TÉCNICO+
FORMAÇÃO AVANÇADA

Subspace clustering for pattern discovery

DASH: Data Science e Análise Não Supervisionada

Rui Henriques, rmch@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa

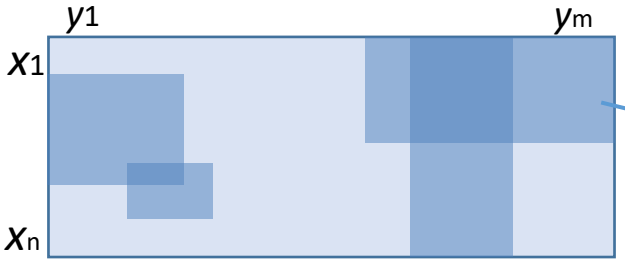
Outline

- Subspace clustering
- Biclustering
  - coherence
  - quality
  - structure
  - evaluation
  - searches
- Triclustering
- Deep learning
- Appendix

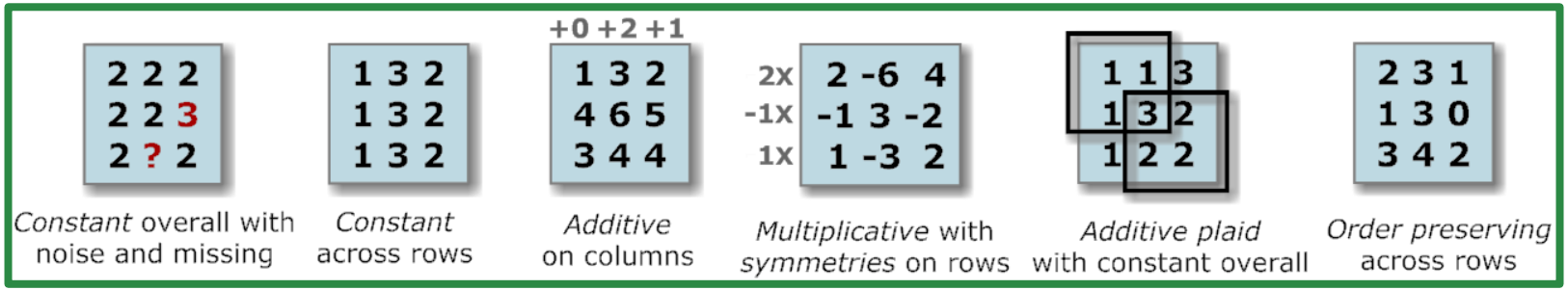
Why subspace clustering?

- Find **patterns on real-valued data**
  - classic pattern mining (e.g. FIM, ARM) suffer from discretization drawbacks
- Find **less-trivial patterns** with non-constant homogeneity
  - classic pattern mining is only able to find constant patterns (i.e. simple repititions)
- Other applications
  - well-established role in **predictive tasks** using discriminative patterns
  - **dimensionality reduction** by reducing data into a set of informative or/and discriminative subspaces
  - **imputation** of missings taking into consideration a subspace's homogeneity
  - ...

Motivation



homogeneous and statistically significant



Motivation

- **GLOBAL stance: clustering**
  - Group observations correlated according to values of all variables
  - **Problem?**
    - High-dimensional data: hundreds to thousands of variables

      Similarity on all variables can be misleading if strong correlation occurs only on a subset

      x1 = 0.2 0.1 0.2 0.1 **0.4 0.7 0.4** 0.3

      x2 = 0.1 0.3 0.1 0.2 **0.4 0.7 0.4** 0.2

      x3 = 0.3 0.1 0.2 0.3 **0.4 0.7 0.4** 0.1

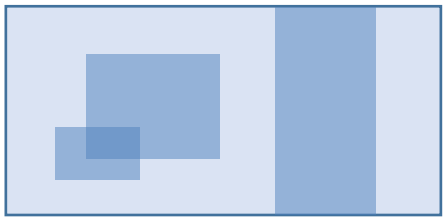
      Dissimilar? What about this coherent pattern

- **Solution? LOCAL stance: subspace clustering**
  - Group observations correlated on a subset of all variables (subspace)
  - Non-exhaustive and overlapping groups of observations

Definition

- Subspace clustering can be applied to different data structures, including:
  - [***biclustering***] simple multivariate data
  - [*n-way subspace clustering*] tensor data (e.g. ***triclustering*** for three-way data)

- Given a dataset, D, with a set of observations $X = \{\mathbf{x}_1, .., \mathbf{x}_N\}$, a set of variables $Y = \{y_1, .., y_M\}$, and elements $a_{ij} \in R$ relating observation x_i and variable y_j:
  - A **bicluster** $B = (I, J)$ defines a pattern in D, where $I = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in}) \subset X$ is a subset of observations (*pattern support*) and $J = (yj_1, \ldots, y_{jm}) \subset Y$ is a subset of variables (*subspace*)
  - The **biclustering task** aims to identify a set of biclusters $\boldsymbol{B} = \{B_1, .., Bs\}$ such that each bicluster $B_k = (I_k, J_k)$ satisfies specific criteria of *homogeneity* and *statistical significance*



Applications: social domains

- **Social networks**
  - communities of individuals with shared interests or correlated activity (X=Y=individuals)
  - aggregation of contents (X) based on correlated accessors' profile, comments and tags (Y)
- **Text data:** group content-related documents to support searches, suggestions and tagging (X=documents, Y=features)
- **(e-)commerce:** browsing patterns (X=users, Y=webpage accesses)
- **Education**: performance analysis (X=students/professors, Y=topics/features)
- **Financial/trading:** subsets of indicators producing similar profitability for subsets of trading points (X=buy and sell signals, Y=stock market ratios)
- **Collaborative filtering data:** groups of users with similar rating patterns and behavior on a subset of available actions (X=users, Y=items/actions)

Applications: biomedicine

- **Omic data:** functional processes and pathways (X=genes/proteins/metabolites, Y=conditions)
- **Physiological data:** coherent sliding features on a subset of stimuli-elicited responses; groups of patients with shared local patterns (X=signals, Y=features)
- **Clinical data:** health trends and risk profiles from health records; individuals with similar treatments, diagnoses and tests (X=individuals, Y=clinical features)
- **Genomic mutations:** correlated mutations (Y) for specific populations (X)
- **Biological networks:** modules of genes, proteins or metabolites (X=Y=biological entities) with cohesive local interaction using adjency matrices

Motivation

Consider **gene expression** data analysis (where X=samples/individuals and Y=genes)

- ***clustering*** groups samples or genes ⇒ limited relevance! **Why?**
  - only a small set of the genes participates in a cellular process of interest
  - an interesting cellular process is active only in a subset of the conditions
  - a single gene may participate in multiple pathways that may or not be coactive under all conditions
- ***biclustering*** groups genes that show similar activity patterns under a subset of samples => **current way of extracting new knowledge**
  - SARS-CoV-2 knowledge advances on regulatory responses and vaccines
  - breakthroughs on cancer mechanisms and therapies

