Slimming down a high-dimensional binary datatable: relevant eigen-subspace and substantial content

Alain Lelu
Université de Franche-Comté / LASELDI
LORIA / KIWI
alain.lelu@univ-fcomte.fr

Introduction: the K* problem

- Big issue = determining the *number of relevant dimensions* in a dataset
  - Data analysis (PCA, CA, ...): how many axes are akin to be interpreted ?
  - Indexing large document collections:
    - **Text collections** (Latent Semantic Indexing = SVD of the document-word matrix): how many singular vectors to take into account for computing similarities ?
    - **Image, video and audio collections**: 1) Define a distance between multimedia items, 2) Out of the pairwise distance table, infer the intrinsic dimension of the data ← for optimizing the data storage and response times to similarity queries (cf. project DISCO ← CNAM/CEDRIC lab.)

Our objective

- Solve the K* problem in the case of **binary** datatables
- and specifically: **whatever the row and column distributions** – e.g. including the very common case of subject-feature matrices, with « Zipfian », power-law distributions of the features:
  - Text mining
  - Biological datasets
  - Graph mining
  - ...

Determining K*: state-of-the-art

- Heuristics:
  - « Scree-break » in the diagram of the successive eigenvalues: visual, or second differences (Cattell 1966)
  - 95% of the total inertia.
  - ...
- Model-based parametric tests ← *assume the type of underlying distributions*

Our solution = use a randomization test (Manly 1997)

- « TourneBool » method (Cadot 2006):
  - Generate a sufficient sample (X1, X2, ... Xp) of randomized versions of the original matrix X0 (e.g. 200 matrices) subject to the same distributional constraints = **same margins as X0**.
  - Extract the full sequence of singular values of X0, in decreasing order.
  - For each k-order eigen-space, starting from k = 1, compare the k-th singular value of X0 to the sorted set of corresponding k-th singular values in the sample:
    - if the current singular value $\lambda_k(X0) \geq$ the randomized one located at the significance threshold (e.g.: the third one at the 99% threshold), it is deemed *significantly diverging from randomness*, and the algorithm goes on with k = k + 1.
  - When stopping, K* = k

Sub-problem: how to generate randomized versions of a binary matrix, subject to prescribed margins ?

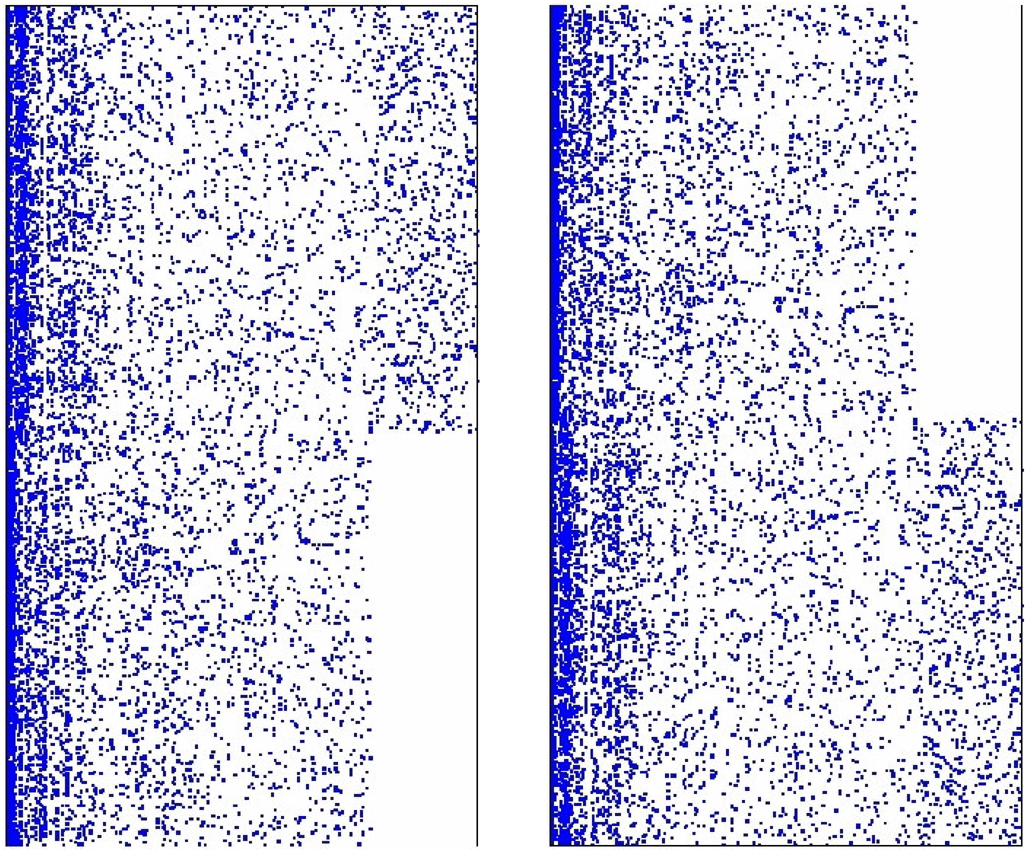
- an elementary *flip-flop exchange* does not alter the margins: $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \rightarrow \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$
- [Cadot 2006]: any matrix with same margins as X0 yields from X0 by a finite set of *cascading flip-flops* → applications in data mining
- A principle (re-)invented several times in different fields: ecology (Connor 1979), psychometrics (Snijders 2004), combinatorics (Ryser 1964), graphs (Milo 2008)

