Clustering : une vision unifiée pour une utilisation éclairée - Partie 4 : Traitement de la grande dimension & co-clustering

C Biernacki

Clustering : une vision unifiée pour une utilisation éclairée
–
Partie 4
Traitement de la grande dimension & co-clustering
(leçon plus avancée)

C. Biernacki

Ateliers statistiques de la SFdS, 15 et 16 juin 2023, IHP, Paris

Outline

Motivation

High dimensional (HD) data sets are now frequent:

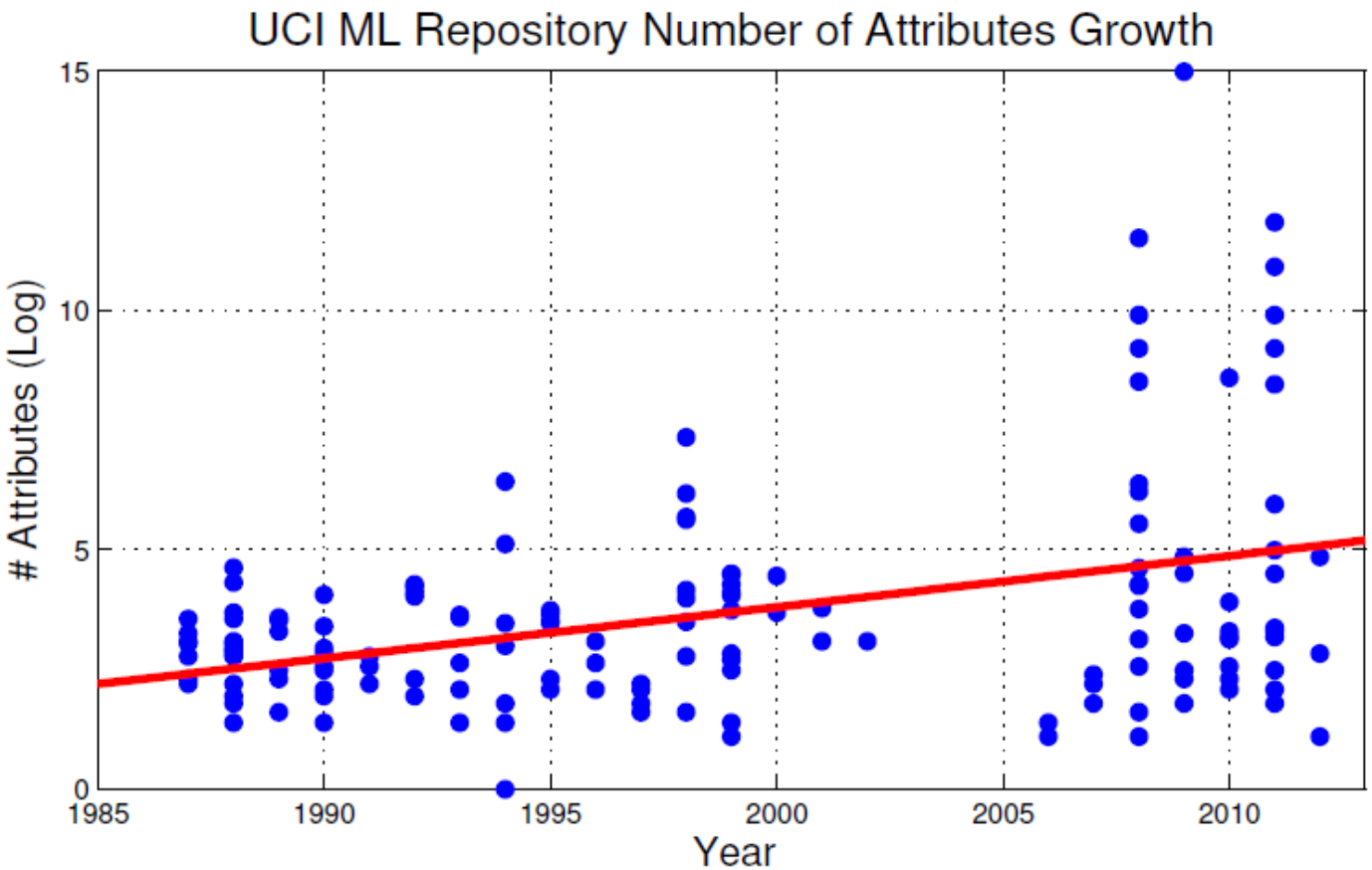
- Marketing: $d \sim 10^2$
- microarray gene expression: $d \sim 10^2$–10^4
- SNP data: $d \sim 10^6$
- Curves: depends on discretization but can be very high
- Text mining
- ...

Clustering has to be applied for HD datasets for the same reasons as the lower dimensional datasets:

- Data summary
- Data exploratory
- Preprocessing for more flexibility of a forthcoming prediction step

But clustering is even more important since visualization in the HD setting can be hazardous...

Today': exponential growing of dimension[1]



[1] S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, **29**

HD data: definition (1/2)

An attempt in the non-parametric case

Dataset $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_j$ described by d variables, where $n = o\left(e^d\right)$

Justifications:

- To approximate within error ϵ a (Lipschitz) function of d variables, about $(1/\epsilon)^d$ evaluations on a grid are required [Bellman, 61]
- Approximate a Gaussian distribution with fixed Gaussian kernels and with approximate error of about 10% [Silverman, 86]

$$\log_{10} n(d) \approx 0.6(d - 0.25)$$

For instance, $n(10) \approx 7.10^5$

HD data: definition (2/2)

An attempt in the parametric case

Dataset $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_j$ described by d variables and a model $\mathbf{m}$ with ν parameters, where $n = o(g(\nu))$, with g a given function

Justification:

