Binary data flows visualizations on factorial axes

Data streams

Alfonso Iodice D'Enza and Francesco Palumbo
iodicede@unina.it, Palumbo@unimc.it

European Workshop on Data Stream Analysis

Outline

Definition of data stream (DS)

Data stream mining

Data stream mining is the process of knowledge extraction from data produced at a high rate . The relation structures characterizing data are transient and they should be detected in real time and data should not be stored for a long term (Muthukrishnan, 2003).

- stock-market exchange
- network traffic
- sensor data
- ...

Binary data streams

Association/affinity analysis of binary data streams

The focus is on binary strings recording the presence/absence of a set of attributes (items):

- a binary stream can record the pages of an e-commerce web-site visited by a user in a single session
- the presence/absence of attributes can refer to the different stata of a production process being monitored

DS techniques

Categories of DS techniques (Gaber et al. 2007)

- data-based: techniques aiming at reducing the amount of streams to be analyzed
- task-based: existing algorithms adaptation from static data to dynamic data
- mining techniques: properly aimed to extract knowledge form data streams

Assumptions

Data

- data consist of binary records
- the flow consist of $n \times p$ (n records, p attributes) matrices produced in each time frame (regular data-flow)

Assumptions

Data

- data consist of binary records
- the flow consist of $n \times p$ (n records, p attributes) matrices produced in each time frame (regular data-flow)

Items association structure

- items associations are assumed to be *a-priori* known: the *starting condition*
- the starting condition is equivalent to the concept of Null Hypothesis in the statistical inductive knowledge paradigm

Aim of the proposal

Association/affinity analysis of binary data streams

The general aim of the proposal is to *monitor* through visualization the transient association structure among sets of items (attributes): in particular, changes in the association patterns are observed with respect to a *starting* situation. The starting association structure can be user-defined:

- referred to previous data
- *status* of co-occurrence of attributes for which the process is stable

Binary data streams: tabular data

Matrix formalization of binary streams

Consider a binary record as a p-dimensional vector storing the presence/absence of a set of observed attributes: a set of n binary records (n number of records in a single time-frame), corresponds to an indicator matrix $\mathbf{Z}_{(n \times p)}$

Binary data streams: tabular data

Matrix formalization of binary streams

Consider a binary record as a p-dimensional vector storing the presence/absence of a set of observed attributes: a set of n binary records (n number of records in a single time-frame), corresponds to an indicator matrix $\mathbf{Z}_{(n \times p)}$

Matrix formalization of binary streams

	$item_1$	$item_2$	$\ldots$	$item_p$
$record_1$	1	1	$\ldots$	1
$record_2$	0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$record_n$	0	0	$\ldots$	1

Multidimensional data analysis tools

Association/affinity analysis of binary data streams

- The proposed strategy is in the Multidimensional Data Analysis (MDA) framework, and more specifically exploits Correspondence Analysis techniques to obtain a synthetic graphical representation of the attributes association.

Multidimensional data analysis tools

Association/affinity analysis of binary data streams

- The proposed strategy is in the Multidimensional Data Analysis (MDA) framework, and more specifically exploits Correspondence Analysis techniques to obtain a synthetic graphical representation of the attributes association.
- Approach key-elements:
  - each attribute or item is transformed from binary to quantitative coding
  - visualization

Key-elements: quantification

Quantification of binary variables

Use of MCA on binary data is such a quantification of the starting binary variables into a reduced number of latent variables. Advantages in monitoring associations:

1. remove noise and redundancies in data
2. show on factorial display the association structures (multiple associations)
3. reduce the computational costs of upcoming frames processing through supplementary projection of new data

MCA of binary data stream: steps

Multiple correspondence analysis is a suitable MDA technique aiming to visually represent the deviation from the independence condition of a finite set of categorical variables observed with respect to n statistical units.

1. *step1*: determination of the starting (reference) situation through MCA of the matrix $\tilde{\mathbf{Z}}$, with $\tilde{\mathbf{Z}}$ being the disjunctive coded version of $\mathbf{Z}$.
2. *step2*: supplementary projection of the upcoming data on the factorial display obtained in *step1*. This step is repeated at each new frame
3. *step3*: after a fixed number of time-frames user can choose whether to update the starting situation, that is to repete *step1* once and keep going with the procedure

Key-elements: quantification

MCA of binary data stream: computations

- MCA is a Correspondence Analysis of a Burt table given by

$$\mathbf{B} = \tilde{\mathbf{Z}}^{\mathsf{T}}\tilde{\mathbf{Z}}$$

with $\mathbf{P} = \frac{\mathbf{B}}{\textit{grand total}}$ being the correspondence matrix, with row and column margins denoted by $\mathbf{r}$ and $\mathbf{c}$, respectively.

- A reduced rank approximation of $\mathbf{P}$ is given by the SVD of its centered version $\mathbf{Q}$, with general element

$$\mathbf{Q} = \{q_{ij}\} = \frac{(p_{ij} - r_i r_j)}{\sqrt{r_i r_j}}.$$

MCA of binary data stream: computations (2)

- the singular value decomposition of $\mathbf{Q}$ is

$$\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}},$$

with $\mathbf{U}$ and $\mathbf{\Lambda}$ the aigenvector and eigenvalue matrices.

- the *principal co-ordinate* of the i^{th} row (column) point on the s^{th} dimension is obtained through

$$f_{is} = a_{is}\lambda_s,$$

with a_{is} being the corresponding *standard co-ordinate*, that is $a_{is} = \frac{u_{is}}{\sqrt{r_i}}$, λ_s being the s^{th} eigenvalue and u_{is} being the i^{th} element of the corresponding eigenvector.

Key-elements: visualization

Graphical display

- It permits to visualize the association (correspondence) within a set of attributes and the difference within a set of records in terms of distance
- A set of methods that permit to reduce the dimension of a data matrix with respect to a least-squares criterion
- It permits to linearly combine a set of variables into a subset of latent variables (factors)
- Factors are constrained to be orthogonal

Supplementary projection of upcoming streams

Computations for new data

New streams are projected on the farctorial map as supplementary information. In particular, we define $\mathbf{Z}^*$ to be the indicator matrix of the data streams of the new time-frame. There are two possible computations to obtain the supplementary coordinates for the new streams (Nenadic and Greenacre, 2006):

- the position of supplementary items profiles is obtained through a weighted average of the standard coordinates of the row-points
- the position of supplementary items profiles is obtained through a weighted average of the principal coordinates of the row-points

Supplementary projection of upcoming streams

Computation based on the indicator $\mathbf{Z}^*$

- the position of the j^{th} new item coordinate on the s^{th} dimension is

$$f_{js}^* = \sum_{i=1}^{n} \frac{z_{ij}^*}{z_{.j}^*} a_{is}$$

with $z_{.j}^*$ being the column mass of the supplementary item.

Supplementary projection of upcoming streams

Computation based on the Burt matrix

- consider $\mathbf{C}^* = \tilde{\mathbf{Z}}^\mathsf{T}\tilde{\mathbf{Z}}^*$ to be the cross-tabulation between new streams and starting streams; the position of the j^{th} new item coordinate on the s^{th} dimension is

$$\tilde{f}_{js} = \sum_{i=1}^{p} \frac{c^*_{ij}}{c^*_{\cdot j}} \tilde{a}_{is}$$

with $c^*_{\cdot j}$ being the column mass of the supplementary item and $\tilde{a}_{is}$ being the standard coordinate referring to the corresponding active row of the Burt matrix.