Randomize module in TourneBool

- Algorithm:
  - Choose a number r of flip-flops to execute
  - Copy X0 to Xc
  - Repeat r times
    - randomly choose (with replacement) a row pair and a column pair
    - if the zeros and ones alternate at the vertices of the rectangle in Xc, then modify Xc moving each value to its complement to 1, else do nothing
  - Store Xc
- r is chosen so as the Hamming distance X0-Xc stabilizes (no bias = memory of X0 in Xc), e.g. several times nnz.

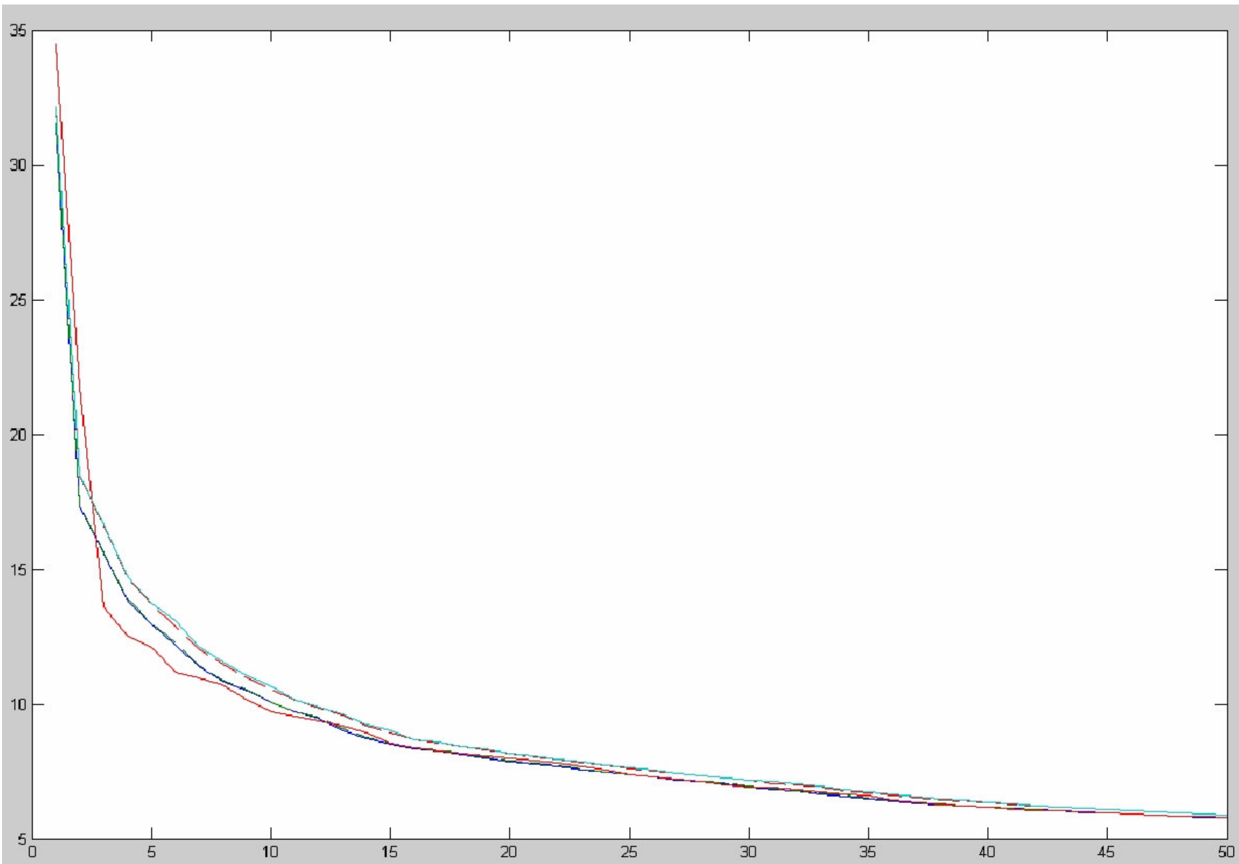
Validating on artificial data - 1

– Building a two-intertwined cluster structure (1500* 836) with a power-law distribution of the binary features