- We consider the heteroscedastic Gaussian mixture with of true parameter $\boldsymbol{\theta}^*$ with K^* components. We note $\hat{\boldsymbol{\theta}}$ the Gaussian MLE with K^* components. We have g linear from the following result [Michel, 08]: it exists constants κ, A and C such that

$$\mathrm{E}_{\mathbf{x}}[\mathrm{Hellinger}^2(\mathrm{p}_{\boldsymbol{\theta}^*}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\hat{K}}})] \leq C \left[\kappa \frac{\nu}{n} \left\{ 2A \ln d + 1 - \ln \left(1 \wedge \left[\frac{\nu}{n} A \ln d \right] \right) \right\} + \frac{1}{n} \right].$$

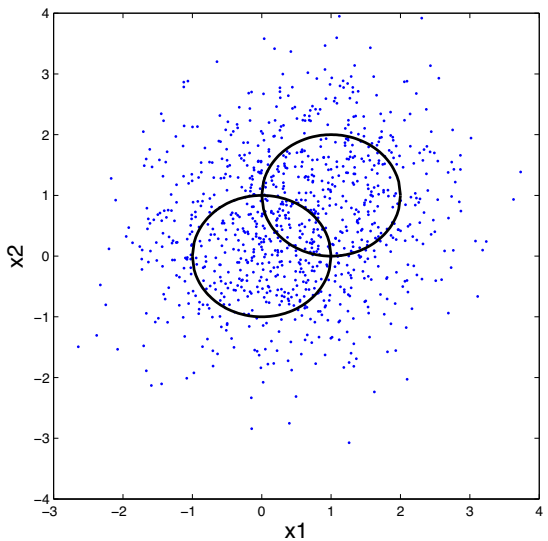
But ν can be high since $\nu \sim d^2/2$, combined with potentially large constants.

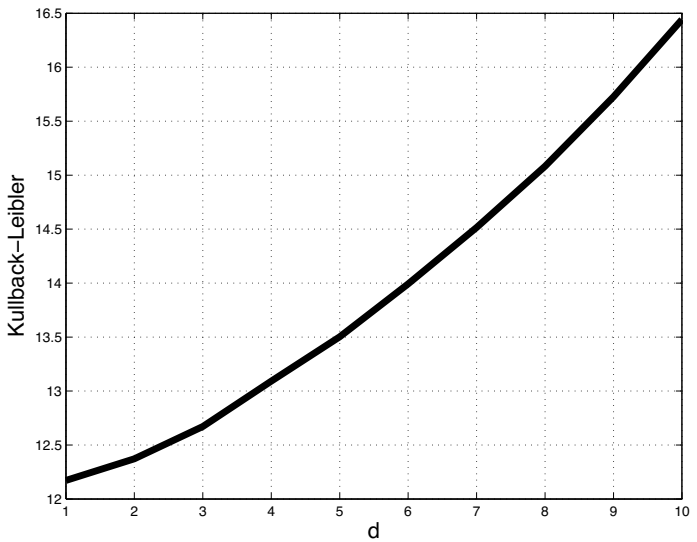
HD density estimation: curse

A two-component d-variate Gaussian mixture:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1 | z_{11} = 1 \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}), \quad \mathbf{X}_1 | z_{12} = 1 \sim \mathrm{N}_d(\mathbf{1}, \mathbf{I})$$

Components are more and more separated when d grows: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{I}} = \sqrt{d}\ldots$





. . . but density estimation quality decreases with d

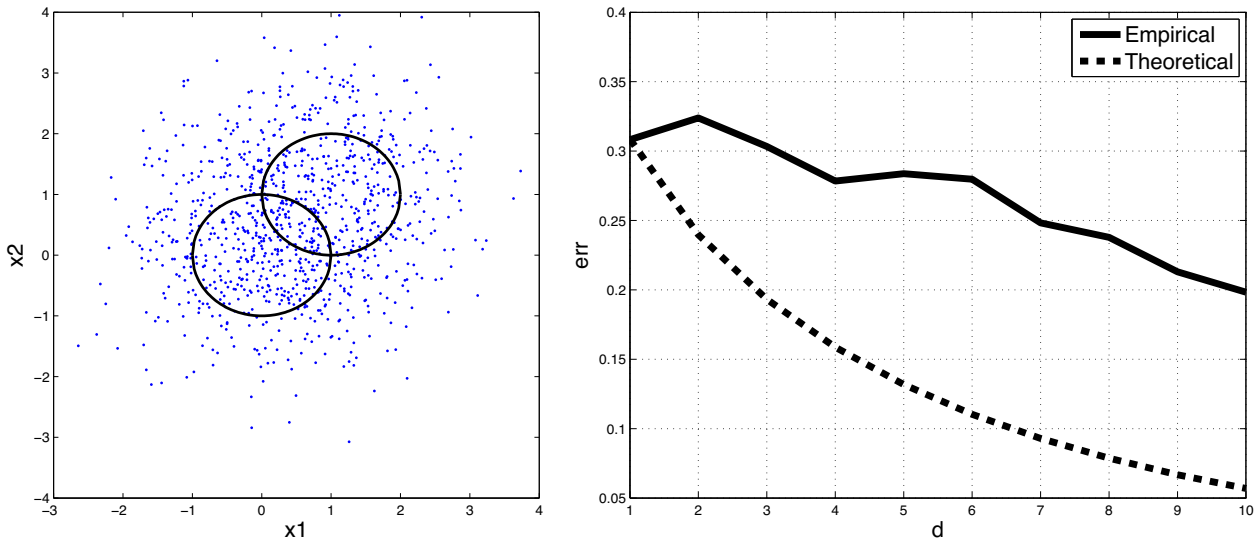
HD clustering: blessing (1/2)

A two-component d-variate Gaussian mixture:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1 | z_{11} = 1 \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}), \quad \mathbf{X}_1 | z_{12} = 1 \sim \mathrm{N}_d(\mathbf{1}, \mathbf{I})$$

