10 and 100×10 respectively
- We built 10-1 = 9 contingency tables :

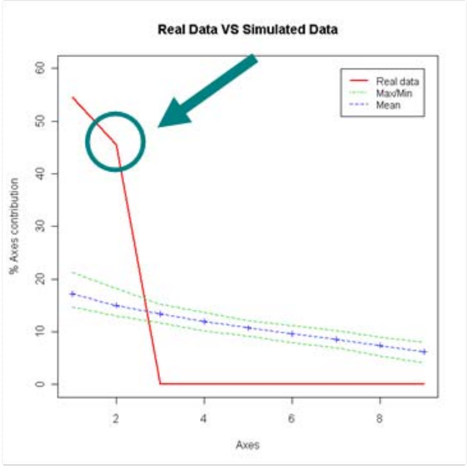
    The i-th matrix, i = 1,…,9, was constructed by reconstituting the SVD but setting all the eigenvalues from i to 9 equal to zero

Results: scree plots and decision rule

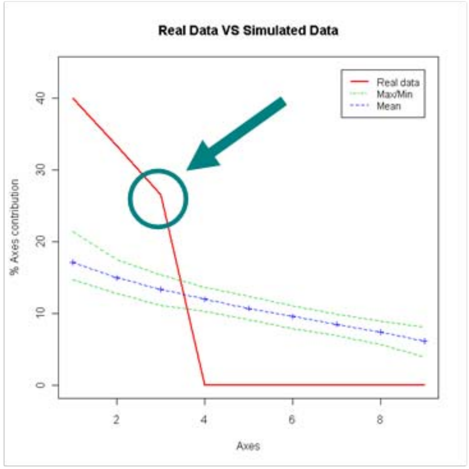
1 axe



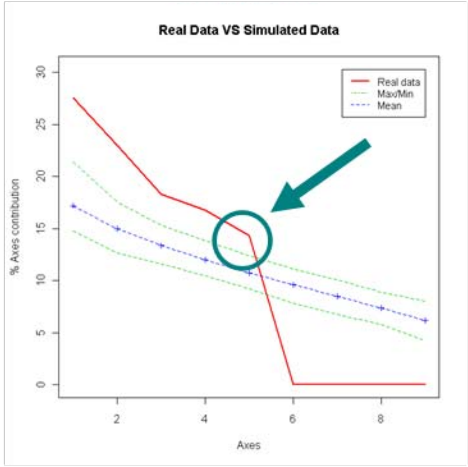
2 axes



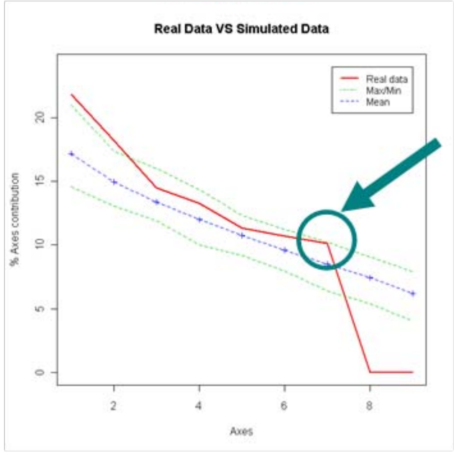
3 axes



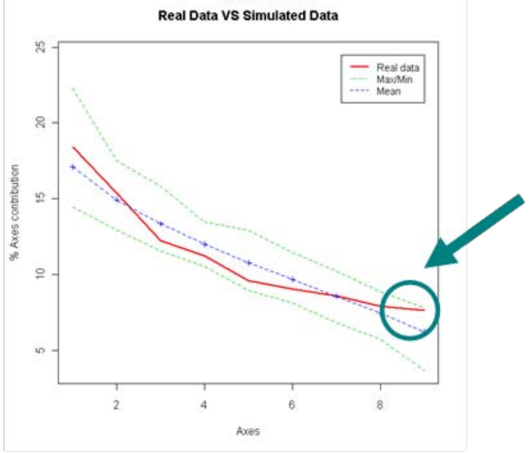
5 axes



7 axes



9 axes



Interpretation

- The simple rule of thumb is confirmed: the number of axes selected by the rule is the correct number
- Notice that in the last graph (*k* = 9) the curve is entirely comprised within the simulation band, suggesting that no data reduction is possible and we have to use all the dimensions