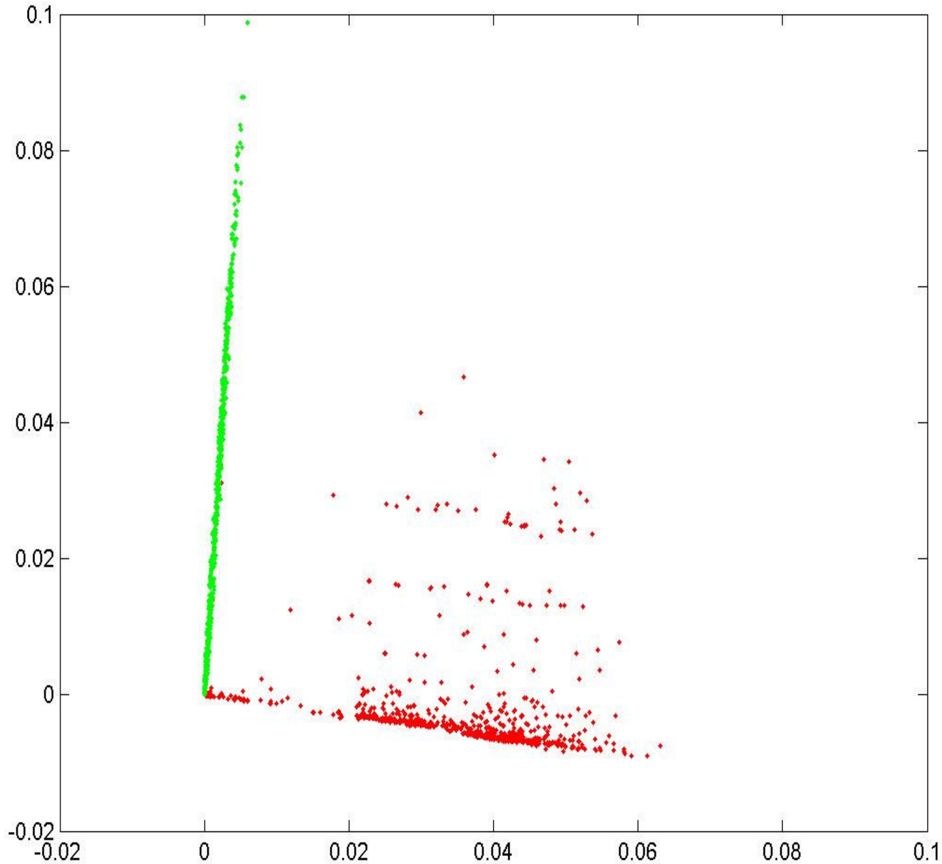
Validating on artificial data - 2

– Results: scree-plot of the 50 first eigenvalues → 2 relevant ones at the 99% confidence theshold (in red: $\lambda_k(X0)$).