Each variable provides equal and own separation information

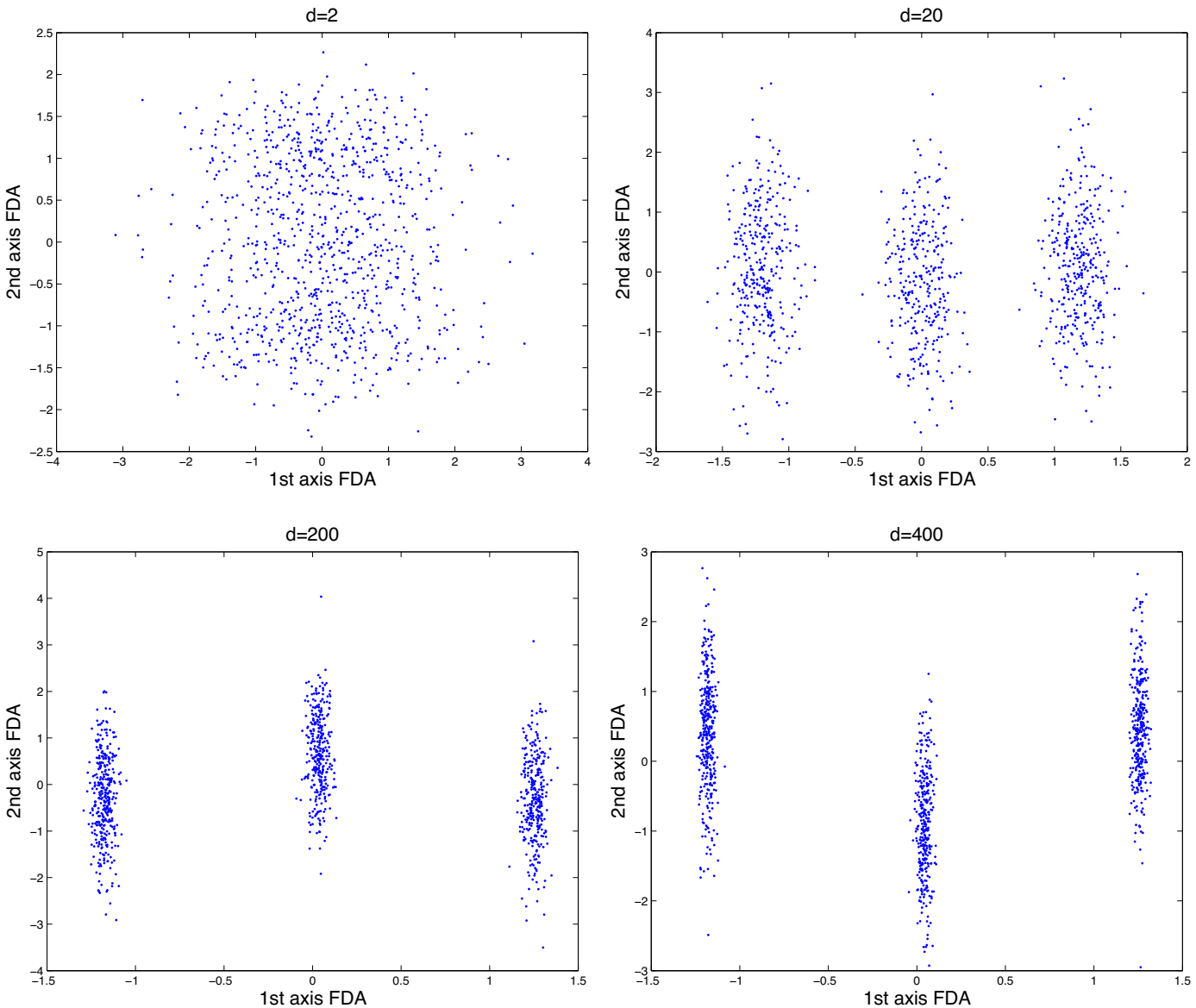
Theoretical error decreases when d grows: $\mathrm{err}_{theo} = \Phi(-\sqrt{d}/2)\ldots$



. . . and empirical error rate decreases also with d!

HD clustering: blessing (2/2)

FDA



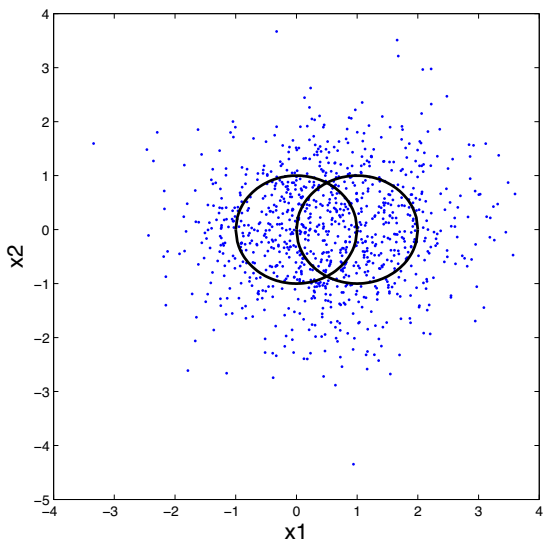
HD clustering: curse (1/2)