b) Choice of number of clusters by AIC/BIC

- We associate a statistical model to each clustering of rows and columns
- We assume a multinomial distribution for a contingency table and calculate *AIC* and *BIC* by each step of the clustering algorithm
- We have calculated the likelihood ratio statistics of each model with respect to the saturated model as well as a $\chi 2$ approximation

Simulation

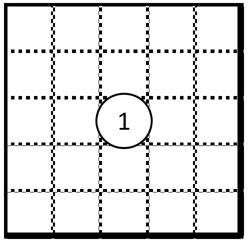
- We have simulated 5x5 contingency tables of total size 10,000, according to several multinomial models, corresponding to 1,2,3 and 4 blocks. Each table has been generated 1000 times.

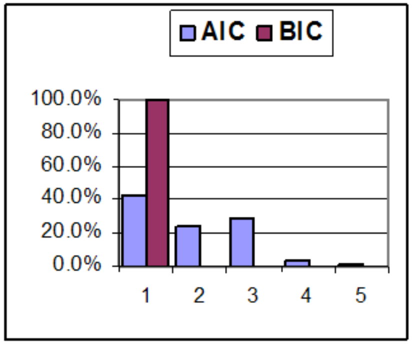
Results

# Row Blocks	Prob of blocs	Detection of real structure			
x # Col Blocks	(fixed: N_tot=10,000)	AIC	BIC	X2_aic	X2_bic
1 x 1	1	42.3%	100.0%	41.5%	100.0%
2 x 1	0.6; 0.4	52.3%	100.0%	51.7%	100.0%
3 x 1	0.3; 0.6; 0.1	56.2%	100.0%	54.8%	100.0%
4 x 1	0.14; 0.20; 0.28; 0.38	56.5%	100.0%	56.8%	100.0%
2 x 2	0.70; 0; 0.3; 0	96.7%	100.0%	96.0%	100.0%
2 x 2	0.60; 0.05; 0.30; 0.05	64.9%	100.0%	54.5%	100.0%