Example of application on synthetic data

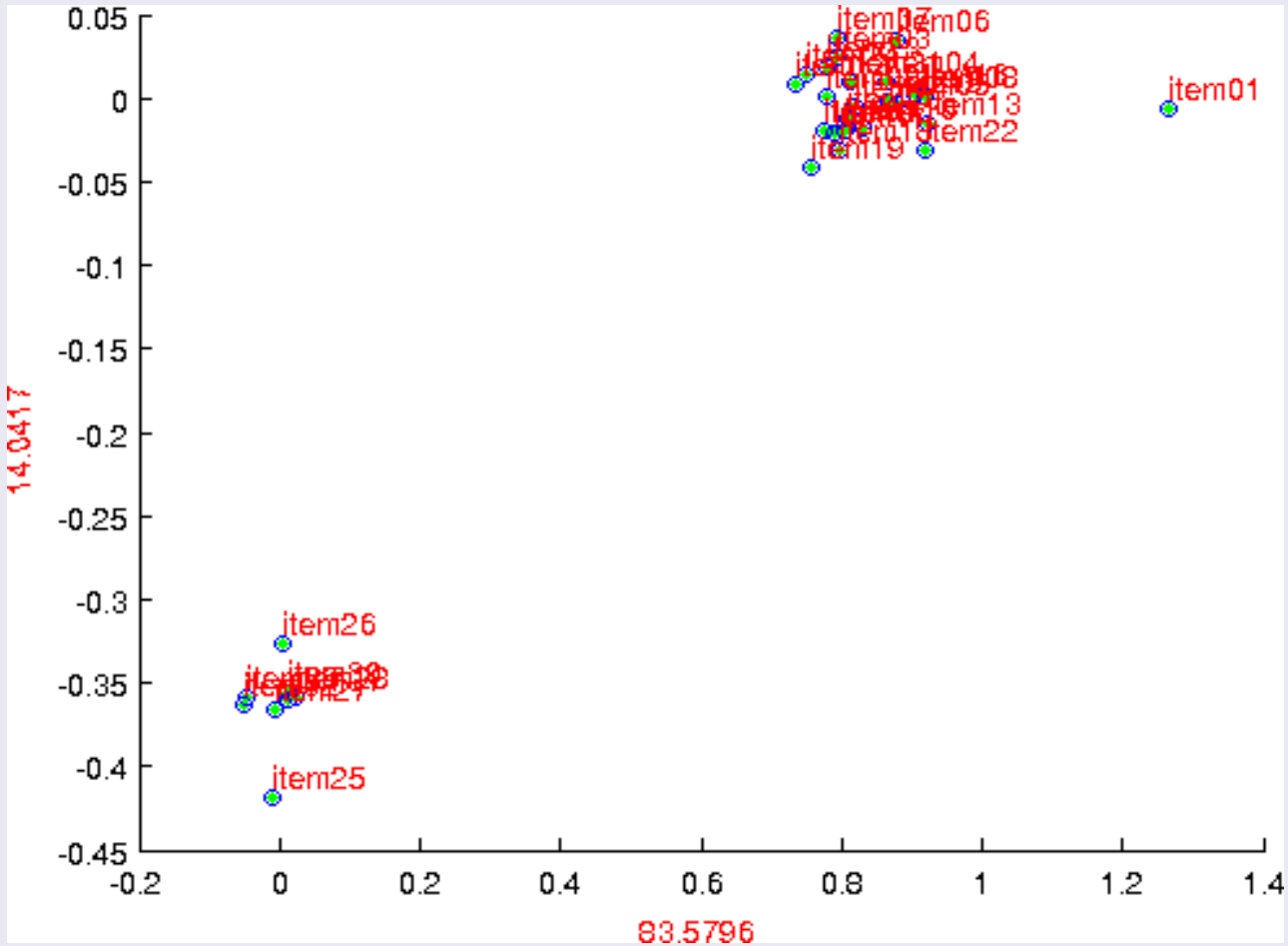
Simulated data

We generated a data set of $n = 500$ binary streams described by $p = 32$ items. There is a strong assciation structure between two blocks of items: the first 24 items for the first block, the remaining 8 items in the second block. Then consider eight consecutive time frames structured as follows:

- data of the first four time frames are generated according to the same association structure
- at each following time frame such association structure is modified by swapping two items from one block to another: that is to swap two binary columns of $\mathbf{Z}^*$ more for each new frame.

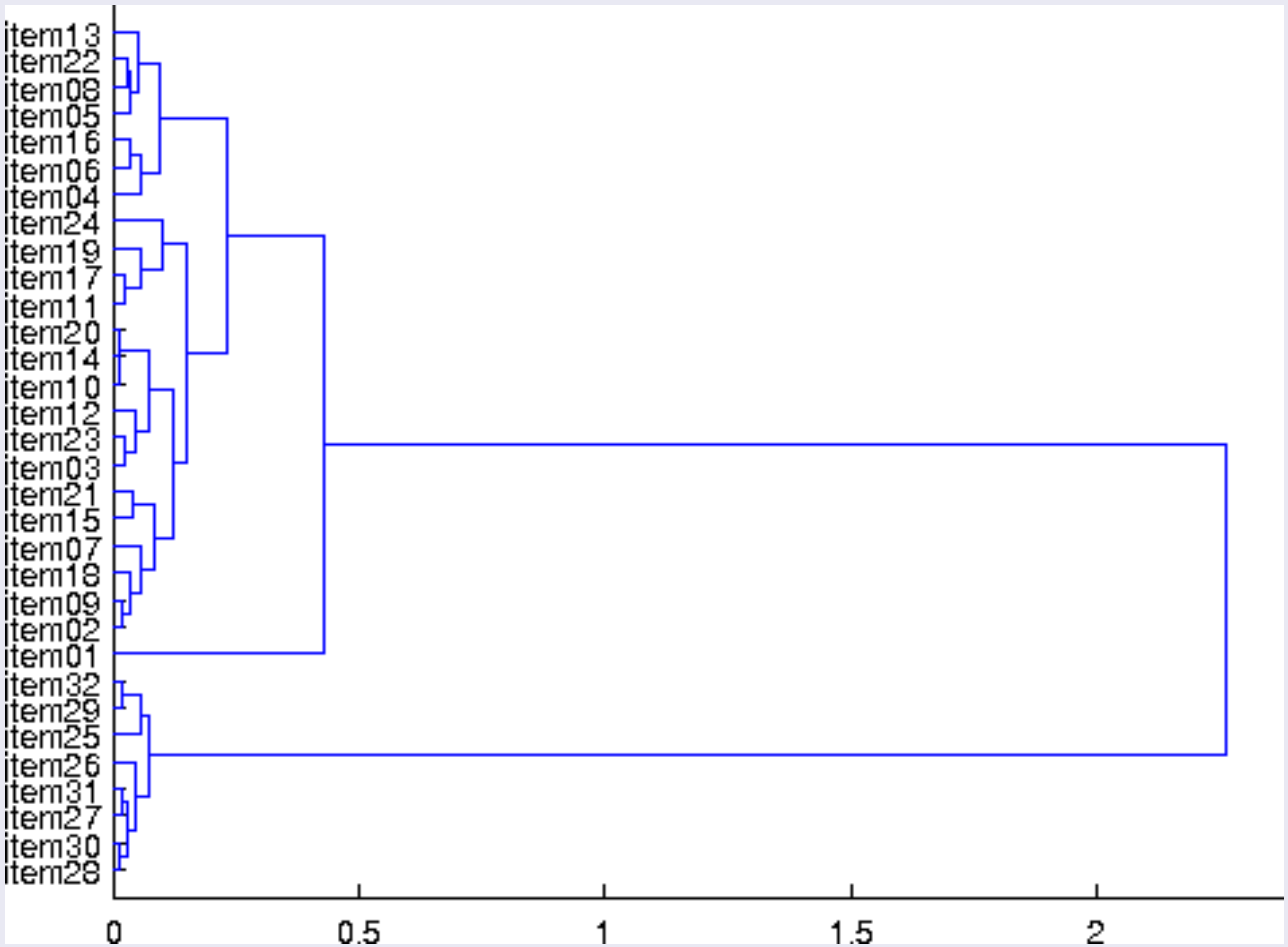
Example of application on synthetic data

Starting association structure



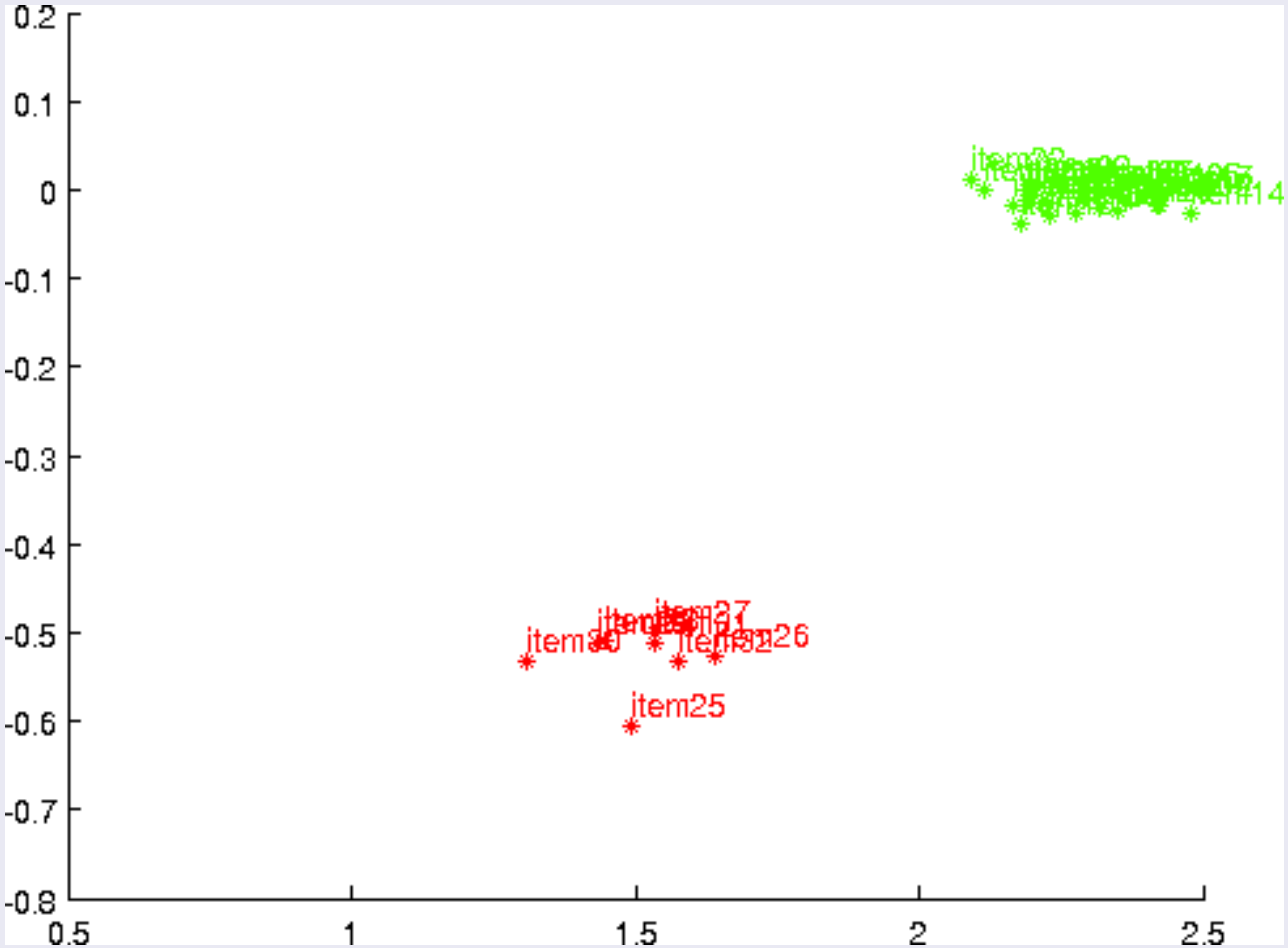
Example of application on synthetic data