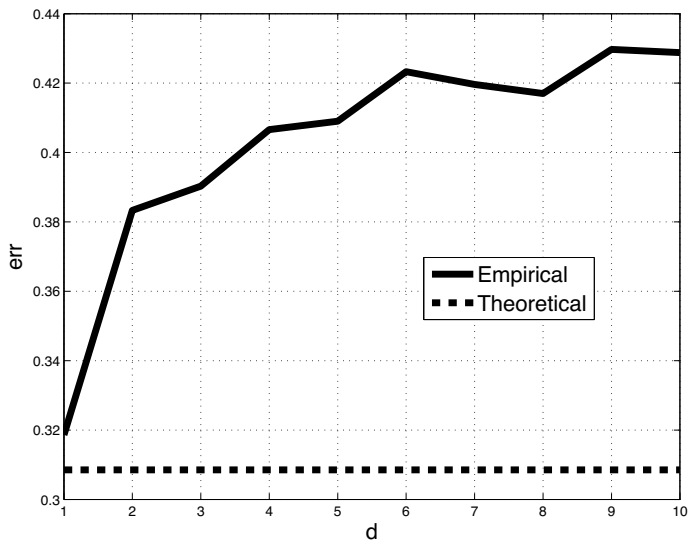
Many variables provide no separation information

Same parameter setting except:

$$\mathbf{X}_1 | z_{12} = 1 \sim \mathrm{N}_d((1\ 0\ \ldots\ 0)', \mathbf{I})$$

Groups are not separated more when d grows: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{I}} = 1 \ldots$





. . . thus theoretical error is constant $(= \Phi(-\frac{1}{2}))$ and empirical error increases with d

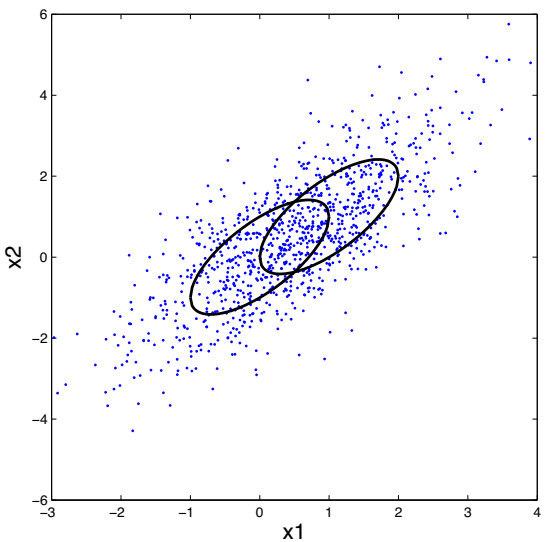
HD clustering: curse (2/2)

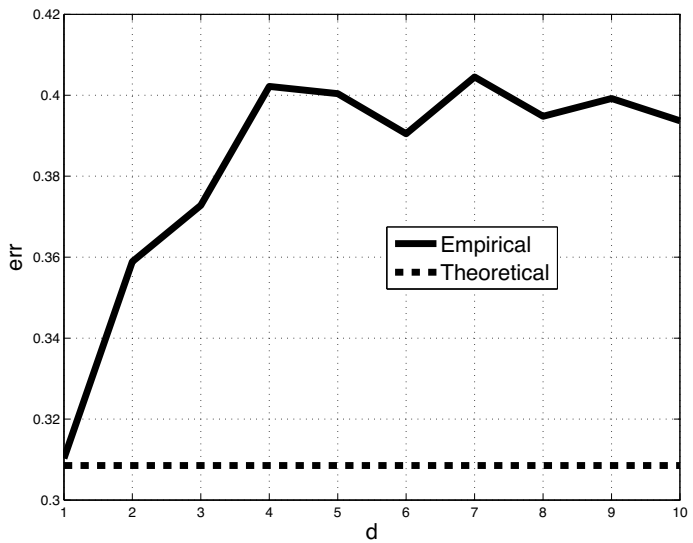
Many variables provide redundant separation information

Same parameter setting except:

$$\mathbf{X}_1^j = \mathbf{X}_1^1 + \mathrm{N}_1(0,1) \quad (j = 2, \ldots, d)$$

Groups are not separated more when d grows: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}} = 1 \ldots$





. . . thus err_{theo} is constant $(= \Phi(-\frac{1}{2}))$ and empirical error increases (less) with d

The trade-off bias/variance

The fundamental statistical principle

Always minimize an error err between truth ($\mathbf{z}$) and estimate ($\hat{\mathbf{z}}$)

- Gap between true ($\mathbf{z}$) and model-based ($\mathcal{Z}_\mathbf{p}$) partitions: $\mathbf{z}^* = \arg\min_{\tilde{\mathbf{z}} \in \mathcal{Z}_\mathbf{p}} \Delta(\mathbf{z}, \tilde{\mathbf{z}})$
- Estimation $\hat{\mathbf{z}}$ of $\mathbf{z}^*$ in $\mathcal{Z}_\mathbf{p}$: any relevant method (bias, consistency, efficiency...)
- Fundamental decomposition of the observed error $\text{err}(\mathbf{z}, \hat{\mathbf{z}})$:

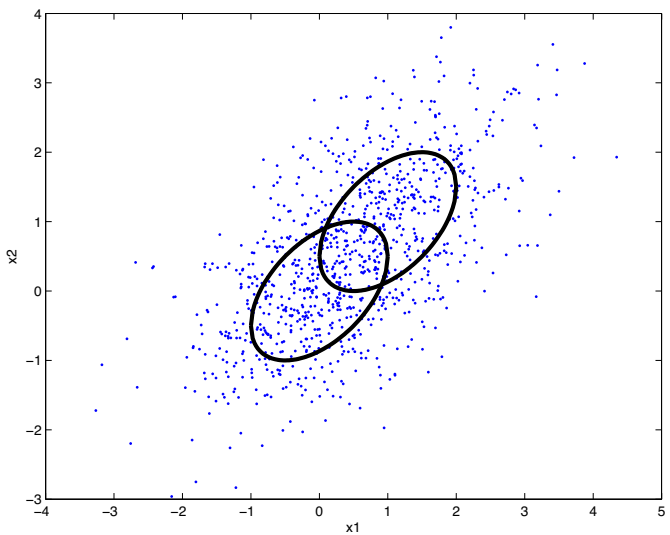
$$\begin{aligned} \text{err}(\mathbf{z}, \hat{\mathbf{z}}) &= \Big\{\text{err}(\mathbf{z}, \mathbf{z}^*) - \text{err}(\mathbf{z}, \mathbf{z})\Big\} + \Big\{\text{err}(\mathbf{z}, \hat{\mathbf{z}}) - \text{err}(\mathbf{z}, \mathbf{z}^*)\Big\} \\ &= \Big\{\text{bias}\Big\} + \Big\{\text{variance}\Big\} \\ &= \Big\{\text{error of approximation}\Big\} + \Big\{\text{error of estimation}\Big\} \end{aligned}$$