Outline

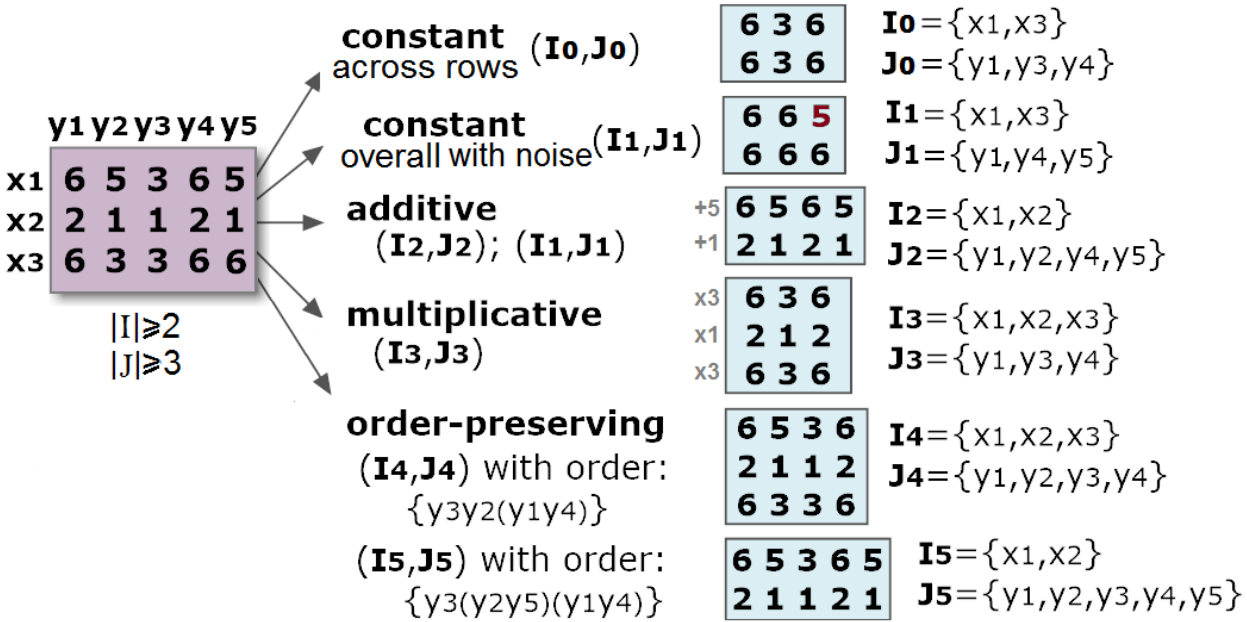
- Subspace clustering
- **Biclustering**
  - **coherence**
  - **quality**
  - **structure**
  - **evaluation**
  - **searches**
- Triclustering
- Deep learning
- Appendix

Homogeneity

- Let us recall: the **biclustering task** aims to identify a set of biclusters such that each bicluster satisfies specific criteria of…
  - **homogeneity:** patterns of interest
    - the placed homogeneity determines the ***structure*** (positioning), ***coherence*** (correlation) and ***quality*** (noise tolerance) of biclusters
  - **statistical significance:** non-spurious patterns,
    i.e. biclusters should not occur by chance (unexpectedly frequent)
    - non-significant bicluster discovered: *false positive*
    - significant bicluster not discovered: *false negative*

Coherence

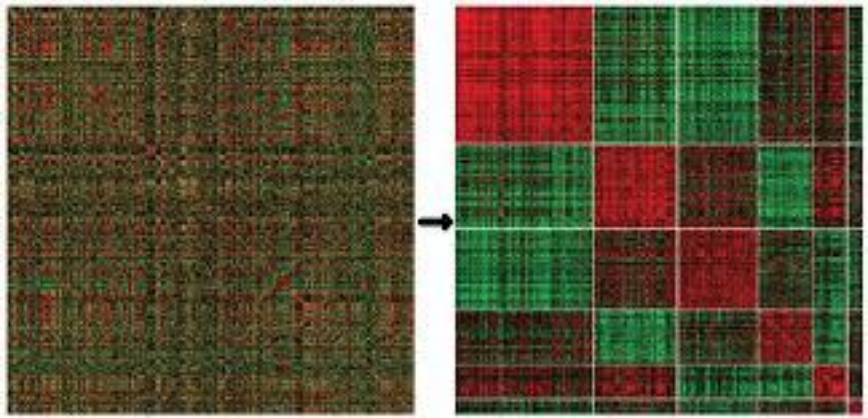
- *The allowed form of* correlation is termed **coherence assumption**



Coherence

- Homogeneity commonly guaranteed through a **merit function**
  - e.g. the variance of the values in a subspace

    Low variance can be used to find biclusters with constant values



Constant model

- **Constant** values
  - **overall** (low-variance)

    $a_{ij} = c + \eta_{ij}$

  - **on variables**

    $a_{ij} = c_j + \eta_{ij}$

  where…
  - c_j is a constant
  - η_{ij} the noise factor
  - $a_{\boldsymbol{IJ}}$ average (I, J)
  - $a_{\boldsymbol{Ij}}$ average $(I, \{yj\})$

1.2	0.5	0.3	**1.3**
1.3	0.5	0.1	**1.2**
1.2	0.6	0.6	**1.1**
1.1	1.3	0.8	**1.2**

$a_{IJ} = 1.2$

$\eta_{11} = .0 \quad \eta_{14} = .1$

$\eta_{21} = .2 \quad \eta_{24} = .0$

$\eta_{31} = .0 \quad \eta_{34} = -.1$

$\eta_{41} = -.1 \quad \eta_{44} = .0$

1.2	**0.5**	0.3	**1.3**
1.3	**0.5**	0.1	**1.2**
1.2	**0.6**	0.6	**1.1**
1.1	1.3	0.8	1.2

$a_{I2} = .53$

$\eta_{12} = .03$

$\eta_{22} = .03$

$\eta_{32} = .07$

Additive model

$$a_{ij} = c_j + \gamma_i + \eta_{ij}$$

- c_j is the value of variable y_j
- γ_i is the adjustment for observation x_i

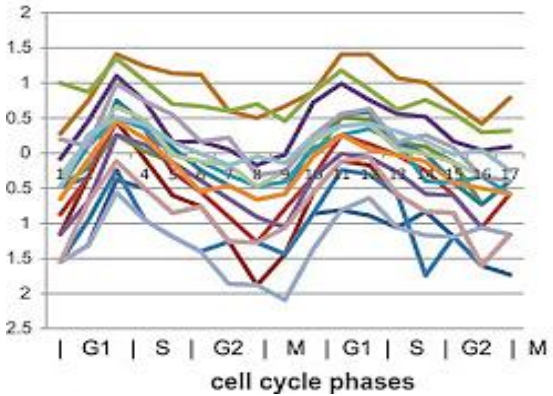
	c_1=1	c_2=3	c_3=2	
γ_1= +0	**1**	**3**	**2**	5
γ_2= +3	**4**	**6**	**5**	1
γ_3= +2	**3**	**5**	**4**	2
	4	1	2	3

- bicluster **pattern** ϕ_B is the expected values in the absence of adjustments γ_i and noise factors η_{ij}
  - e.g. $\phi_B = \{c_1 = 1, c_2 = 3, c_3 = 2\}$

Additive model

Why?