Dendrogram of items



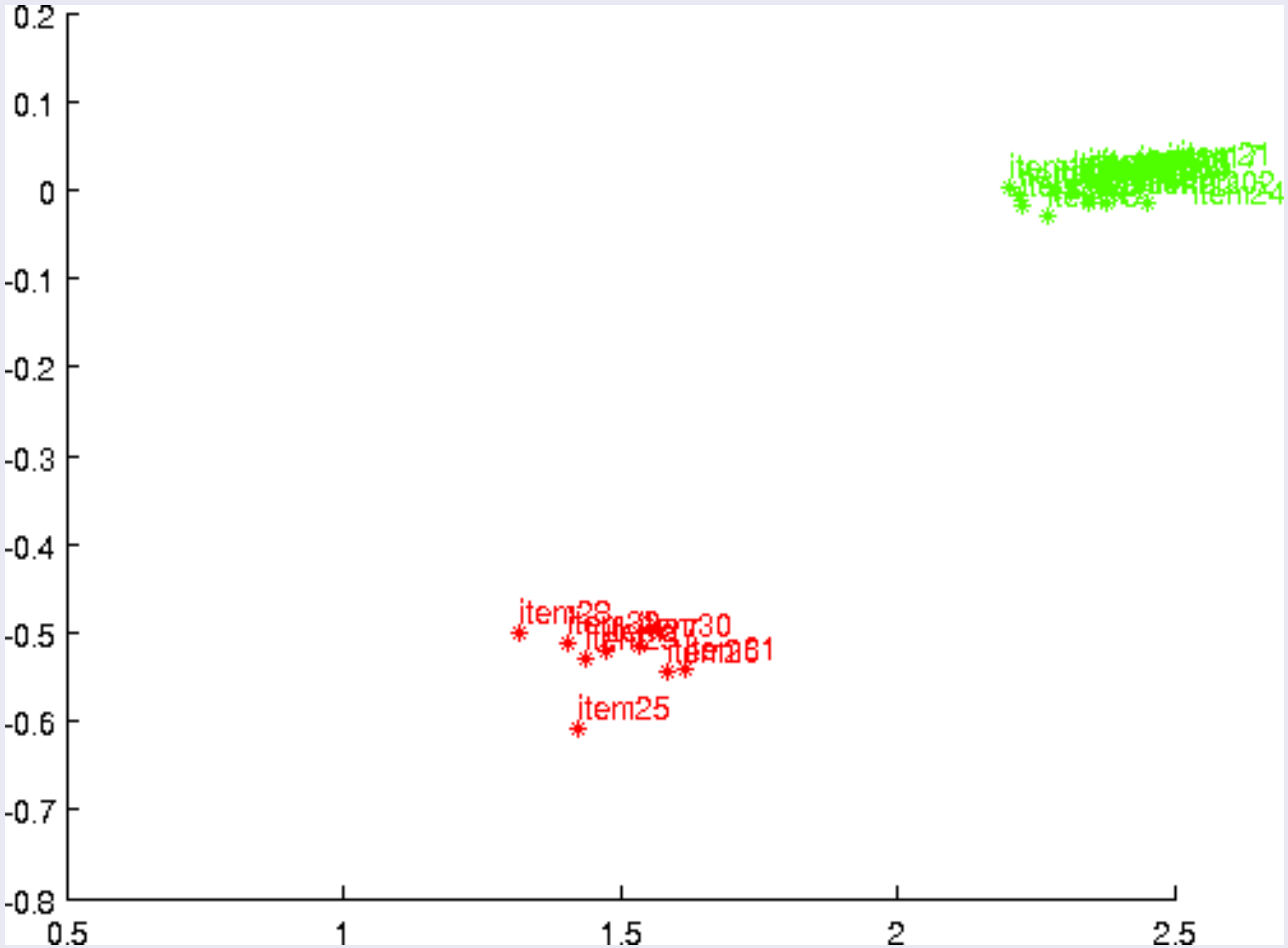
Example of application on synthetic data

T1 association structure



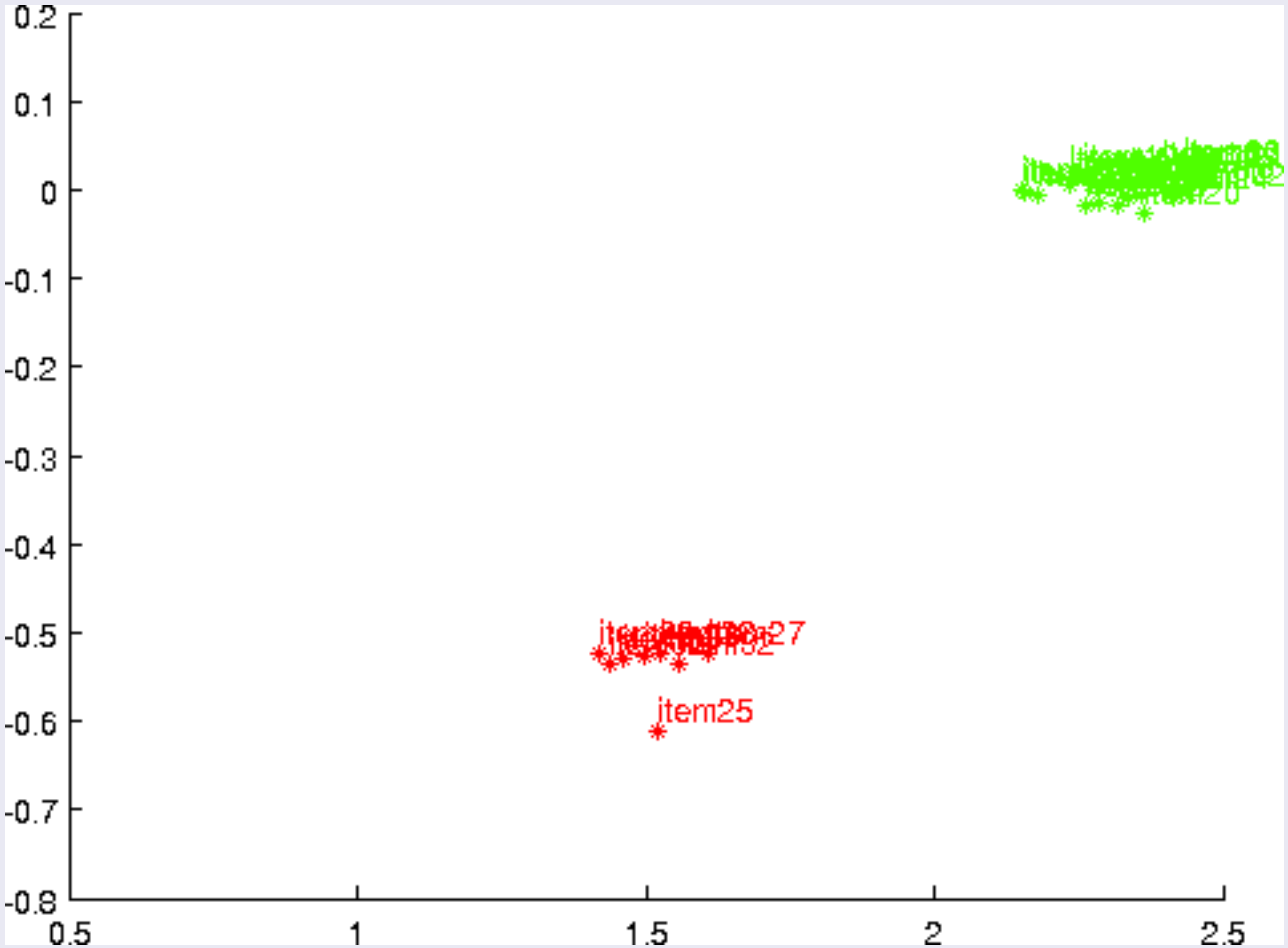
Example of application on synthetic data

T2 association structure



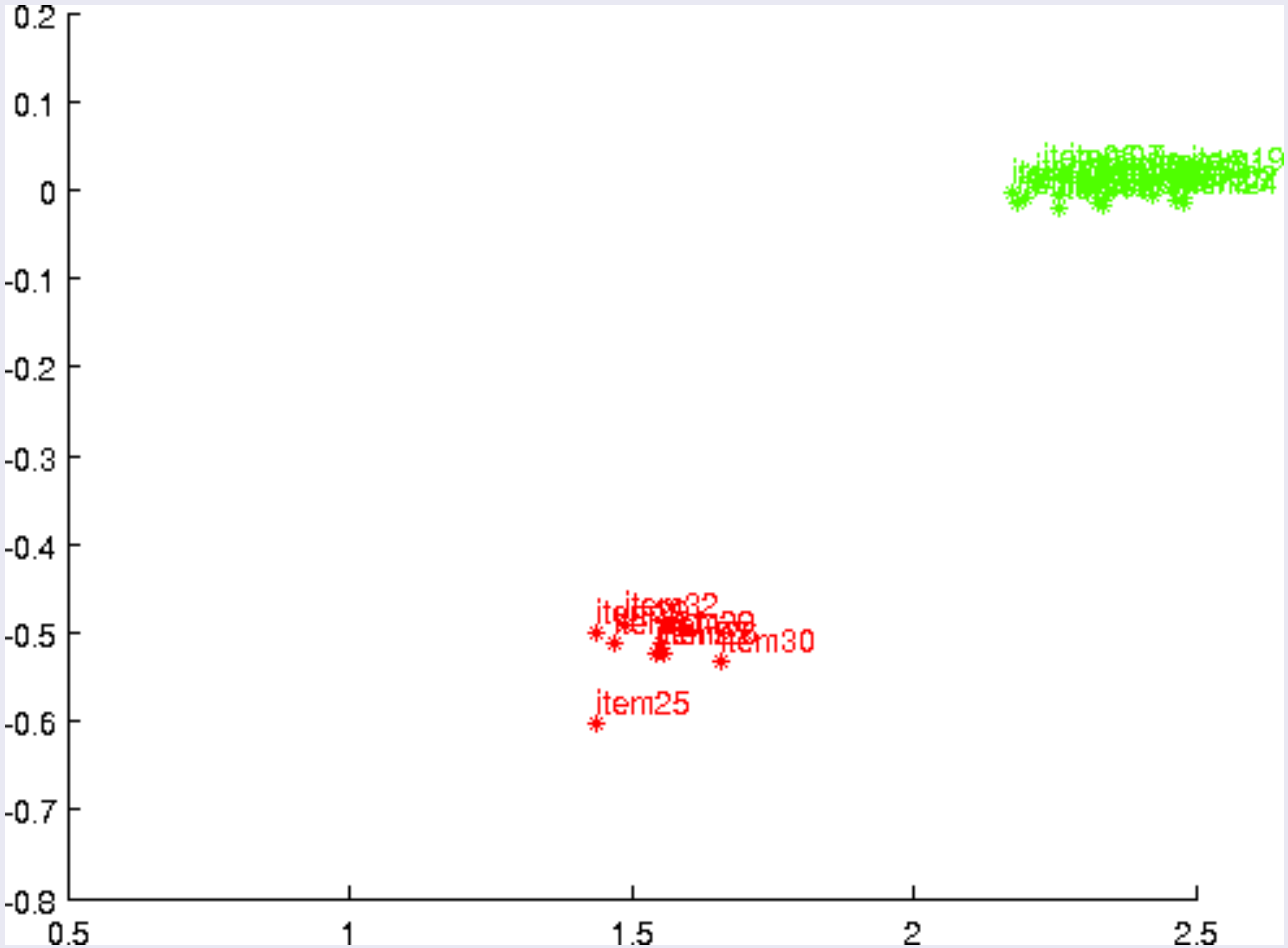
Example of application on synthetic data