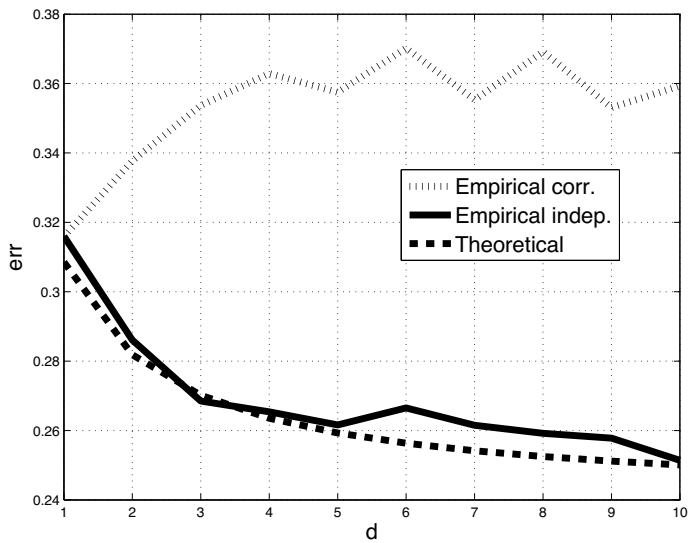
Bias/variance in HD: reduce variance, accept bias

A two-component d-variate Gaussian mixture with intra-dependency:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1 | z_{11} = 1 \sim \mathrm{N}_d(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{X}_1 | z_{12} = 1 \sim \mathrm{N}_d(\mathbf{1}, \mathbf{\Sigma})$$

- Each variable provides equal and own separation information
- Theoretical error decreases when d grows: $\mathrm{err}_{theo} = \Phi(-\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{\Sigma}^{-1}}/2)$
- Empirical error rate with the (true) intra-correlated model worse with d
- Empirical error rate with the (false) intra-independent model better with d!





Some alternatives for reducing variance

- Dimension reduction in non-canonical space (PCA-like typically)
- Dimension reduction in the canonical space (variable selection)
- Model parsimony in the initial HD space (constraints on model parameters)

But which kind of parsimony?

- Remember that clustering is a way for dealing with large n
- Why not reusing this idea for large d?

Co-clustering

It performs parsimony of row clustering through variable clustering

Gaussian mixture of factor analysers (1/2)

Definition

[Ghahramani and Hinton, 97], [McLachlan *et al.*, 03]

$$\boldsymbol{\Sigma}_k = \mathbf{B}_k \mathbf{B}_k' + \omega_k \boldsymbol{\Lambda}_k$$

where

- $\mathbf{B}_k$ is a *loadings* $d \times q$ non-square real matrix ($1 \leq q \leq q_{\max}$, $q_{\max} < d$)
- ω_k is a positive real number
- $\boldsymbol{\Lambda}_k$ is a $d \times d$ diagonal positive definite matrix such that $|\boldsymbol{\Lambda}_k| = 1$

Interpretation $\mathbf{X}_1 \in \mathbb{R}^d$ is generated by a latent variable $\mathbf{Y}_1 \in \mathbb{R}^q$

$$\mathbf{X}_{1|\mathbf{Y}_1, Z_{1k}=1} = \mathbf{B}_k \underbrace{\mathbf{Y}_1}_{\text{factor}} + \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_k$$

where $\mathbf{Y}_1 \perp \boldsymbol{\varepsilon}_k$ and

$$\mathbf{Y}_1 \sim \mathrm{N}_q(\mathbf{0}, \mathbf{I}) \quad \text{and} \quad \boldsymbol{\varepsilon}_k \sim \mathrm{N}_d(\mathbf{0}, \omega_k \boldsymbol{\Lambda}_k)$$

Gaussian mixture of factor analysers (2/2)

Complexity

$$\nu = (K-1) + Kd + Kq(d-(q-1)/2) + Kd$$

Model selection

- Models in competition: $\mathbf{m} = (q, K)$ + 12 parsimonious versions [McNicholas and Murphy, 10]
- Classical criteria can be used

Package

PGMM: `http://cran.r-project.org/web/packages/pgmm/index.html`

HD Gaussian models (1/2)

Definition

[Bouveyron *et al.*, 07]

$$\mathbf{\Sigma}_k = \mathbf{D}_k \mathbf{\Delta}_k \mathbf{D}_k'$$

where

- $\mathbf{D}_k$ is the orthogonal matrix of the eigenvectors of $\mathbf{\Sigma}_k$
- $\mathbf{\Delta}_k$ is a diagonal matrix containing the eigenvalues of $\mathbf{\Sigma}_k$