- medical field: handle individual differences, different stages of disease progression
- biological field: responsiveness of genes, experimental differences
- social field: individual differences regarding activity



Multiplicative model

- Similar to the additive model:
  - on observations

    $a_{ij} = c_j \gamma_i + \eta_{ij}$
  - on variables

    $a_{ij} = c_i \gamma_j + \eta_{ij}$

	c_1=1	c_2=-3	c_3=2	
γ_1= 2	**2**	**-6**	**4**	5
γ_2= -1	**-1**	**3**	**2**	1
γ_3= 1	**1**	**-3**	**2**	2
	4	1	2	-3

- c_j (or c_i) is the value of variable y_j (or observation x_i)
- γ_i (or γ_j) is the adjustment for observation x_i (or variable y_j)

Order-preserving model

- A bicluster following an **order-preserving** model is (**I**, **J**) where the values on each observation in **I** across the **J** variables are ordered according the same permutation **π**

	y_1	y_2	y_3	y_4
x_1	**19**	**12**	6	**14**
x_2	**10**	**7**	13	**9**
x_3	**11**	**6**	17	**8**
x_4	**13**	**4**	1	**11**

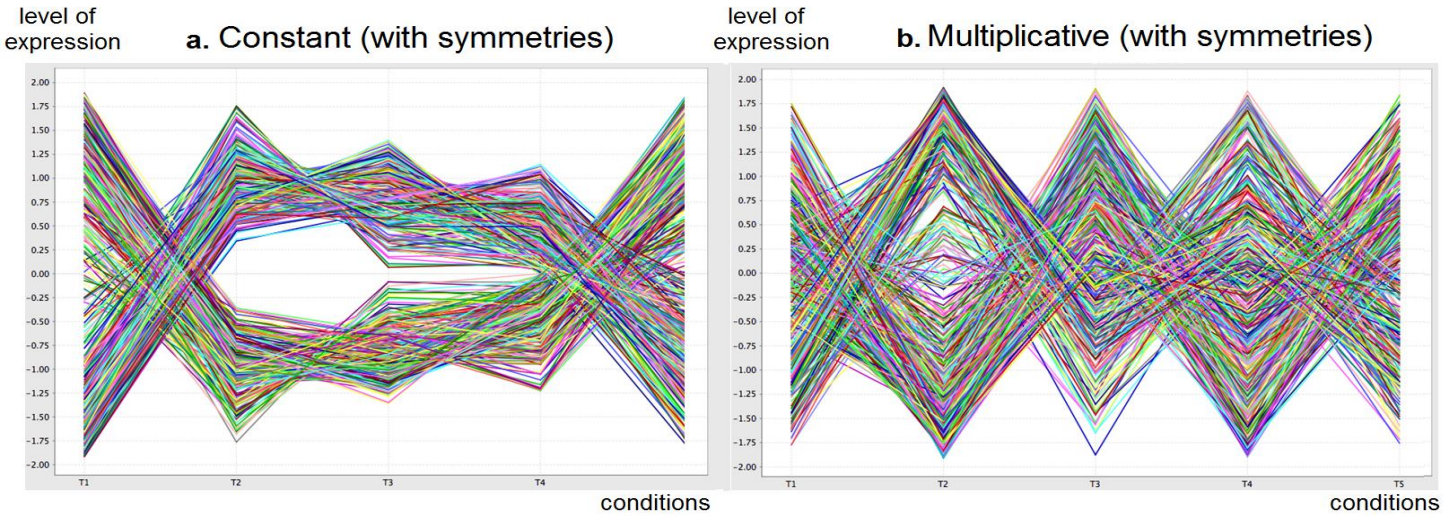
$$\boldsymbol{\pi} = \boldsymbol{y_1} \geq \boldsymbol{y_4} \geq \boldsymbol{y_2}$$
$$B = (\{x_1, x_2, x_3, x_4\}, \{y_1, y_2, y_4\})$$

Why?

- More flexible and noise robust
- Focus on orderings instead of absolute values (preferences, difficulties)

Symmetries

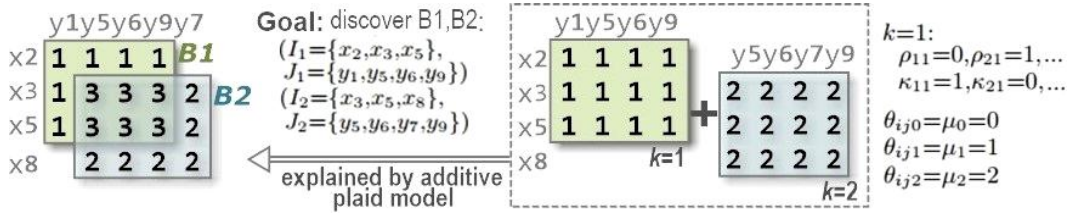
- Symmetries can be accommodated on observations: $a_{ij} \times k_i \; where \; k_i \in \{1, -1\}$
  - e.g. activation and repression regulatory patterns

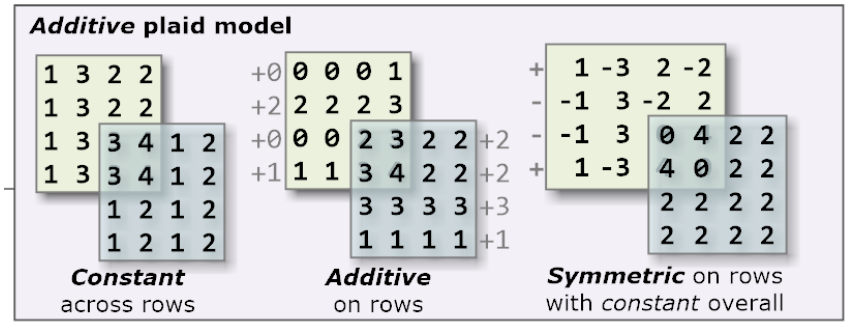


Plaid model

The **plaid assumption** considers the **cumulative effect** of the contributions from multiple biclusters on areas where their observations and variables **overlap**

Why? Synergistic behaviors (e.g. social networking, comorbidity effects, multi-purpose genes)





Merit functions

- *Variance* of values ⇒ constant overall
- *Mean square residue*, H ⇒ additive

$$r(a_{ij}) = a_{ij} - aiJ - a_{Ij} + aIJ \text{ (residue of an element in an additive subspace)}$$

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} r(a_{ij})^2 \text{ (measure of additive coherence)}$$

- *Pearson* correlation ⇒ additive/multiplicative
- *Cosine* measure ⇒ order-preserving

Many others…

Coherence strength

- Let $\bar{A}$ be the amplitude of the range of values in a matrix A
- Given A, the **coherence strength** is a range $\delta \in [0, \bar{A}]$, such that $a_{ij} = c_j + \gamma_i + \eta_{ij}$ (or other) and $\eta_{ij} \in [-\delta/2, \delta/2]$
  - e.g. $a_{13} = 1.4$, $a_{23} = 1.8$ and $a_{33} = 1.6$ can be seen as constant column if $\delta = 0.4$
- Increasing coherence strength
  - more tolerance to noise
  - larger biclusters
- Decreasing coherence strength?
- How to fix coherence strength?

Quality

The **quality** of a set of biclusters is defined by the **type** and **amount** of accommodated noise

- **amount** of noisy elements: e.g. 30% of noisy elements
- **type** of noise: e.g. consider only small deviations from $[\eta_{ij} - \delta/2, \eta_{ij} + \delta/2]$

2.	2.	2.	2.
2.	4.	2.	2.
2.	2.	2.	2.
2.	2.	0.	2.

12.5% of noisy elements

2.	2.	2.	2.
2.	1.8	2.	1.7
2.4	2.	2.	2.
2.	2.	2.3	2.