T3 association structure



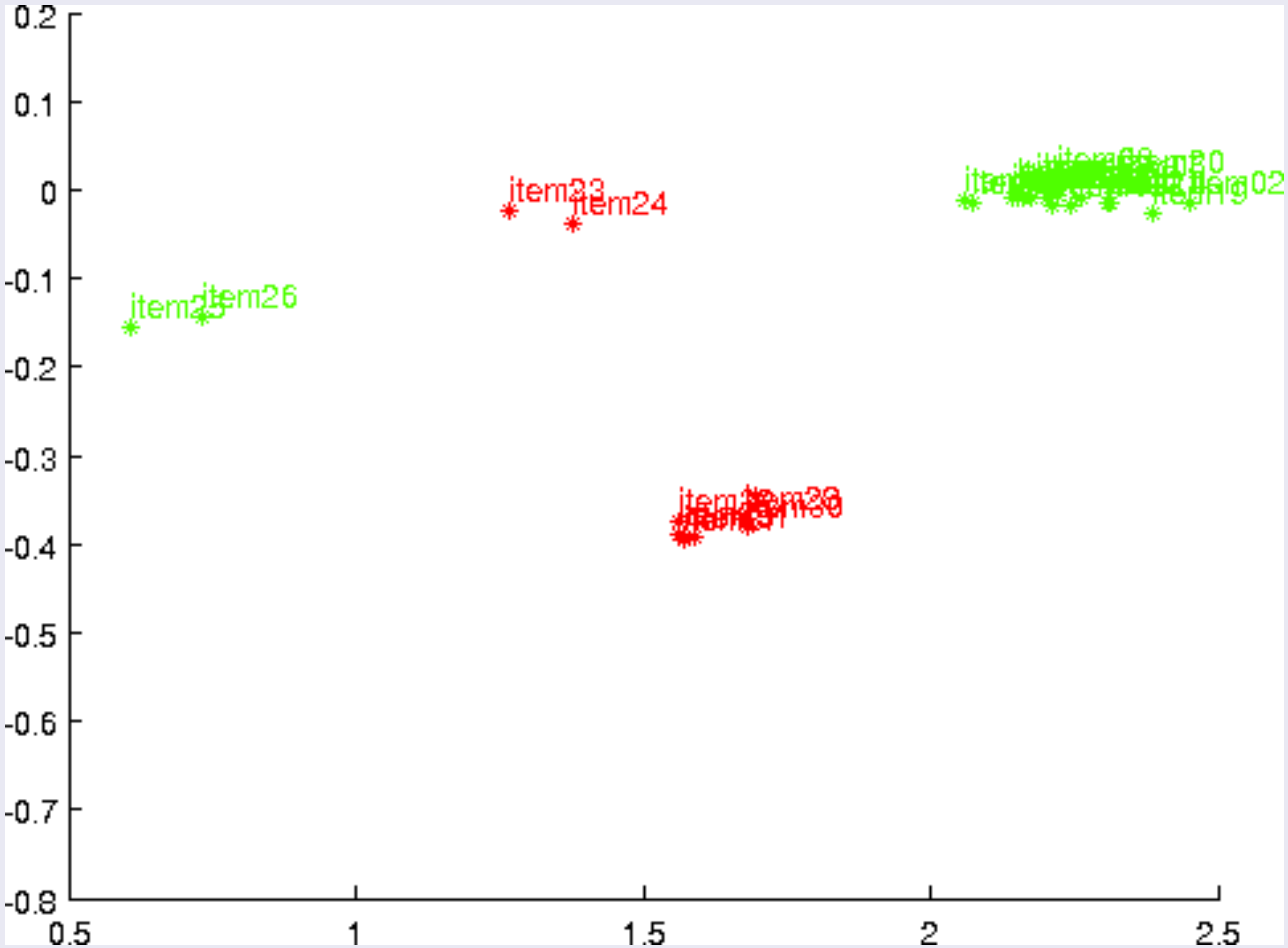
Example of application on synthetic data

T4 association structure



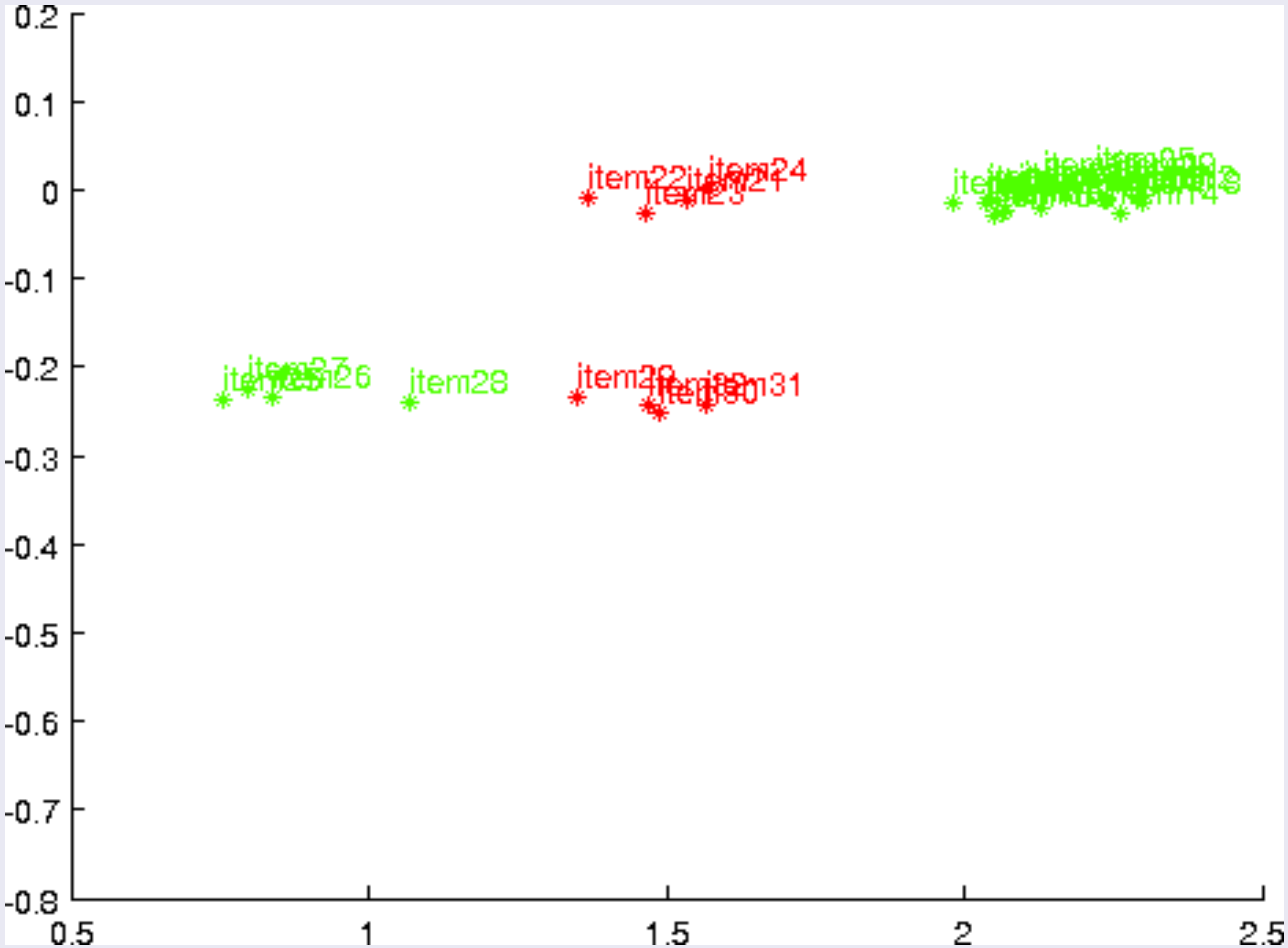
Example of application on synthetic data