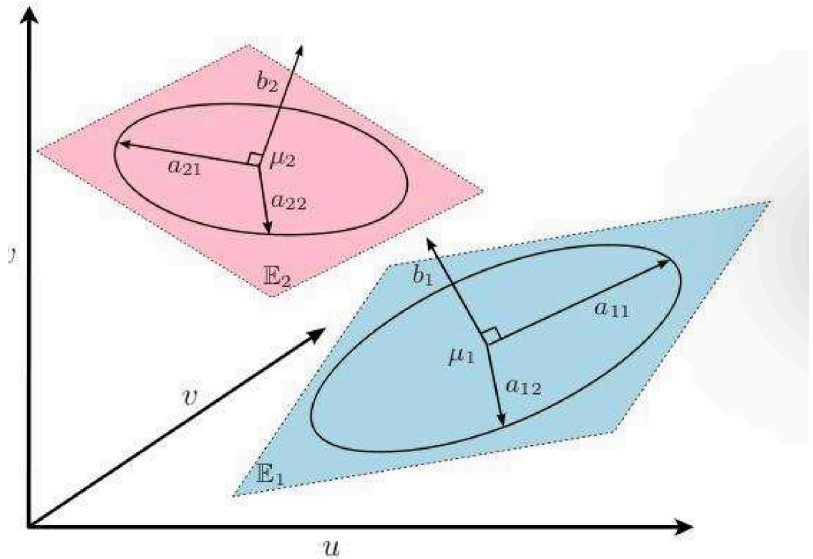
$$\mathbf{\Delta}_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{k\delta_k} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} \delta_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (d - \delta_k) \end{matrix}$$

with $a_{kj} \geq b_k$, for $j = 1, ..., \delta_k$ and $\delta_k < d$

Complexity

$$\nu = (K-1) + Kd + \sum_{k=1}^{K} \delta_k [d - (\delta_k + 1)/2] + \sum_{k=1}^{K} \delta_k + 2K$$

HD Gaussian models (2/2)



Model selection

- Models in competition: $\mathbf{m} = ((\delta_1, \ldots, \delta_K), K) + 8$ parsimonious versions [Bouveyron *et al.*, 07]
- Since it could be combinatorics, detect empirically a break in the eigenvalue scree of the empirical covariance matrix of each group
- Classical criteria can be used for the remaining models

Package

RMIXMOD: `http://cran.r-project.org/web/packages/Rmixmod/index.html`

Gaussian "variable selection" (1/2)

Definition

[Raftery and Dean, 06], [Maugis *et al.*, 09a], [Maugis *et al.*, 09b]

$$p(\mathbf{x}_1;\boldsymbol{\theta}) = \underbrace{\left\{\sum_{k=1}^{K}\pi_k p(\mathbf{x}_1^S;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}}_{\text{clustering variables}} \times \underbrace{\left\{p(\mathbf{x}_1^U;\mathbf{a}+\mathbf{x}_1^R\mathbf{b},\mathbf{C})\right\}}_{\text{redundant variables}} \times \underbrace{\left\{p(\mathbf{x}_1^W;\mathbf{u},\mathbf{V})\right\}}_{\text{independent variables}}$$

where

- all parts are Gaussians
- S: set of variables useful for clustering
- U: set of redondant clustering variables, expressed with $R \subseteq S$
- W: set of variables independent of clustering

Trick

Variable selection is recasted as a particular variable role

Gaussian "variable selection" (2/2)

Model selection

- Models in competition: $\mathbf{m} = (S, R, U, W, K) \rightarrow$ combinatorics
- Use a backward stepwise algorithm guided by a model selection criterion: $d \approx 30$
- Use alternatively a lasso-like procedure for ranking quickly different sets of clustering related and clustering independent variables [Sedki *et al.*, 14]

$$\text{crit}_{\lambda,\rho} = \ell(\boldsymbol{\theta}; \bar{\mathbf{x}}) - \lambda \sum_{k=1}^{K} \sum_{j=1}^{d} |\mu_{kj}| - \rho \sum_{k=1}^{K} \sum_{(j,j'),j \neq j'} |(\boldsymbol{\Sigma}_k^{-1})_{jj'}|$$

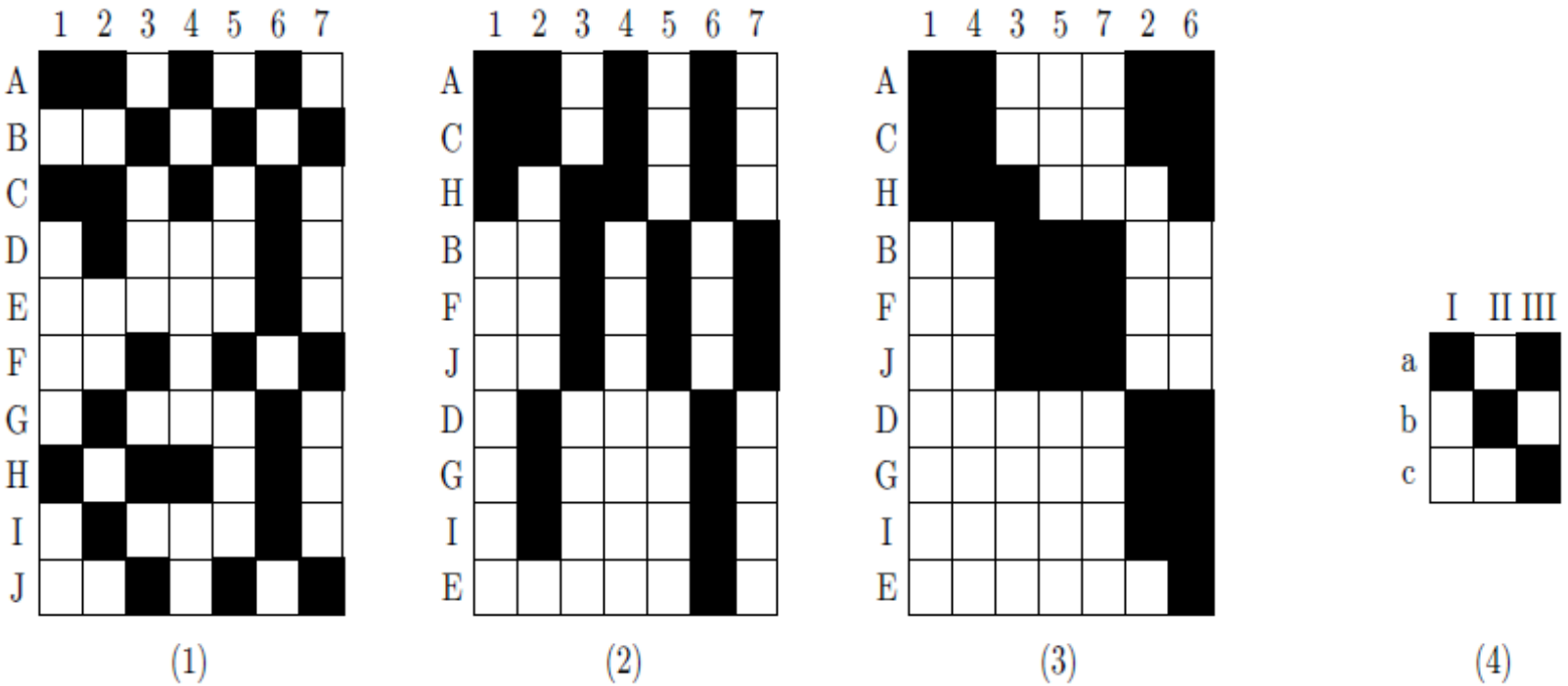
where $\boldsymbol{\theta}$ full Gaussian parameters, $\bar{\mathbf{x}}$ is $\mathbf{x}$ centered and (λ, ρ) are on a grid
A variable j is considered independent of clustering if $\hat{\mu}_{kj}(\lambda, \rho) = 0$ for all k