25% of slight deviations from $[\eta_{ij}$-$\delta/2$, η_{ij}+$\delta/2]$

Structure

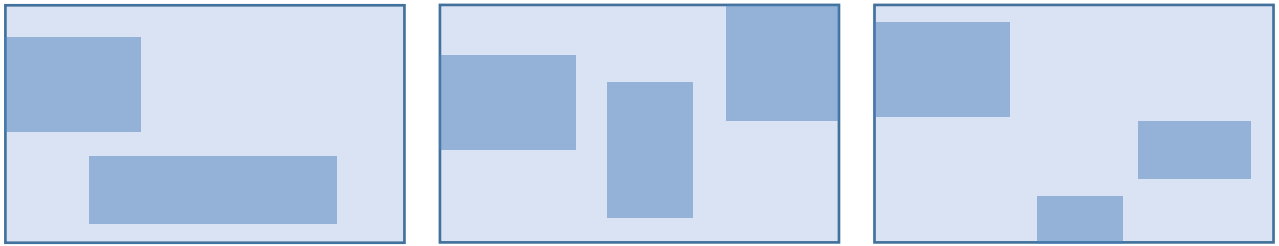
- **Number** of biclusters
- **Size** and **shape**
- **Positioning** constraints:
  - Exhaustive: rows and/or columns
  - Exclusive: rows and/or columns
  - Non-overlapping
  - Others: tree, hierarchical structures

Flexible structure:

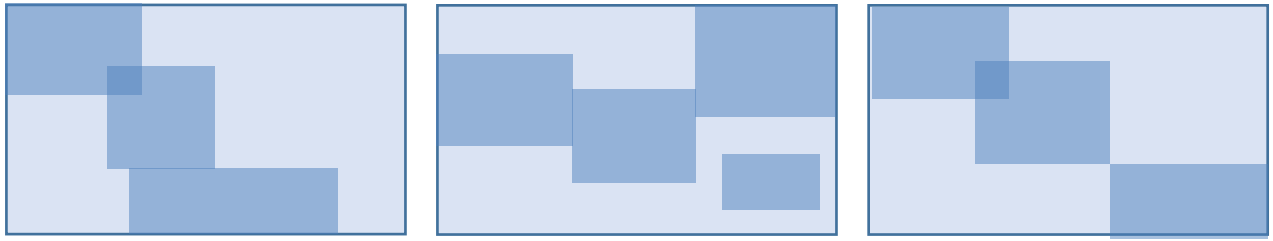
  - non-fixed number of biclusters
  - no constraints on size, shape and positioning

Structures

- **Exclusive** on rows (*left*), columns (*middle*), or both (*right*)

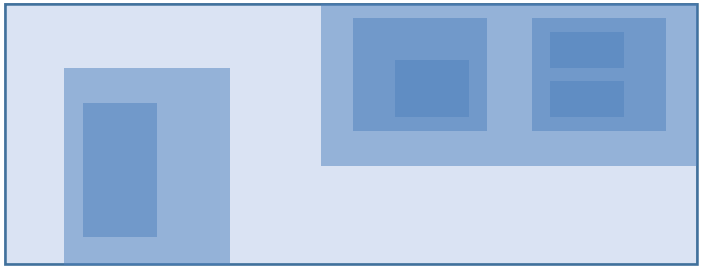


- **Exhaustive** on rows (*left*), columns (*middle*), or both (*right*)

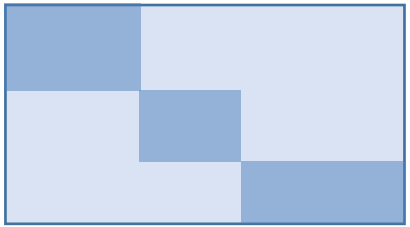


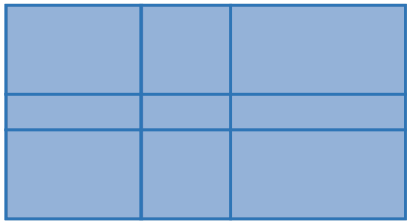
Structures

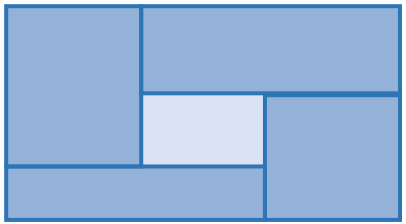
- **Hierarchical**



- **Others**
  - coclustering structure/checkboard
  - diagonal, L, square shape

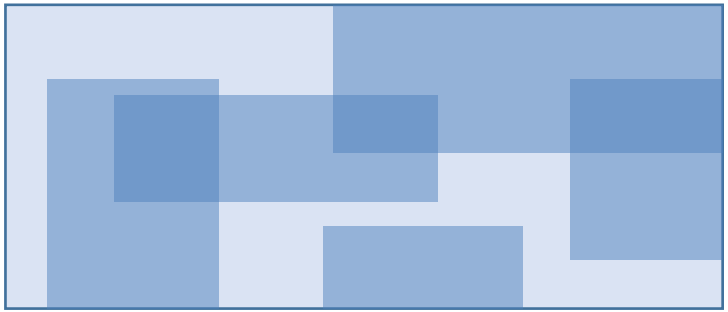






Flexible structures

- Arbitrarily positioned (possibly overlapping) biclusters



- Biclustering solutions with flexible positioning can be associated with redundant patterns
  - searches should further include *dissimilarity* criteria
    (e.g. merging biclusters with high overlapping at mining or postprocessing time)

Statistical significance

Is a bicluster **unexpected**?

- e.g. Is a bicluster defined by 2 observations and 2 variables in large dataset interesting? Or does it occur by chance?
- Solution: assess statistical significance
  - **randomize data** several times and compute the ratio of datasets where we are able to find a similar bicluster
  - **approximate data distributions** and perform a statistical test
    - the null hypothesis is that the bicluster does not occur by chance
    - we can statistically test its support using **binomial test** similarly as we did for itemsets
- Assessment can be considered during the biclustering search
- Refresh memory: *False positive* bicluster? *False negative* bicluster?

Evaluation metrics

Challenges

- no ground truth to evaluate biclusters observed in real data
- metrics only able to assess a single homogeneity criteria
- evaluation metrics used within the biclustering searches – problem?

Options

- **synthetic data** (knowledge of true/hidden biclusters)
  - **objective** metrics: accuracy, precision, completeness
- **real data** (no knowledge of true biclusters)
  - **subjective** metrics: domain-driven relevance scores

Evaluation metrics

Synthetic data (objective metrics)

H is the set of true biclusters and **B** is the set of found biclusters

- ***Clustering*** metrics on one dimension (X and Y separately)
  - silhouette, recall and precision – problems?
- ***Jaccard***-based match scores (MS) to assess the similarity of *B* and *H*
  - *MS*(*B*,*H*) extent to which found biclusters match hidden biclusters
  - *MS*(*H*,*B*) reflects how well hidden biclusters are recovered

$$MS(\mathcal{B},\mathcal{H}) = \frac{1}{|\mathcal{B}|}\Sigma_{(I_1,J_1)\in\mathcal{B}} max_{(I_2,J_2)\in\mathcal{H}} \frac{|I_1\cap I_2|}{|I_1\cup I_2|}$$

- ***Fabia consensus*** is sensitive to the number of biclusters in both sets

$$FC(\mathcal{B},\mathcal{H}) = \frac{1}{|\mathcal{S}_1|}\Sigma_{((I_1,J_1)\in\mathcal{S}_1,(I_2,J_2)\in\mathcal{S}_2)\in MP} \frac{|I_1\cap I_2|\times|J_1\cap J_2|}{|I_1|\times|J_1|+|I_2|\times|J_2|-|I_1\cap I_2|\times|J_1\cap J_2|}$$

Evaluation metrics

Real data (subjective metrics)

- *statistical significance*: unexpected occurrence probability
- *domain relevance*: unexpected probability of participating in studied process
- domain and statistical significance correlated but not always in agreement!