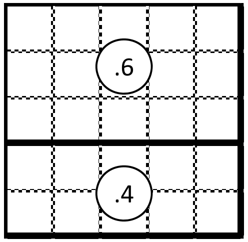
Number of retrieved blocks: histograms

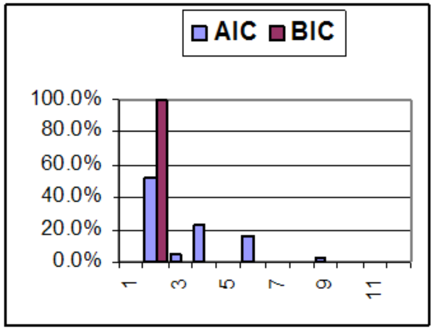
1 x 1 bloc



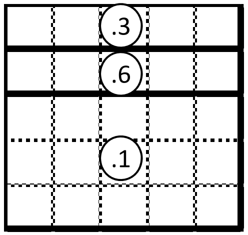


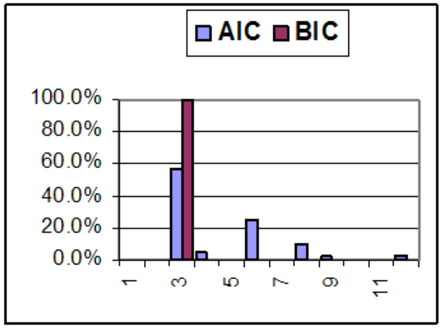
2 x 1 blocs



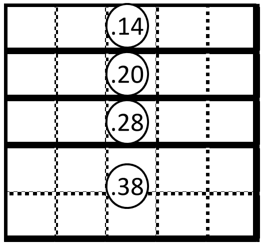


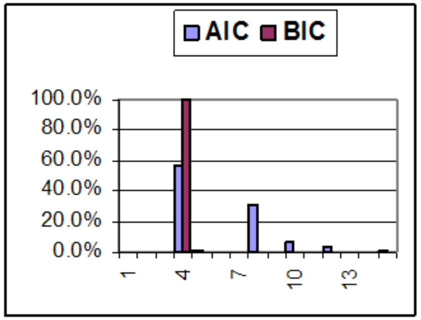
3 x 1 blocs



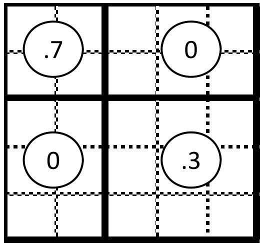


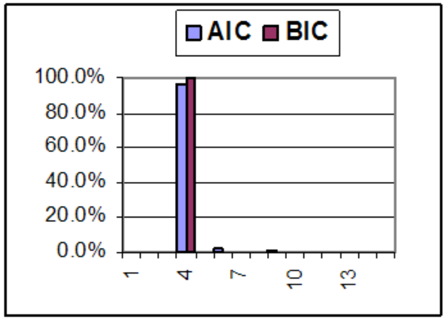
4 x 1 blocs



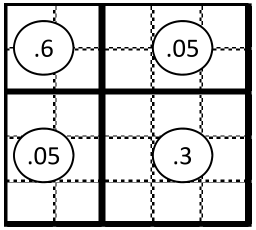


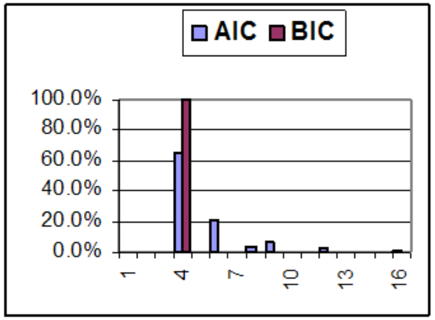
2 x 2blocs





2 x 2 blocs





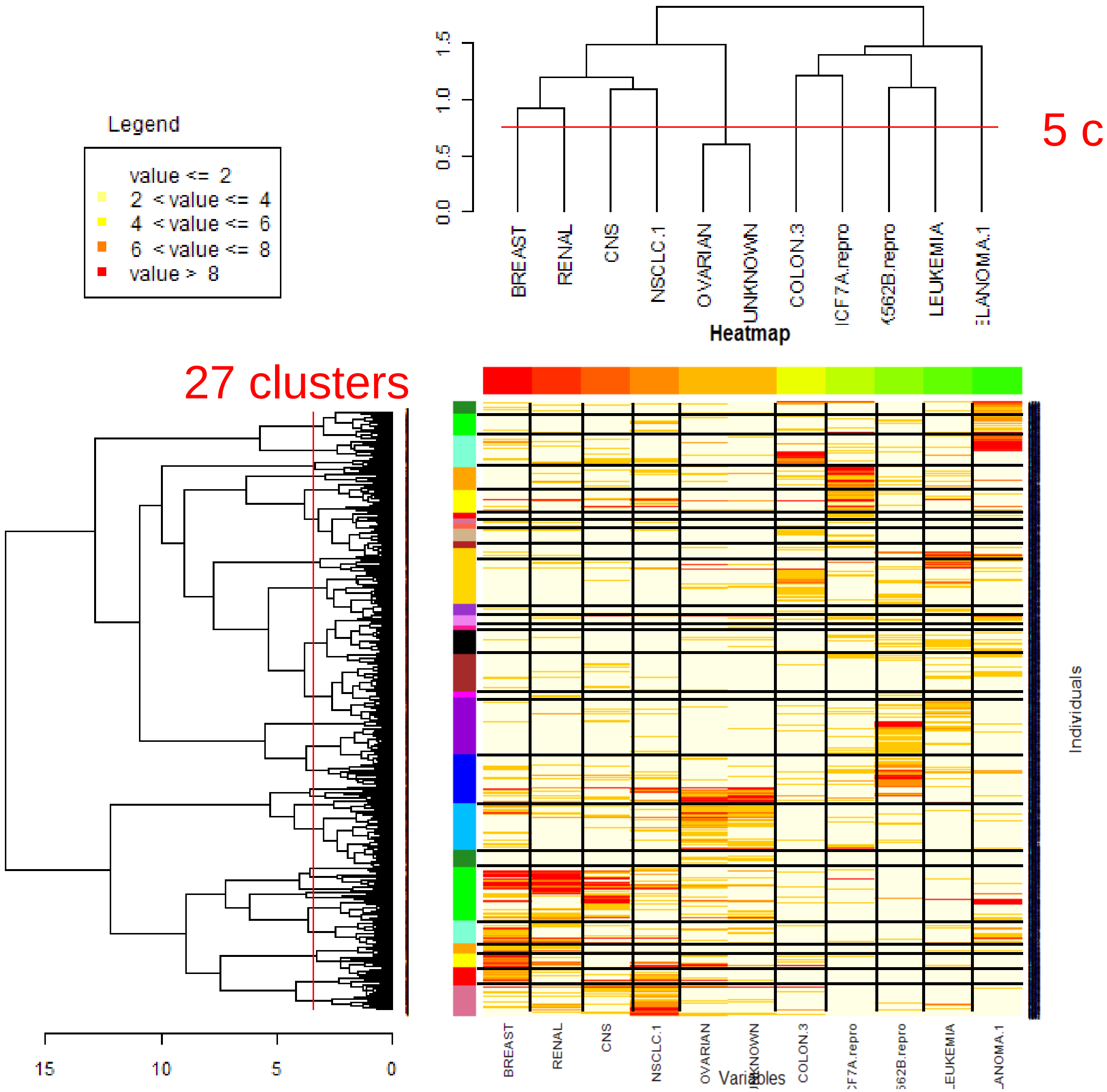
4. An example

- Publicly available NCI microarray data: expression level of 6830 genes in 64 cancerous tissues
- Our selection: 2000 genes (at random) and 11 malignancies:
  - CNS, RENAL, BREAST, NSCLC, UNKNOWN, OVARIAN, LEUKEMIA, K562B.repro, COLON, MELANOMA, MCF7A.repro

Goal of the analysis: *to identify blocks of genes and tissues with distinct profile*

1. Pre-processing the data: *to obtain non-negative entries (the original data are in the logarithmic scale)*
2. Correspondence Analysis: *reduction to 11 dimensional Euclidean space*
3. Ward hierarchical clustering algorithm: *applied to both rows and columns using representation in reduced space*