- Classical criteria are available

Package

- SELVARCLUST:
  `http://www.math.univ-toulouse.fr/~maugis/SelvarClustHomepage.html`
- SELVARMIX:
  `http://cran.r-project.org/web/packages/SelvarMix/index.html`

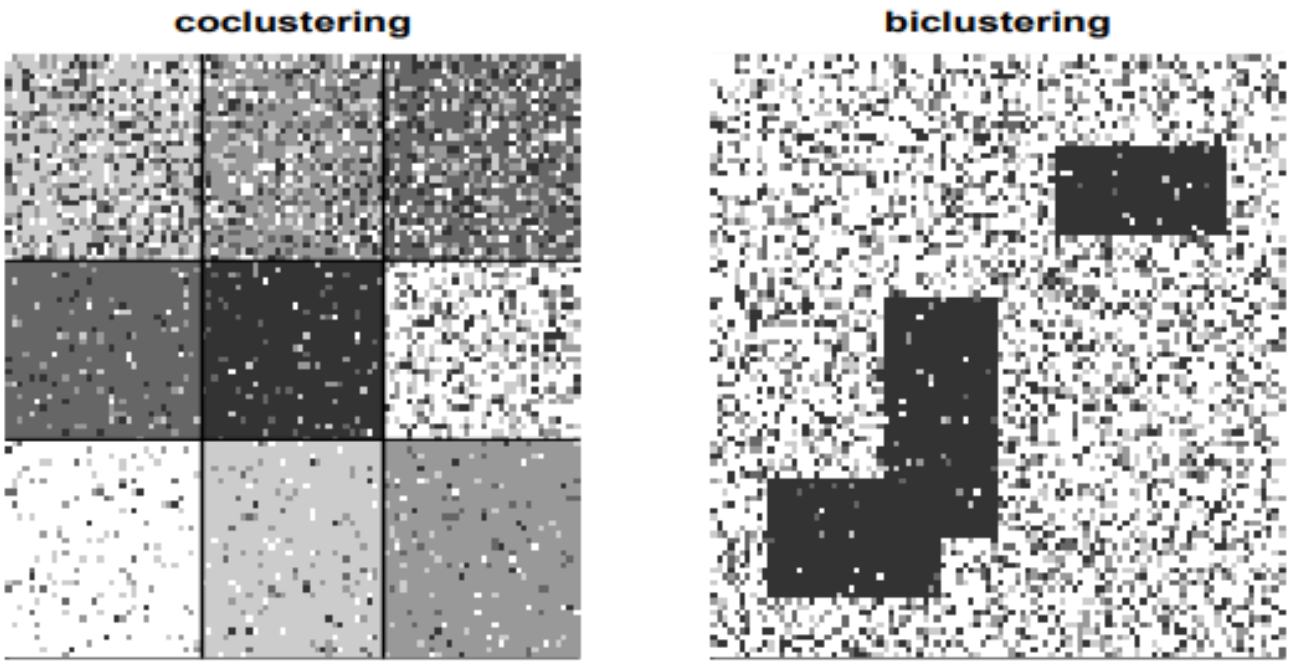
From clustering to co-clustering



[Govaert, 2011]

Bi-clustering

- A generalization of co-clustering
- Look for submatrices of $\boldsymbol{x}$ which are homogeneous
- We do not consider bi-clustering here



Link between co-clustering and PCA/kmeans[2]

Interpreting LBM as an MBC dimension reduction method PCA is certainly the most emblematic reduction dimension method for numerical data sets $\mathbf{x}$ (we consider in the following that $\mathbf{x}$ is centered in columns). It proceeds in two steps. First, it expresses each data unit $\mathbf{x}_i$ of the initial data set $\mathbf{x}$ in a new vector basis $(\mathbf{u}^1, \ldots, \mathbf{u}^d)$, ordered in decreasing value of the preserved variance, with respective coordinates $(a_i^1, \ldots, a_i^d)$ and where each $\mathbf{u}^j$ is defined by a linear combination of the canonical vector basis $(\mathbf{e}^1, \ldots, \mathbf{e}^d)$, namely $\mathbf{u}^j = \sum_{j'=1}^d b_{j'} \mathbf{e}^{j'}$. Second, it selects a reduced number of these new coordinates only (say J), resulting in a new data set of smaller dimension ($J < d$). This PCA sequential procedure for the i-th data individual $\mathbf{x}_i$ can be expressed as follows:

$$\mathbf{x}_i = \sum_{j=1}^{d} x_i^j \mathbf{e}^j = \sum_{j=1}^{d} a_i^j \mathbf{u}^j \approx \sum_{j=1}^{J} a_i^j \mathbf{u}^j. \tag{7}$$