- HOW to assess domain relevance of bicluster (I, J)?
  - *source of annotations*: knowledge bases (e.g. GO) and literature data (e.g. PubMED)
  - statistically test I and J against well-established annotations
    - hypergeometric tests to compute **enrichment p-values** against an annotation database
    - *intuition*: most of entries in I (or J) sharing annotations in a database suggests relevance

Approaches

- **Number of biclusters at a time**
  - discover one bicluster at a time and then mask it
  - discover a set of biclusters at a time

- **Optimality**
  - **exhaustive** and (quasi-)exhaustive searches
    - e.g. find heavy subgraphs in bipartite graphs mapped from data
    - *heavy!* Place restrictions on structure, coherence and quality
  - **approximate** searches (next slide)

Approaches

- **Greedy iterative searches**
  - iteratively add/remove rows and columns to maximize a merit function
    - e.g. CC minimize mean square residue (MSR) until reaching MSR < δ
- **Row and column clustering combination** (e.g. CTWC, ITWC)
  - apply clustering on rows and columns separately
  - use an iterative procedure to combine the two clustering results
- **Divide and conquer searches**
  - break the matrix into submatrices (e.g. find best row or column split)
  - continue the biclustering process on the new submatrices
- **Distribution parameter identification** (e.g. plaid)
  - use mixture to describe biclustering solution and learn its parameters (maximize likelihood)

Outline

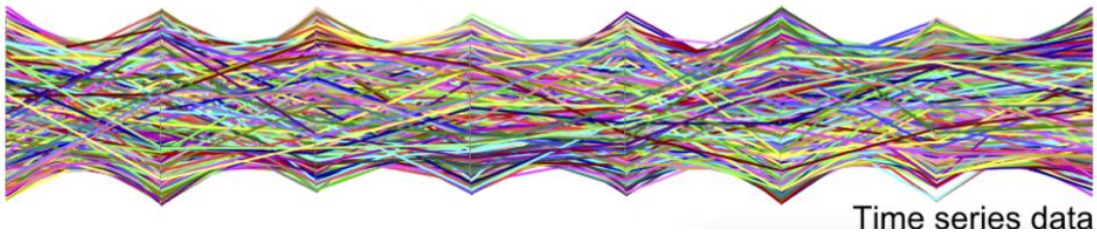
- Subspace clustering
- Biclustering
  - coherence
  - quality
  - structure
  - evaluation
  - searches
- **Triclustering**
- Deep learning
- Appendix

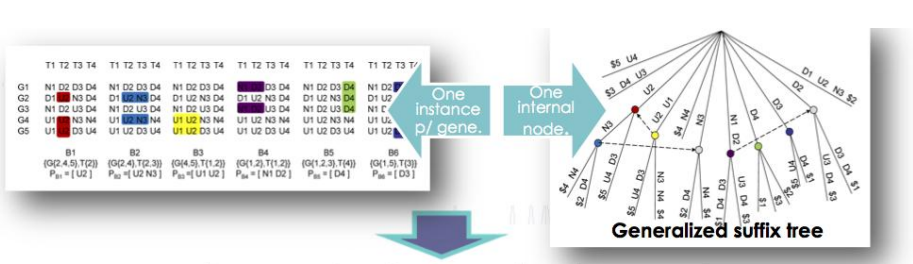
Time series biclustering

- Until here: biclustering as relevant pattern discovery task for multivariate data, yet…
  - applicability to (univariate) time series data is equally pervasive
    - a *bicluster* is defined by a subset of **observations** and **time points**
    - **contiguity** is generally assumed across time points (convex temporal pattern)
    - **temporal misalignments** between observations can be further accommodated
      - e.g. individuals at different stages of a disease

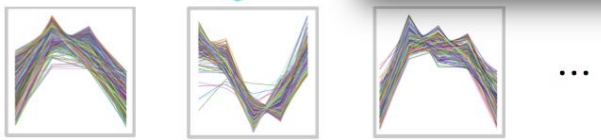
- Illustrative method: CCC identifies patterns in linear complexity time using suffix trees
  - eCCC extension can further allow temporal misalignments and noise