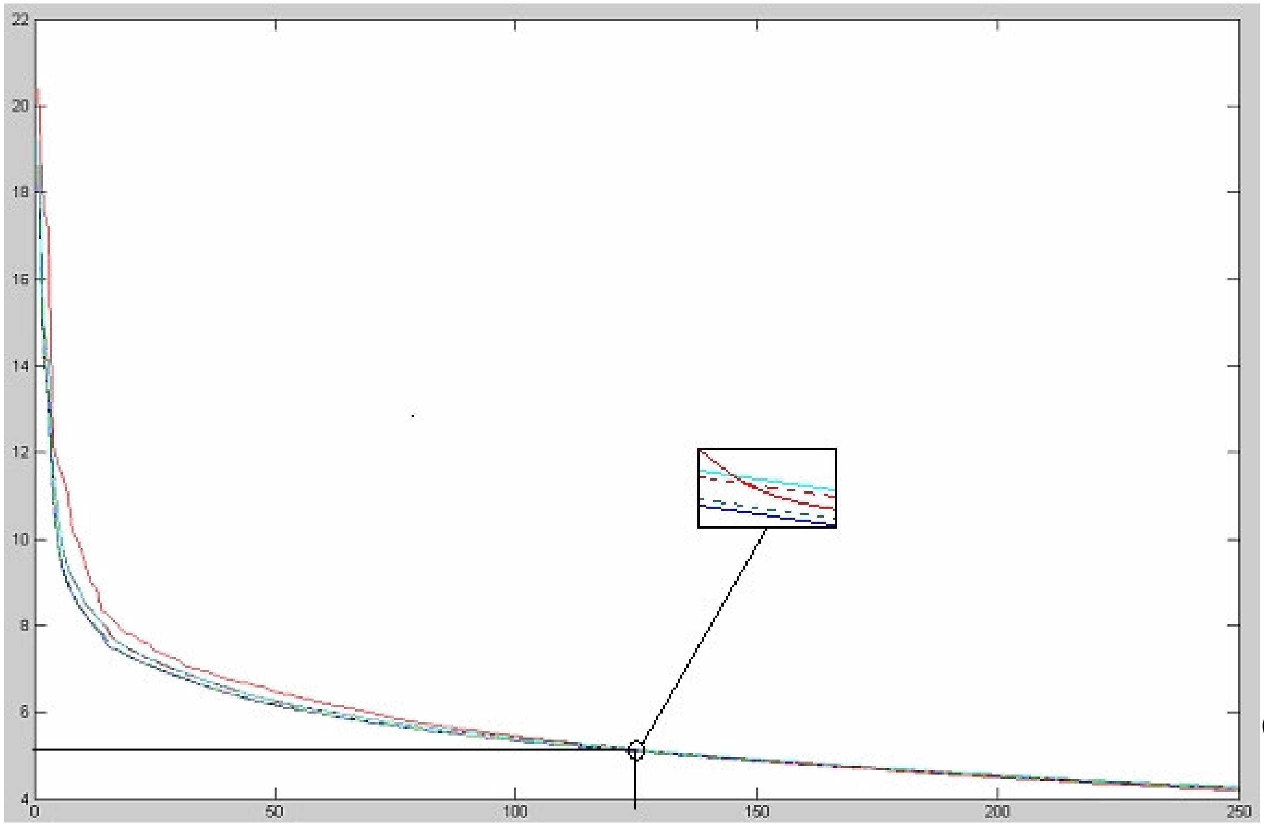
Validating on artificial data - 3

– 2 relevant eigenvalues → 2 « independant » overlapping clusters



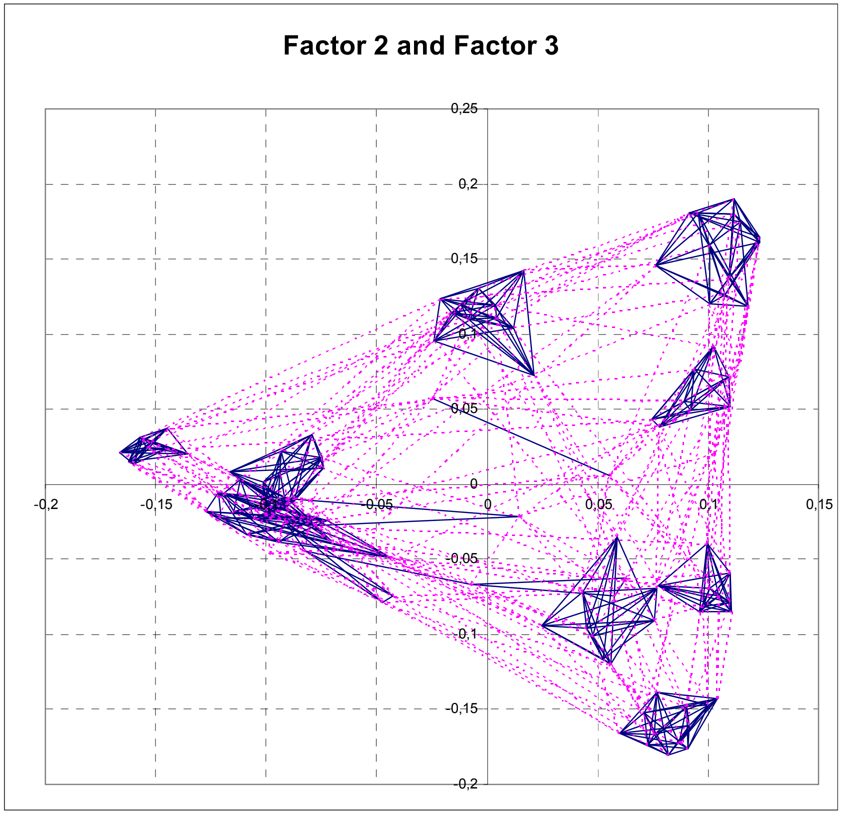
Relevant eigen-subspace of a real-world binary data-table

- Excerpt from the Pascal (INIST) database: Science in Lorraine – 1920 bibliographic entries, 3557 keywords
- Results: scree-plot of the 250 first eigenvalues → 125 relevant ones at the 99% confidence theshold (in red: $\lambda_k(X0)$). Difficult to validate !



Quantitative validation: Girvan-Newman's « Football league » graph dataset - 1

- Data: binary symmetric adjacency matrix of the games between 115 teams, structured in 12 « conferences » →
  - extra constraints for the randomized matrices: symmetry, diagonal with zeros.
  - results: 11-D relevant eigenspace at the 99% confidence theshold



Quantitative validation: Girvan-Newman's « Football league » graph dataset - 2

- Density clustering in this 11-D intrinsic « sphered » space:

# dimensions	F-score
10	.931
11	**.934**
12	.915

Conclusions, perspectives

- Encouraging qualitative and quantitative results as to *K* problem* for binary matrices
- More artificial and real-life datasets have to be worked out
- Very preliminary results as to reconstitution of the data starting from the sole relevant eigenvectors (max. Fscore .80 with threshold .3)
- Parallel implementation possible and needed for large-scale datasets

Thank you for your attention !