Results: the heat map



5 clusters

27 clusters

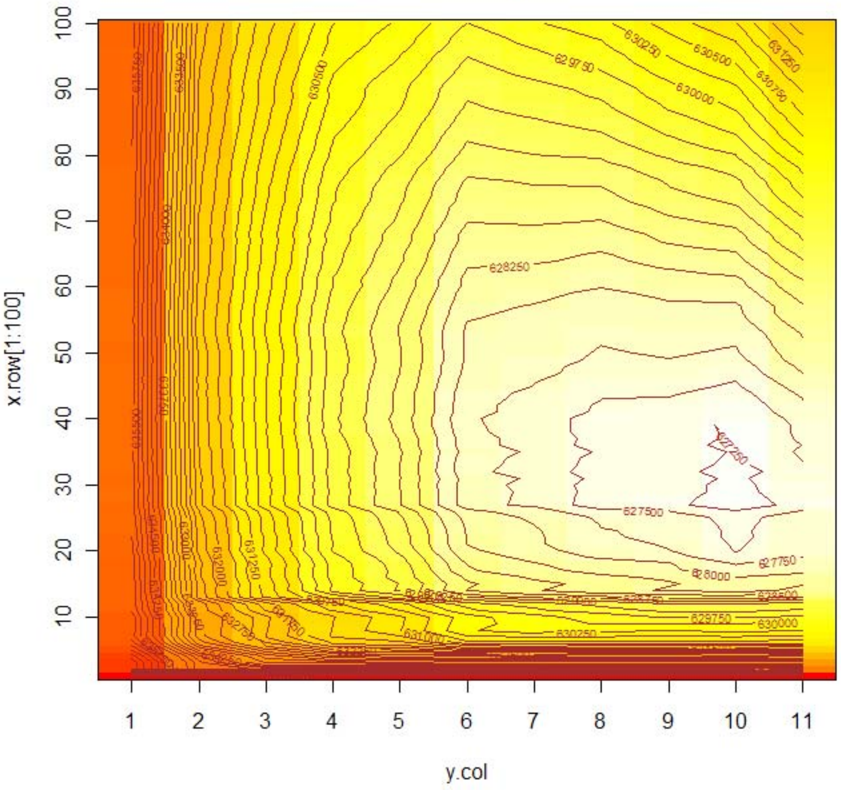
Cutting the dendrograms

1. Calculate the BIC for every possible cut of the row-dendrogram and column-dendrogram
2. Chose the cut corresponding to the minimum value in this table
3. Useful to look at the contour and perspective plots of the BIC table

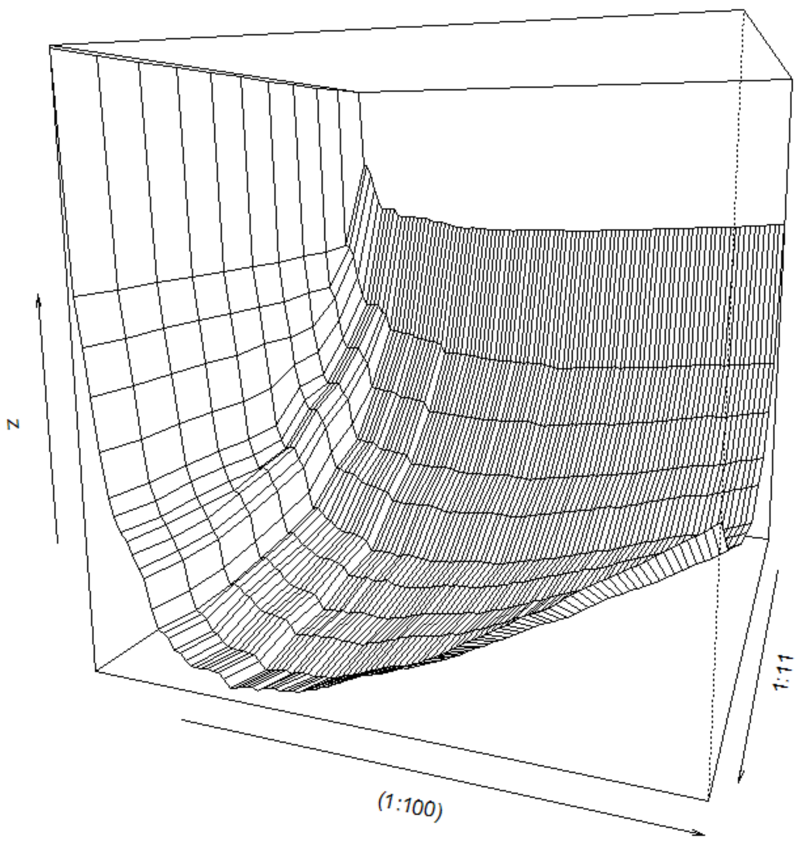
Remark: *Similar AIC table and plots yielded comparable results with a few more clusters (data not shown)*