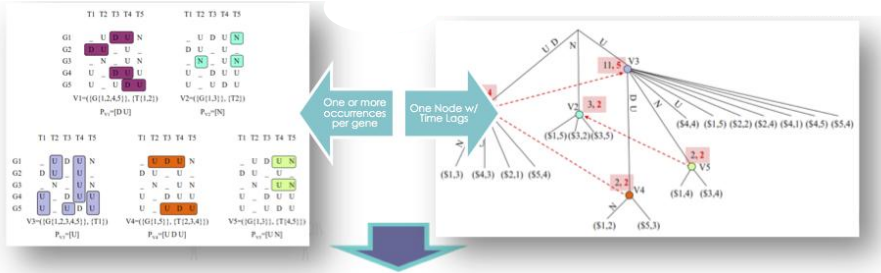
Time series biclustering: CCC algorithm



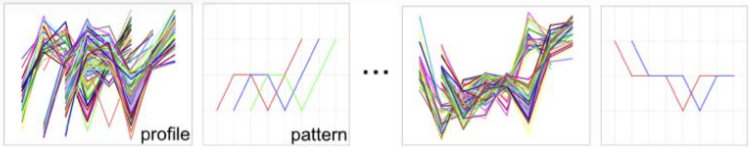


CCC
Biclusters
same pattern



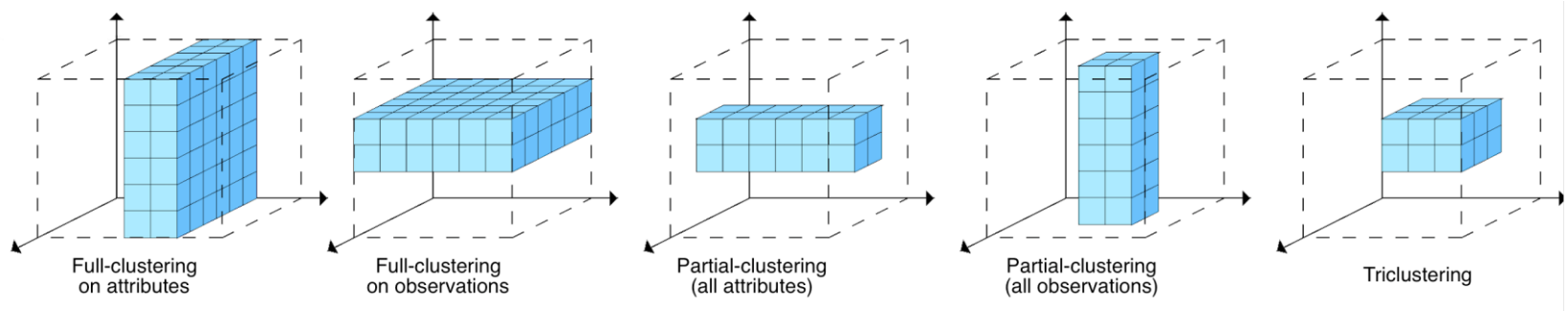


"Late"
Biclusters
same pattern
potential delay

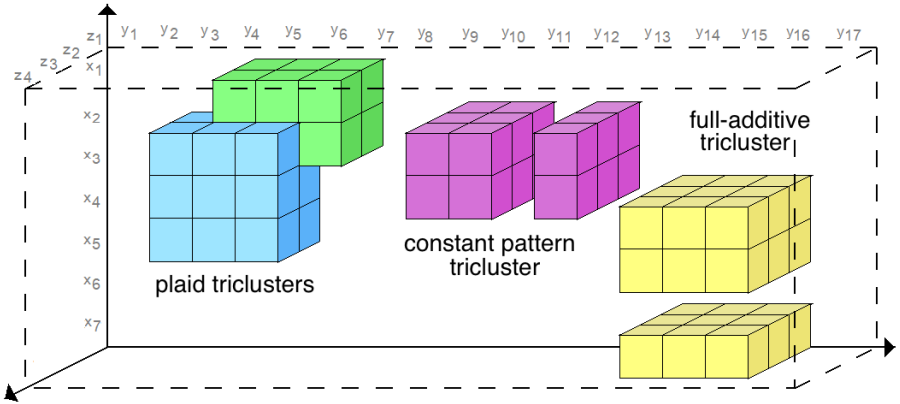


Triclustering

- How to move from univariate to multivariate time series (MTS) data?
  - multivariate data is defined by a set of observations, variables and time points
- Solution: **triclustering**
  - a **tricluster** is a subset of *observations*, *variables* and *time points* with good:
    - homogeneity, e.g. well established temporal pattern on a subset of variables
    - statistical significance, e.g. unexpected high #observations supporting the pattern



Triclustering



	y3	y4	y5
z3 x2	1	3	3
z3 x3	1	1	1
z3 x4	1	1	1
z4 x2	1	3	3
z4 x3	1	1	1
z4 x4	1	1	1

	y4	y5	y6
z2 x1	2	2	2
z2 x2	2	2	2
z3 x1	2	2	2
z3 x2	3	3	2

Triclusters with low variance and cumulative effects on overlapping regions

	y9	y10	y12
z1 x2	2	3	2
z1 x3	2	3	2
z2 x2	2	3	2
z2 x3	2	3	2
z3 x2	2	3	2
z3 x3	2	3	2

Tricluster with constant pattern on observations and contexts

	y14	y15	y16
z1 x4	0	1	1
z1 x5	1	2	2
z1 x7	0	1	1
z2 x4	1	2	2
z2 x5	2	3	3
z2 x7	1	2	2
z3 x4	0	1	1
z3 x5	1	2	2
z3 x7	0	1	1

Tricluster with low mean square residue

Triclustering

- Applications
  - computational **biology**: regulation (genes × conditions × time), protein–protein interaction (proteins × interaction × conditions), drug response profiling (cell lines × drugs × dosages)
  - **recommender systems**: dynamic preference modeling (users × items × time/location/device)
  - **NLP**: topic discovery (documents × terms × time/language), sentiment analysis with context
  - **social networks**: community detection with context, role discovery in dynamic networks
  - **time series**: segmentation (sensors × signals × time windows), EEG analysis, climate monitoring
  - **vision**: video analysis (objects × features × frames), activity recognition, multimodal systems
  - **business**: sales patterns (products × stores × time), fraud detection, basket analysis with season
  - **healthcare**: patient stratification (patients × variables × time), treatment outcome, epidemiology
- *Challenge*: extend previous concepts on biclustering after reading the survey by Henriques et al. (2017)

Outline

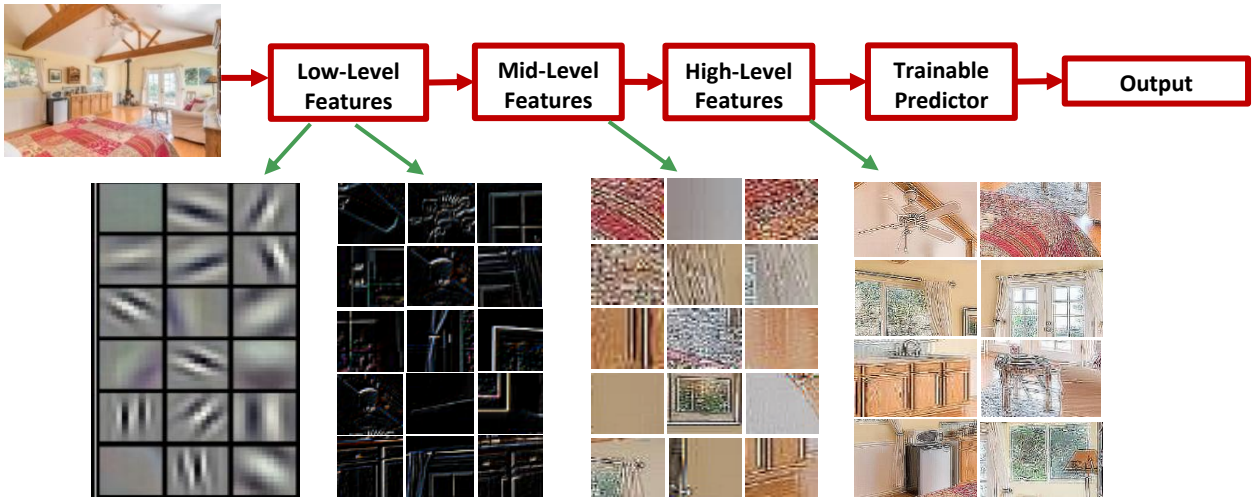
- Subspace clustering
- Biclustering
  - coherence
  - quality
  - structure
  - evaluation
  - searches
- Triclustering
- **Deep learning**
- Appendix

Patterns and deep learning?

Until now: focus placed on *explicit* pattern discovery (classic and subspace clustering)

What about deep learning? Multi-layer learning process to extract rich *implicit patterns*

- **image**: pixels → edges → textures → motifs → parts → objects
- **text**: character → word → word group → clause → sentence → story

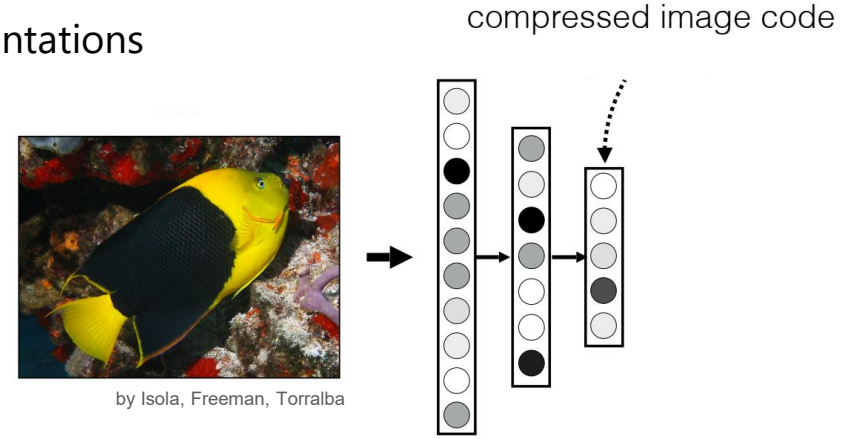


by Param Vir Singh

Patterns and deep learning?

- **Hidden unit ≈ latent pattern**
  - high-weight connections ≈ frequent features
  - activations per observation ≈ pattern support

- Recall the following properties of a good representations
  - *disentangled explanatory factors*
    - each unit should capture a separate, meaningful aspect of the data
  - *loose factor dependencies*
  - *sparsity:* for any observation $\mathbf{x}$, only some factors are relevant



Patterns and deep learning?