T5 association structure



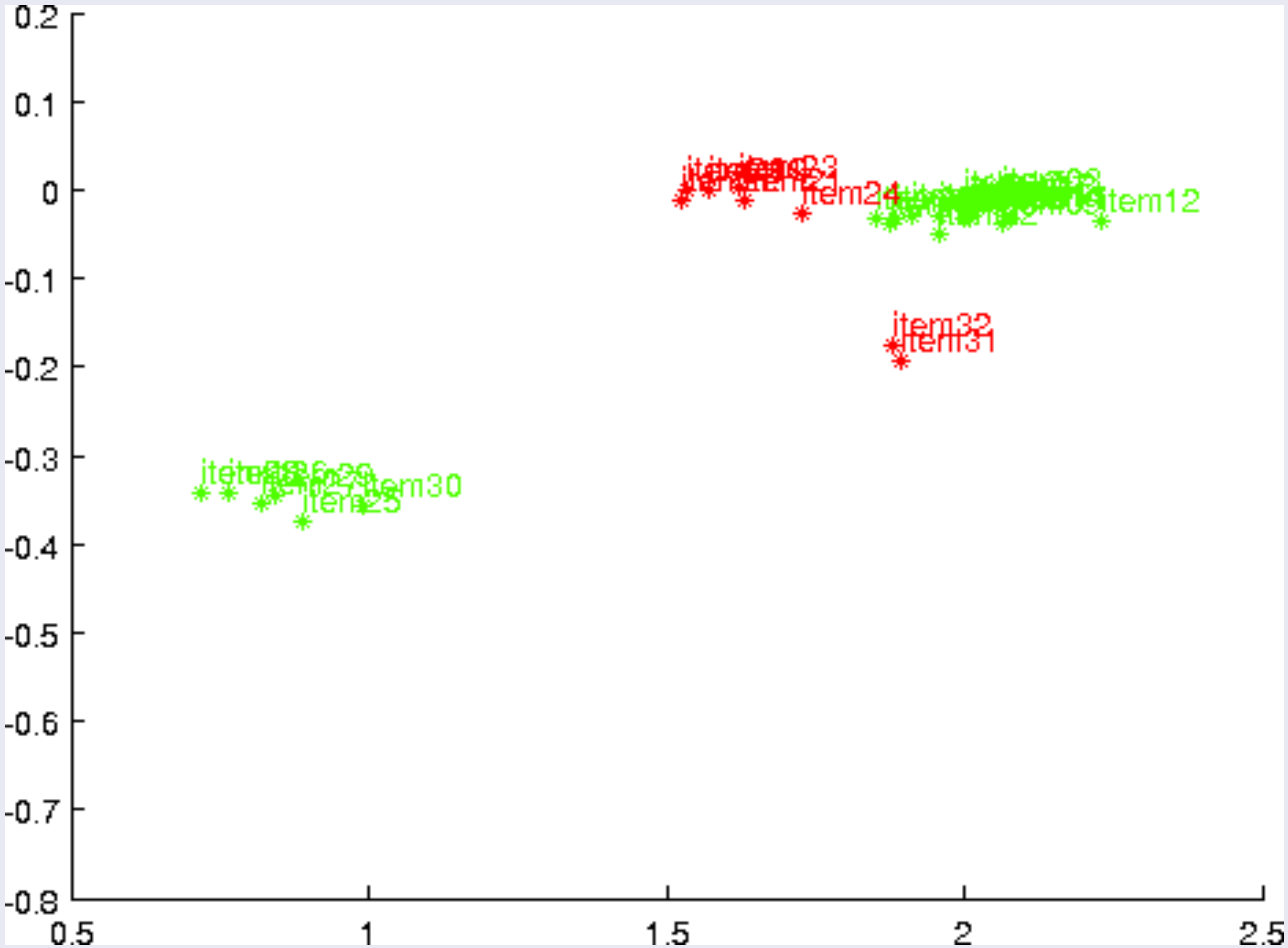
Example of application on synthetic data

T6 association structure



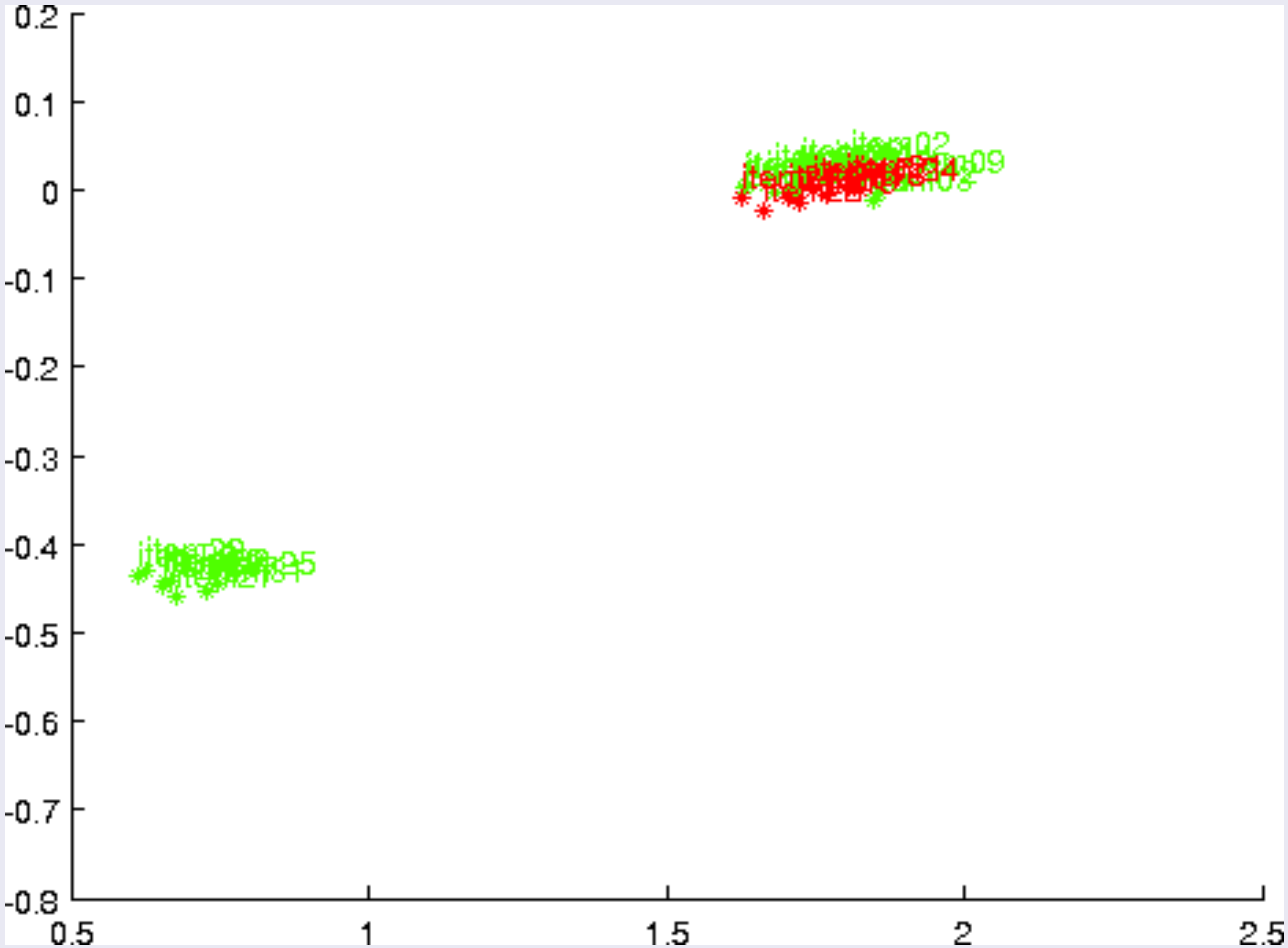
Example of application on synthetic data

T7 association structure



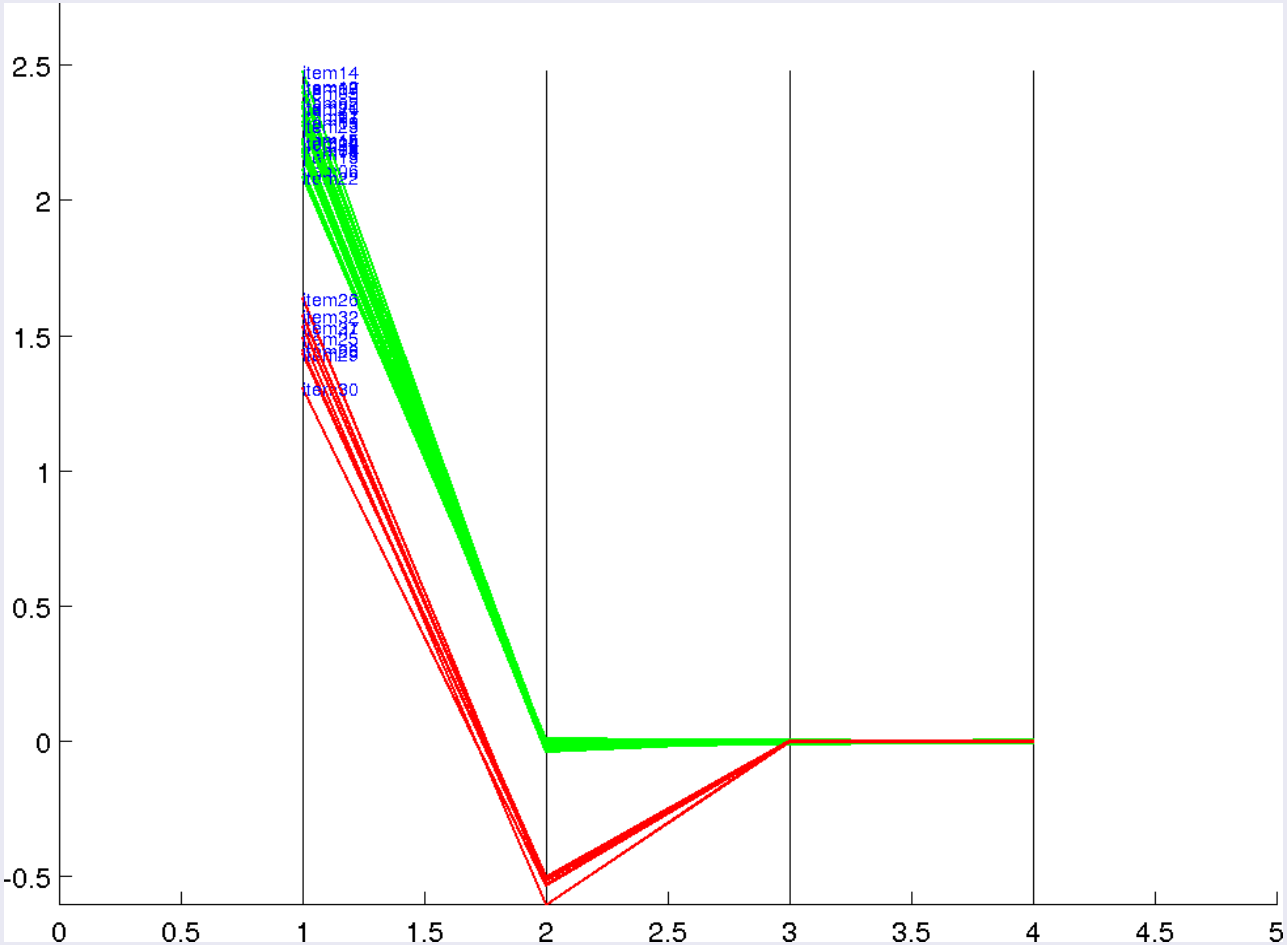
Example of application on synthetic data