Contour and perspective plots

BIC contour plot



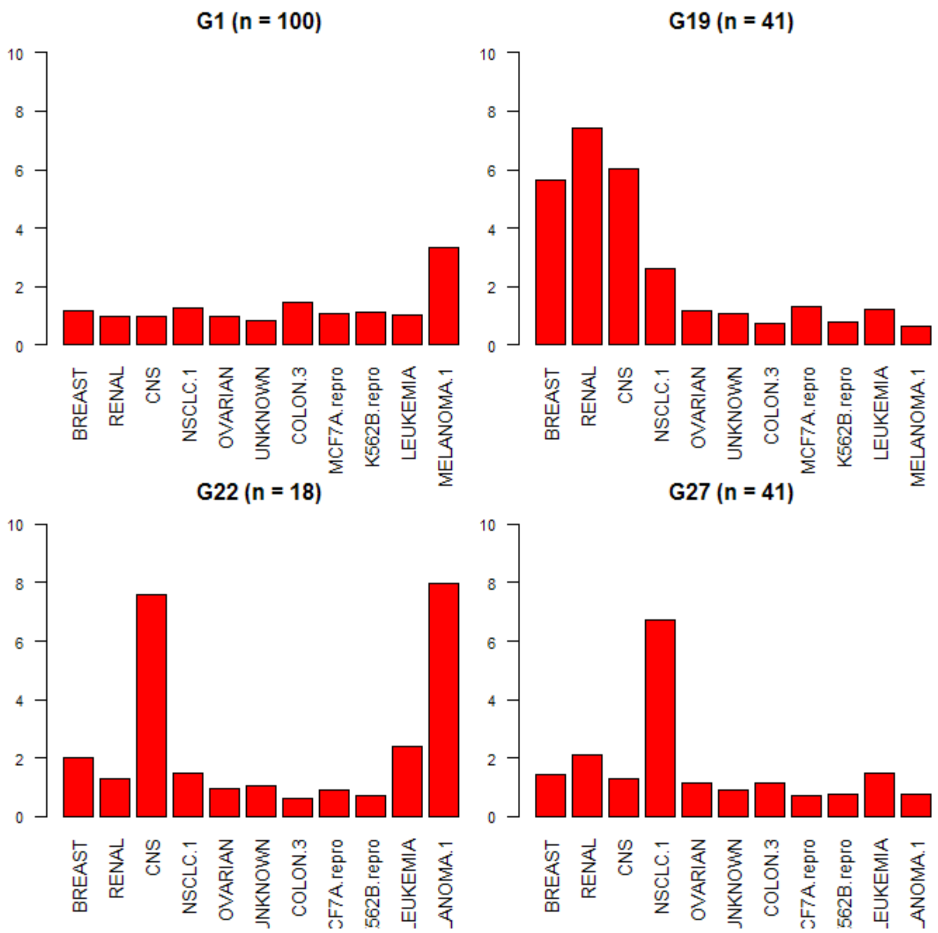
BIC perspective plot (clusters 1 to 100)



Details

- The minimum BIC rule identifies 27 groups of genes and 10 groups of tissues (OVARIAN and UNKNOWN tissues merge according to this clustering)
- The alternative strategy (clustering separately rows and columns) yields also 27 groups of genes but only a unique group of tissues, a result which is not helpful and unintuitive
- The 2-dimensional BIC plots help exclude this solution, since it is represented on the steep portion of the surface

Results: a selection of 4 of 27 distinct profiles



5. Discussion

- Our medium and long-term methodological development aims to produce interpretable two-way classifications of data tables with non-negative entries
- Progress achieved here:
  - a) Validation of the data reduction proposed earlier resulting in improved heat map
  - b) An AIC/BIC based approach to choosing how to cut the dendrograms
  - c) Contour and perspective plots for exploring various possible choices of block clustering

Current and future research

- Systematic simulations
- Comparison with block clustering
  - is one approach clearly superior?
  - Can the two approaches be used in a complementary way, e.g. in choosing the number of clusters
- Other areas for further research:
  - a) Comparison of AIC and the BIC with other criteria for model selection, e.g. ICL, CS and NEC
  - b) Introduction of expert input in validating the clusters
  - c) Comparison of CA with other techniques of scaling;
  - d) Variable and object selection.