- Handling complex neural networks?
  - multimodal architectures
    - extension to shared spaces: latent patterns embody relationships across data modalities
  - transformer-based architectures
    - attention weights ≈ pattern importance
    - attention heads ≈ distinct patterns

- ***Core problem***: neural networks only able to generate *implicit patterns*
  - limited pattern deidentification
  - limited post-hoc explainability on the deidentified patterns

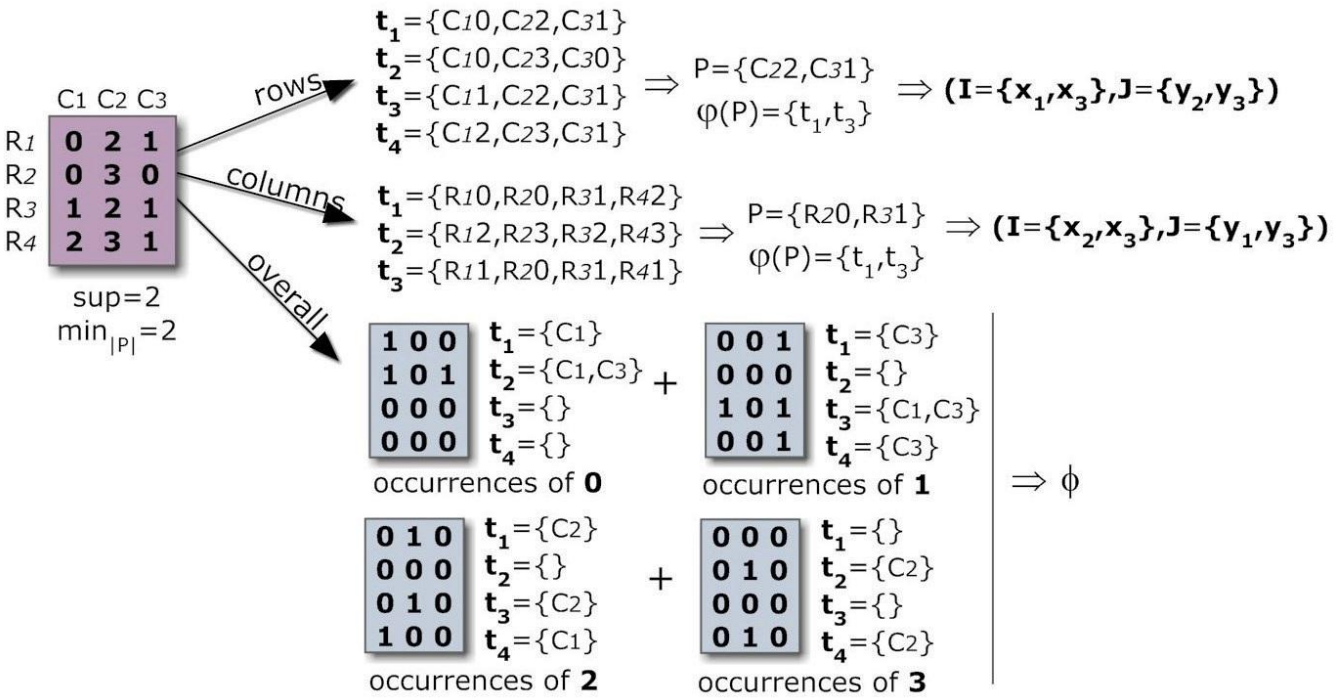
Thank you!

Rui Henriques

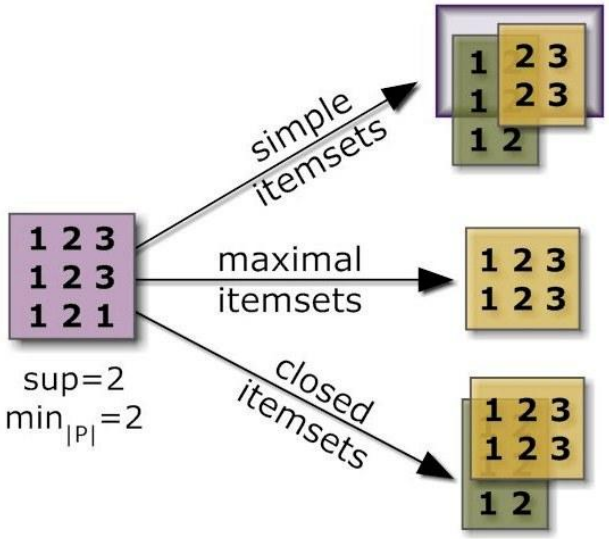
rmch@tecnico.ulisboa.pt

Appendix

BicPAMS: constant biclusters
(using frequent itemset mining)



Maximal biclusters
(≡ closed frequent itemsets)

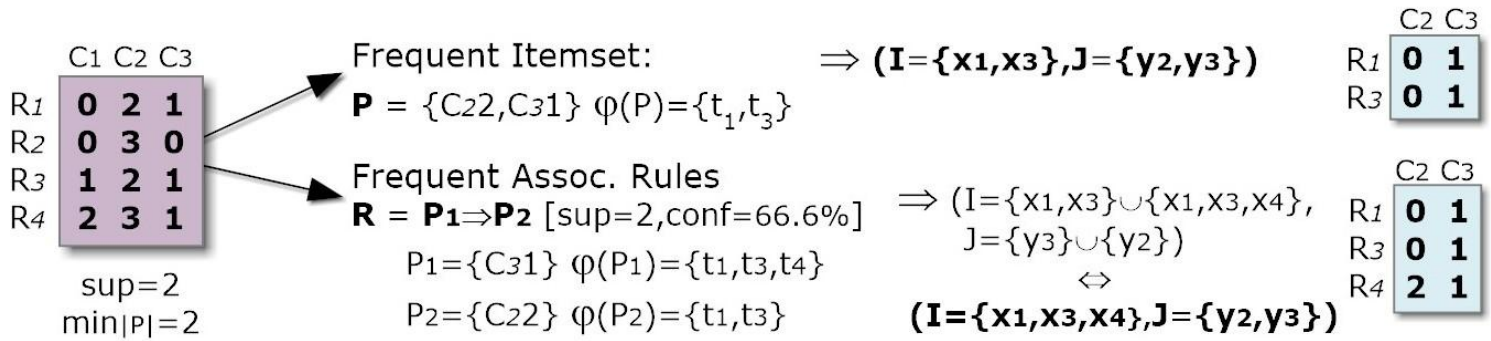


P_1={B2,C3}, $\varphi(P_1)$={t_1,t_2}
P_1={A1,C2}, $\varphi(P_1)$={t_1,t_2,t_3}
P_1={A1,B2,C3}, $\varphi(P_1)$={t_1,t_2}
⇒ include all maximal biclusters