T8 association structure



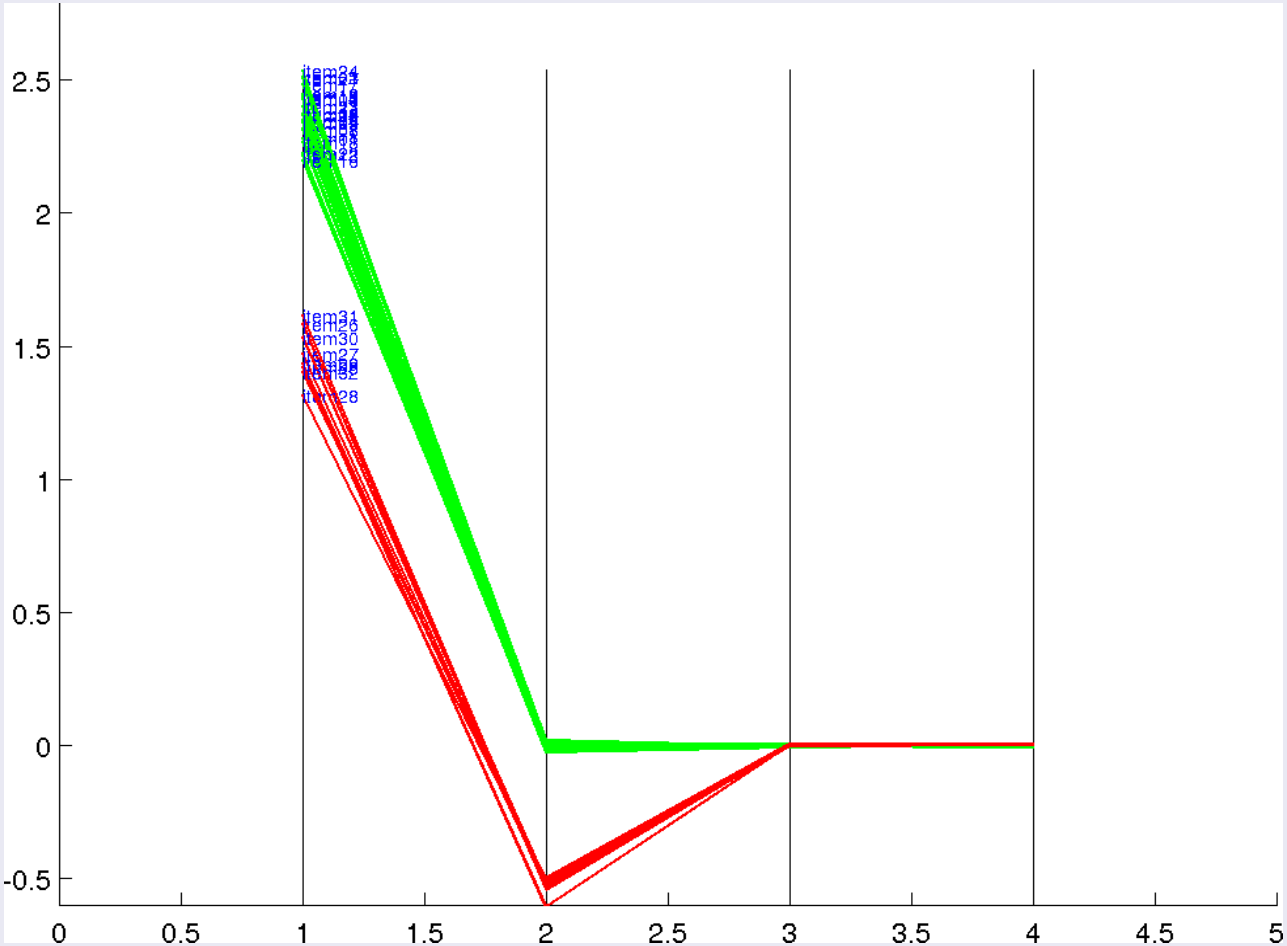
Example of application on synthetic data

Parallel coordinates visualization of T1



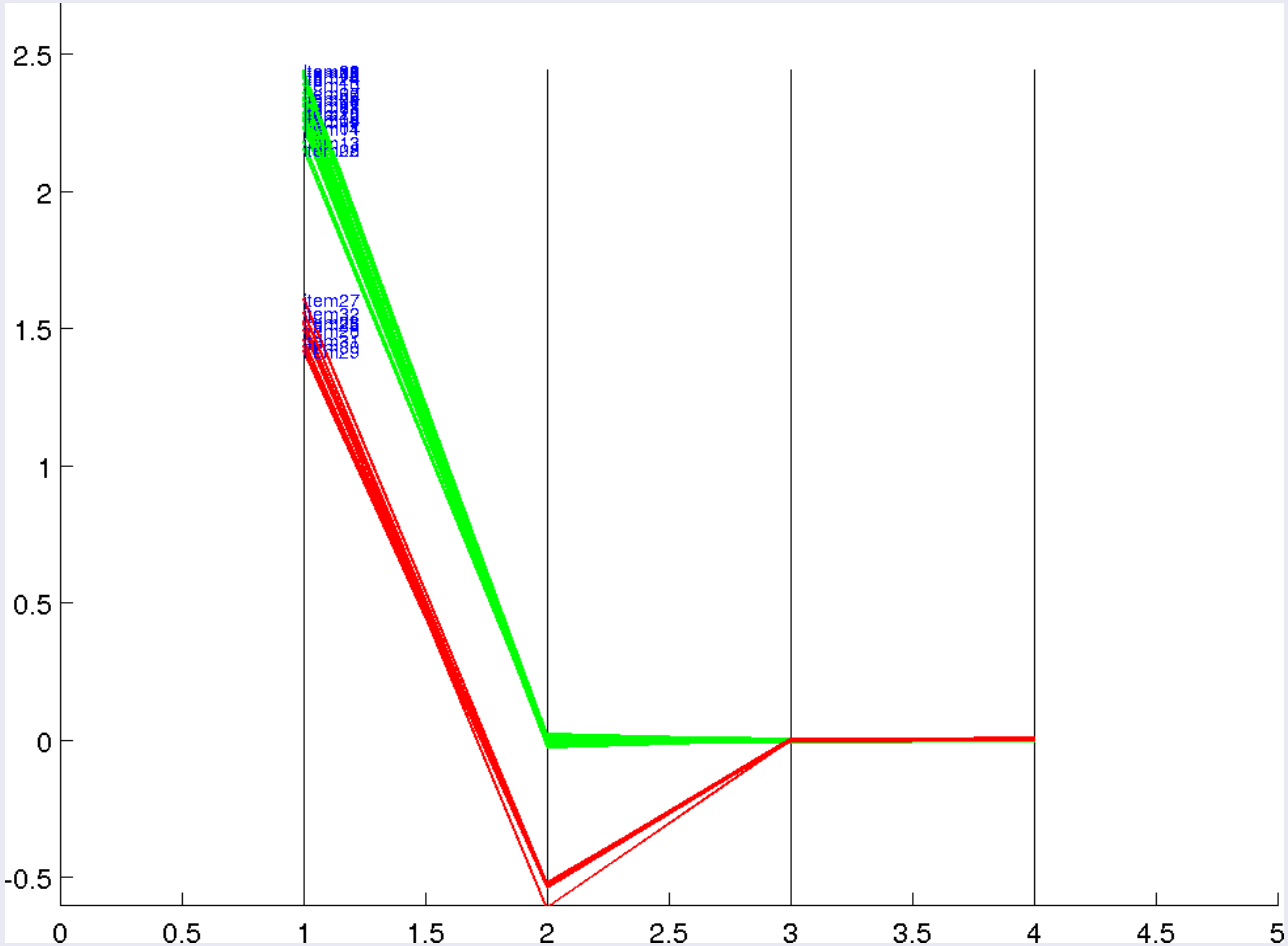
Example of application on synthetic data