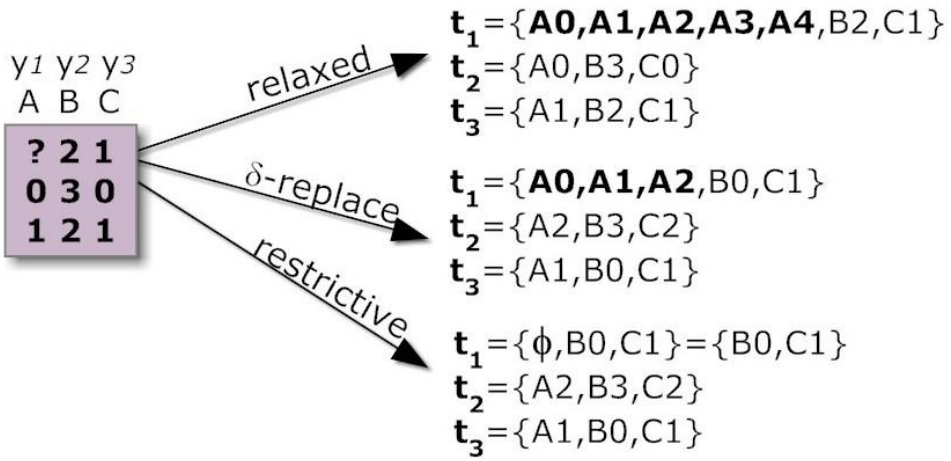
P_1={A1,B2,C3}, $\varphi(P_1)$={t_1,t_2}
⇒ subset of maximal biclusters

P_1={A1,C2}, $\varphi(P_1)$={t_1,t_2,t_3}
P_1={A1,B2,C3}, $\varphi(P_1)$={t_1,t_2}
⇒ exact maximal biclusters

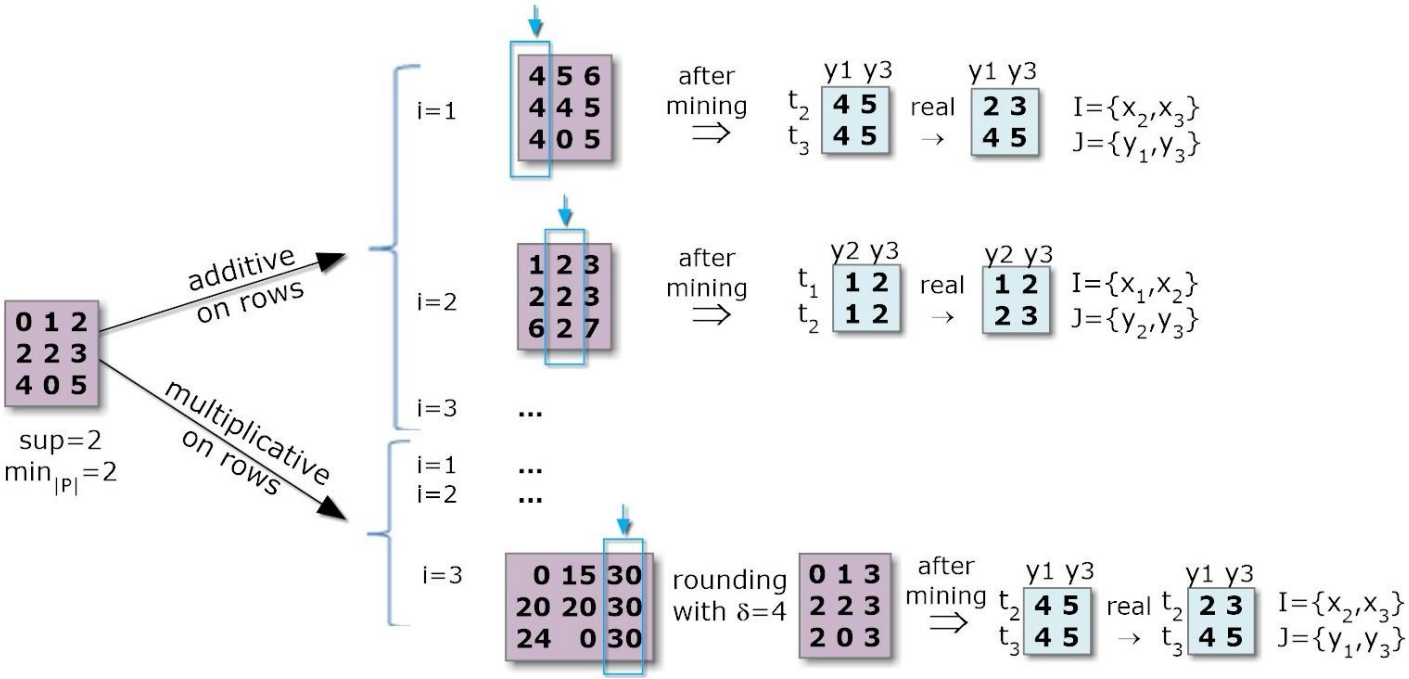
BicPAMS: noise-tolerant biclusters
(using association rules)



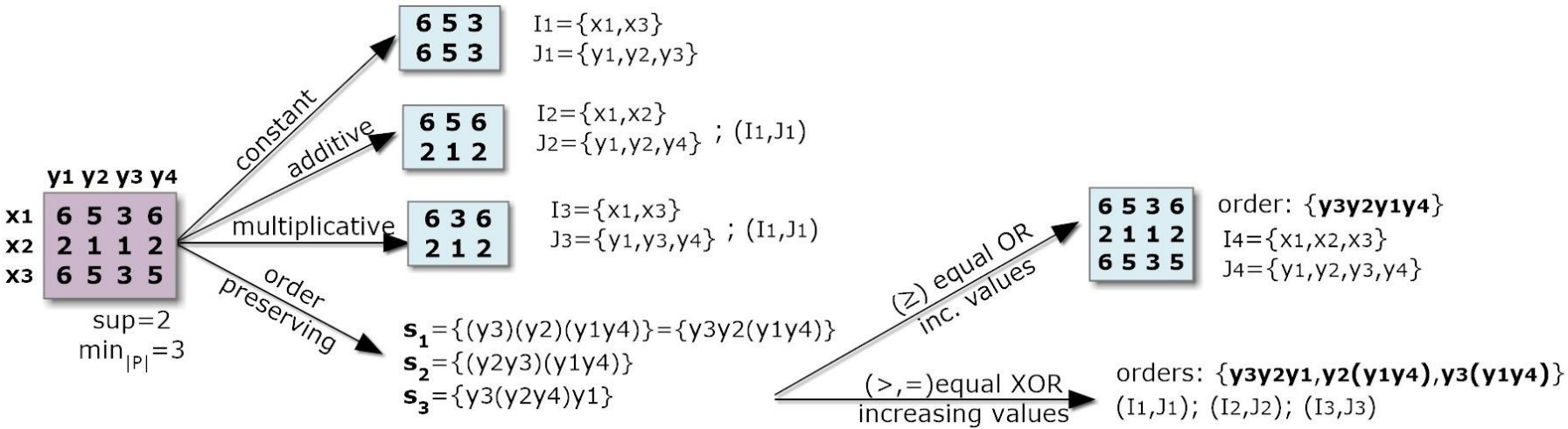
Handling missings



BicPAMS: additive models
(based on shifting factors)



BicPAMS: order-preserving models
(with sequential pattern mining)



Summary

- Subspace clustering is key to find **non-trivial patterns** on **real-valued data**
  - unsupervised exploration of **high-dimensional data** (focus on subspaces)
- Subspace clusters satisfy specific criteria of homogeneity and statistical significance
- **Merit functions** determine the homogeneity (patterns of interest)
  - condition the structure, coherence and quality (noise tolerance) of patterns
- **Statistically significant** patterns are unexpected (low *p*-value)
- **Coherence** defines the underlying assumption (constant, additive, order-preserving…) and strength
- The **structure** is determined by the number, shape and positioning of patterns
- Subspace clustering **searches** can be classified in exhaustive, greedy, and parametric