Parallel coordinates visualization of T2



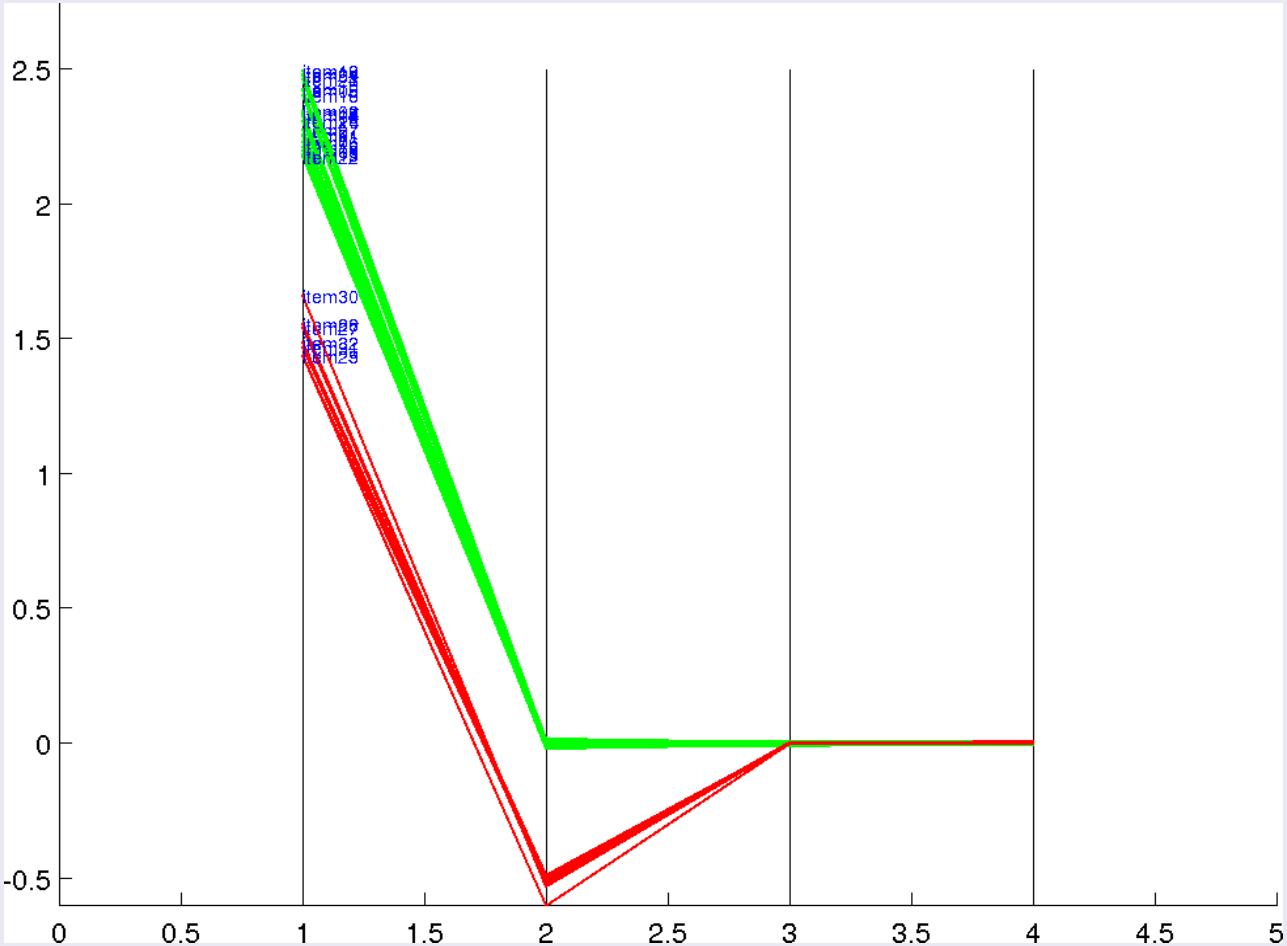
Example of application on synthetic data

Parallel coordinates visualization of T3



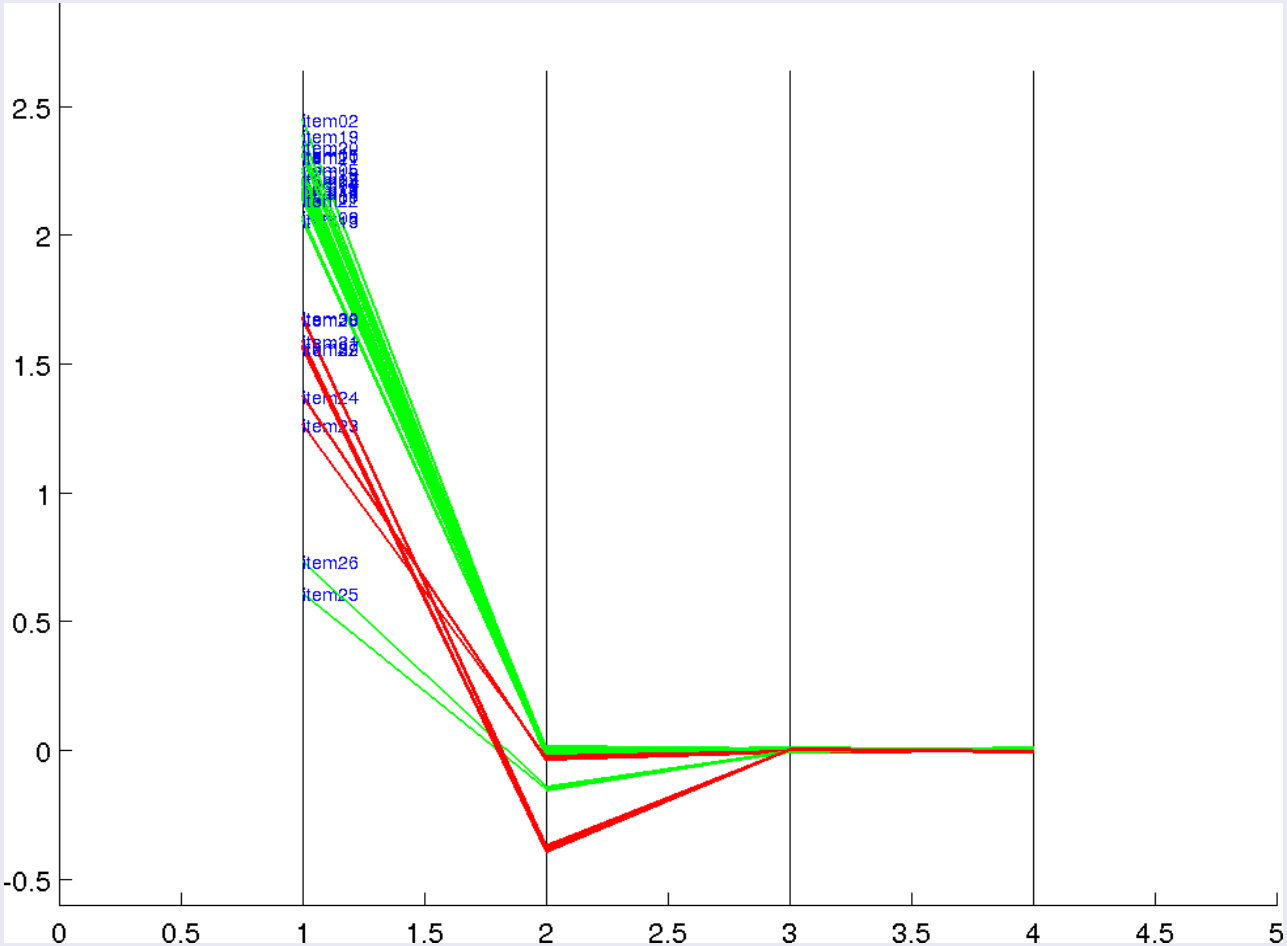
Example of application on synthetic data

Parallel coordinates visualization of T4



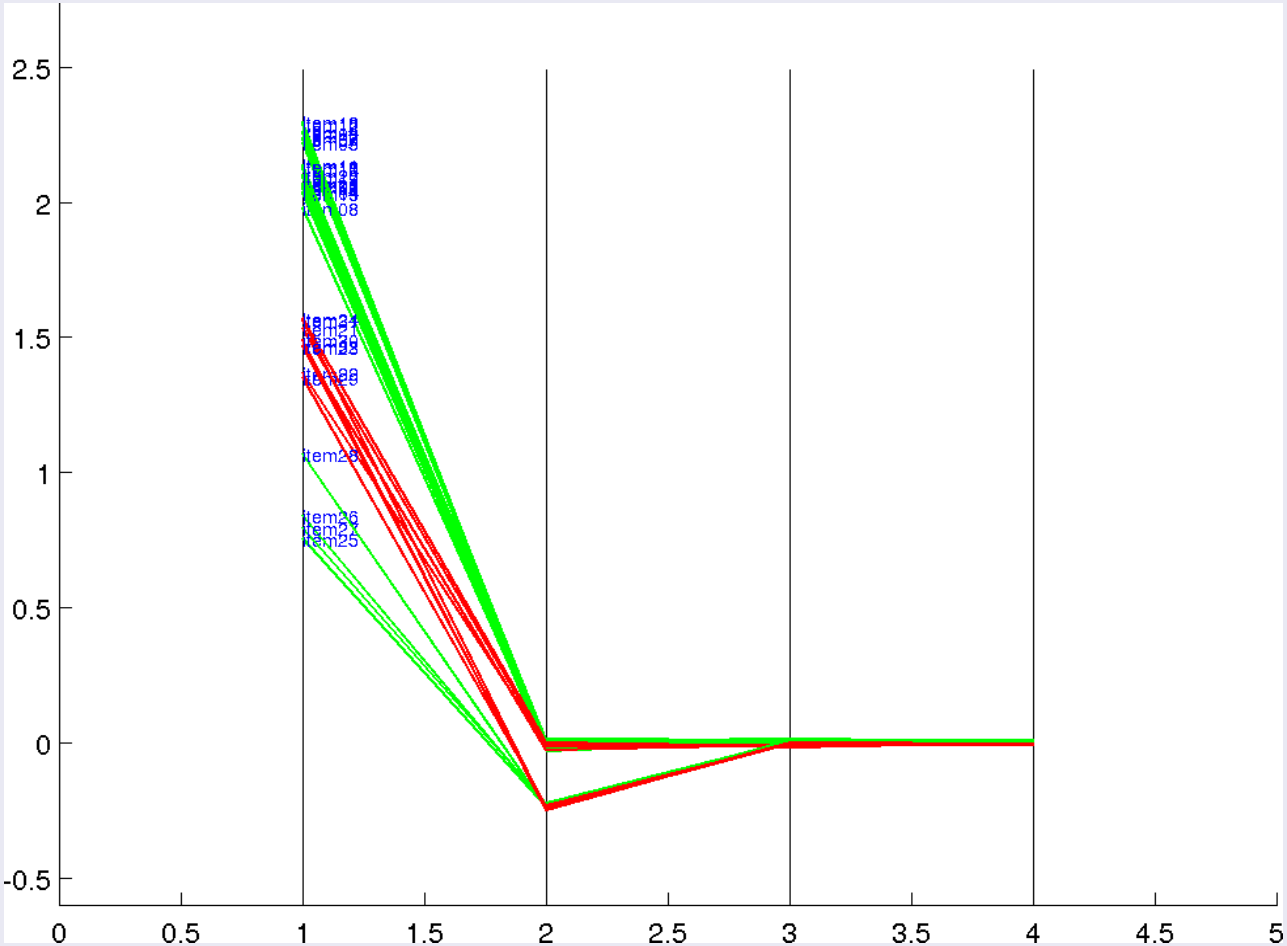
Example of application on synthetic data

Parallel coordinates visualization of T5



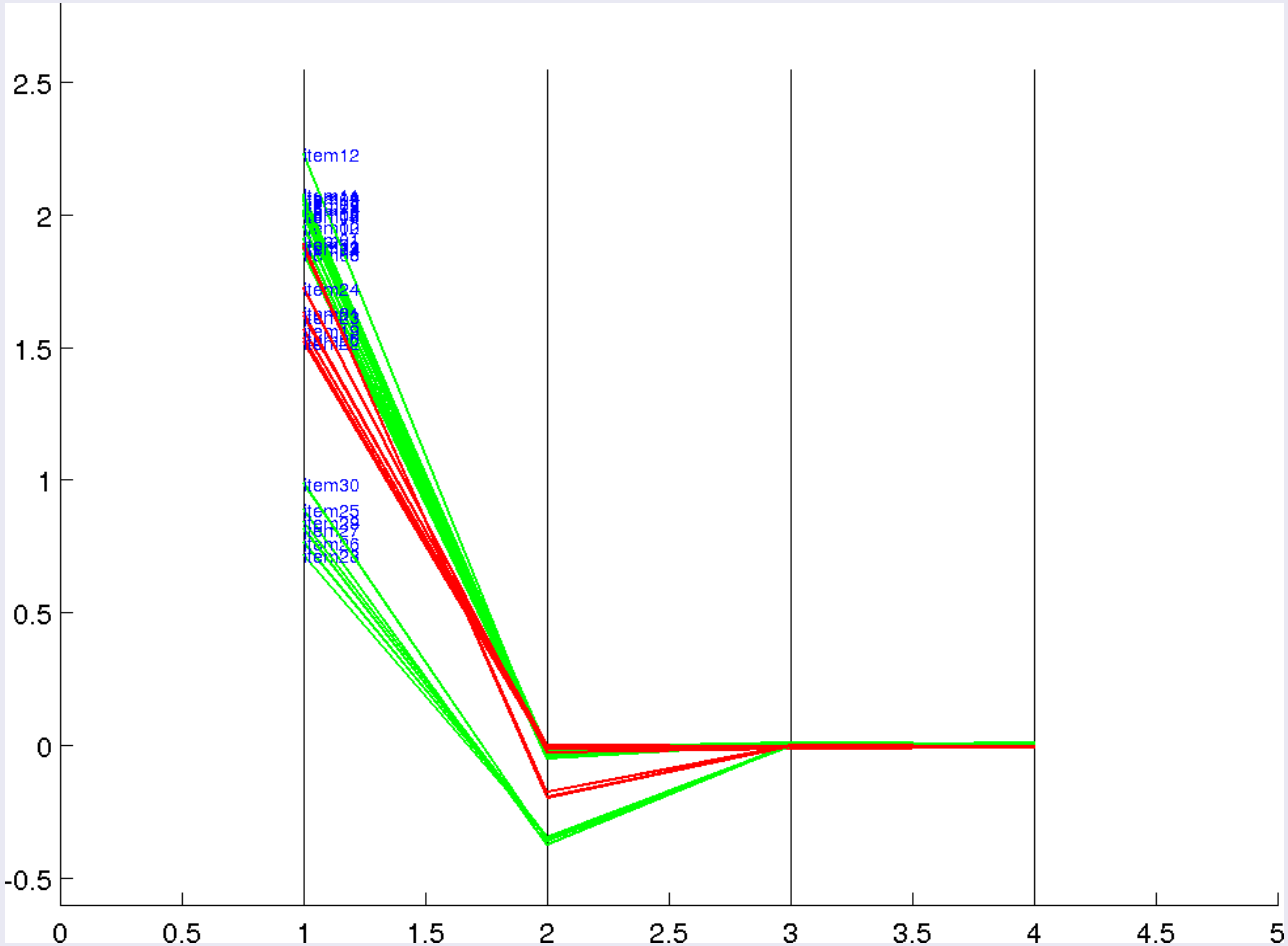
Example of application on synthetic data

Parallel coordinates visualization of T6



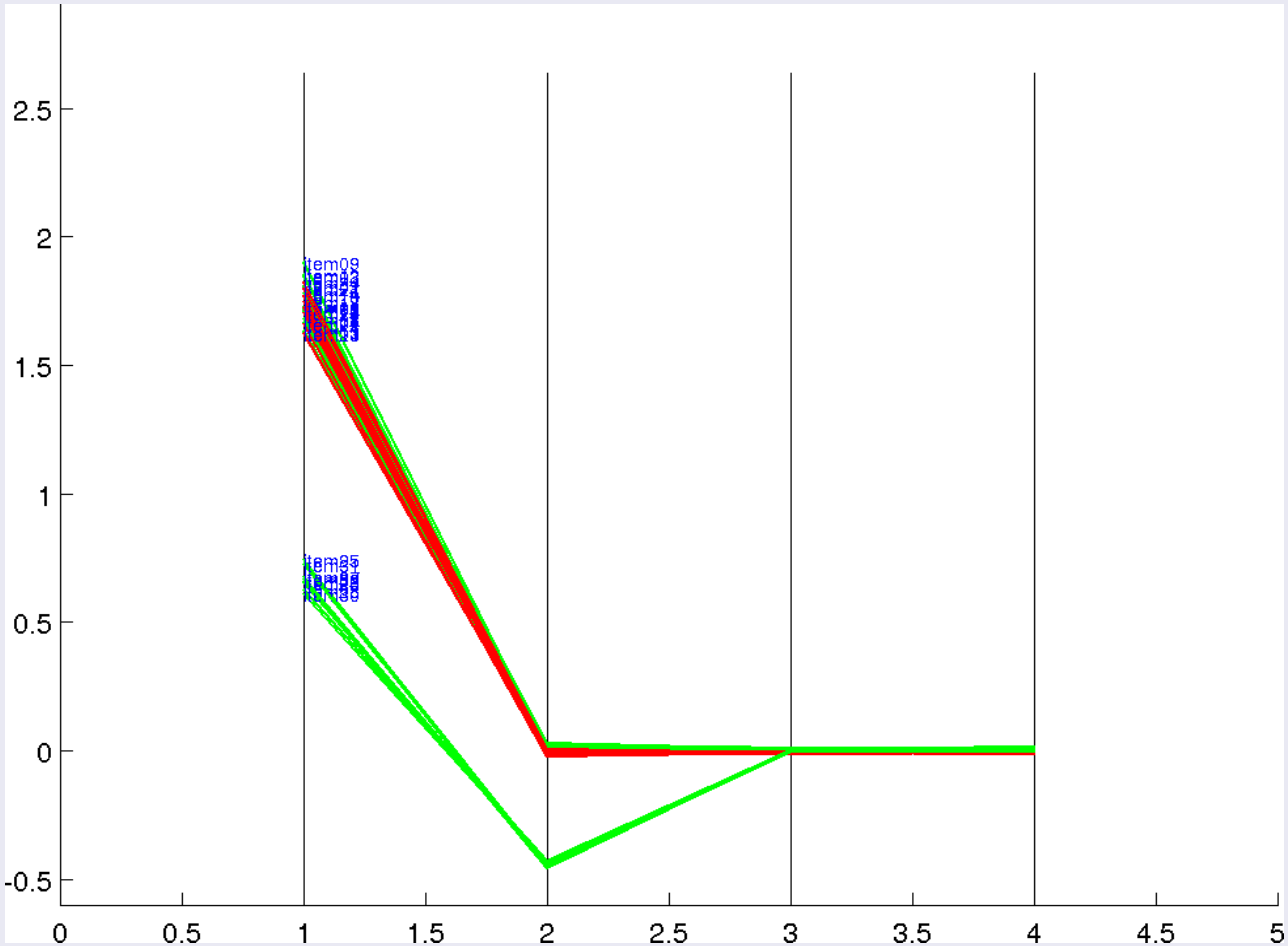
Example of application on synthetic data

Parallel coordinates visualization of T7



Example of application on synthetic data

Parallel coordinates visualization of T8



For Further Reading I


J.-P. Benzècri. Historie et prèhistorie de l'analyse des données. *Callieurs de l'Analyse des Données*, 1, 1976


Benzécri, J. P.: 1979, Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, *Le Cahiers de l'Analyse des Données* **4**(4), 377–378.


Gaber, M. M., Zaslavsky, A. and Krishnaswamy, S.: 2007, *Data Streams: Models and Algorithms*, Springer Verlag, chapter A Survey of Classification Methods in Data Streams.


Greenacre, M. J.: 1984, *Theory and Application of Correspondence Analysis*, Academic Press, London.


Nenadic, O. and Greenacre, M. J.: 2006, *Computations of multiple correspondence analysis with code in R*.


Iodice D'Enza, A.: 2006, *Exploratory Study of Association in Transaction Data Bases*, PhD thesis, Dipartimento di Matematica e Statistica Universitá degli Studi di Napoli Federico II, Napoli.


Iodice D'Enza, A., Palumbo, F. and Greenacre, M.: 2006, Exploratory data analysis leading towards the most interesting simple association rules, *Computational Statistics and Data Analysis* **Accepted, in press.**


Lebart, L., Morineau, A. and Piron, M.: 1995, *Statistique exploratorie multidimensionelle*, Dunod, Paris.


Muthukrishnan, S.: 2003, Data streams: algorithms and applications, ACM-SIAM Symposium on Discrete Algorithms.
citeseer.ist.psu.edu/article/muthukrishnan03data.html.