Consider now the (very) specific co-clustering case reduced to $K = 1$, thus meaning that only a variable clustering in L column clusters is performed. Let $\varepsilon_i^j \sim \mathcal{N}(0, \sigma^2_{\tilde{w}_j})$ in an i.i.d manner, where $\tilde{w}_j$ is the cluster index such that $\tilde{w}_j = \ell \Leftrightarrow w_{j\ell} = 1$. We can write, with $\mathbf{v}^\ell = \sum_{\{j:\tilde{w}_j=\ell\}} \mathbf{e}^j$ and $\mathbf{r}_i = \sum_{j=1}^d \varepsilon_i^j \mathbf{e}^j$:

$$\mathbf{x}_i = \sum_{j=1}^{d} x_i^j \mathbf{e}^j = \sum_{j=1}^{d} \left(\mu_{\tilde{w}_j} + \varepsilon_i^j\right) \mathbf{e}^j = \sum_{\ell=1}^{L} \mu_\ell \mathbf{v}^\ell + \mathbf{r}_i \approx \sum_{\ell=1}^{L} \mu_\ell \mathbf{v}^\ell. \tag{8}$$

It reinforces the interest of co-clustering for the HD case. . .

[2] C. Biernacki, J. Jacques, C. Keribin (2023). A Survey on Model-Based Co-Clustering: High Dimension and

Outline

Notations

- z_i: the cluster of the row i
- w_j: the cluster of the column j
- (z_i, w_j): the block of the element x_{ij} (row i, column j)

- $\mathbf{z} = (z_1, \ldots, z_n)$: partition of individuals in K custers of rows
- $\mathbf{w} = (w_1, \ldots, w_d)$: partition of variables in L clusters of columns
- $(\mathbf{z}, \mathbf{w})$: bi-partition of the whole data set $\mathbf{x}$
- Both space partitions are respectively denoted by $\mathcal{Z}$ and $\mathcal{W}$

Restriction

All variables are of the same kind (see discussion at the end)

A geometric approach

- Example in the continuous case: $\boldsymbol{x} \in \mathbb{R}^{n \times d}$
- It could be possible to define a within-block inertia criterion

$$W(\boldsymbol{z}, \boldsymbol{w}) = \underbrace{\sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n} \sum_{j=1}^{d}}_{\sum_{i,j,k,l}} z_{ik} w_{jl} \|x_{ij} - \mu_{kl}\|^2$$

with μ_{kl} the center of the block (k, l)

$$\mu_{kl} = \frac{1}{n_{kl}} \sum_{i,j} z_{ik} w_{jl} x_{ij}$$

where $n_{kl} = \sum_{ij} z_{ik} w_{jl}$ is the sample size of the block (k, l)

But we know now that it hides some model-based assumptions...

The latent block model (LBM)

- Generalization of some existing non-probabilistic methods
- Extend the latent class principle of local (or conditional) independence
- Thus x_{ij} is assumed to be independent once z_i and w_j are fixed ($\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kl})$):

$$\mathrm{p}(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\alpha}) = \prod_{i,j} \mathrm{p}(x_{ij}; \boldsymbol{\alpha}_{z_i w_j})$$

- $\boldsymbol{\pi} = (\pi_k)$: vectors of proba. π_k that a row belongs to the kth row cluster
- $\boldsymbol{\rho} = (\rho_k)$: vectors of proba. ρ_k that a row belongs to the lth column cluster
- Independence between all z_i and w_j
- Extension of the traditional mixture model-based clustering ($\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kl})$):

$$\mathrm{p}(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,j} \pi_{z_i} \rho_{w_j} \mathrm{p}(x_{ij}; \boldsymbol{\alpha}_{z_i w_j})$$

Distribution for different kinds of data

[Govaert and Nadif, 2014] The pdf $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{z_i w_j})$ depends on the kind of data x_{ij}:

- Binary data: $x_{ij} \in \{0, 1\}$, $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{B}(\alpha_{kl})$
- Categorical data with m levels:
  $\mathbf{x}_{ij} = \{x_{ijh}\} \in \{0, 1\}^m$ with $\sum_{h=1}^{m} x_{ijh} = 1$ and $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{M}(\boldsymbol{\alpha}_{kl})$ with $\boldsymbol{\alpha}_{kl} = \{\alpha_{kjh}\}$
- Count data: $x_i^j \in \mathbb{N}$, $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{P}(\mu_k \nu_l \gamma_{kl})$[3]
- Continuous data: $x_i^j \in \mathbb{R}$, $\mathrm{p}(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{N}(\mu_{kl}, \sigma_{kl}^2)$

[3] The Poisson parameter is here split into μ_k and ν_l the effects of the row k and the column l respectively and γ_{kl} the effect of the block kl. Unfortunately, this parameterization is not identifiable. It is therefore not possible to estimate simultaneously μ_k, ν_l and γ_{kl} without imposing further constraints. Constraints $\sum_k \pi_k \gamma_{kl} = \sum_l \rho_l \gamma_{kl} = 1$ and $\sum_k \mu_k = 1, \sum_l \nu_l = 1$ are a possibility.

Extreme parsimony ability

Model	Number of parameters
Binary	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL$
Categorical	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL(m-1)$
Contingency	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL$
Continuous	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + 2KL$

Very parsimonious so well suitable for the (ultra) HD setting

$$\text{nb. param.}_{\text{HD}} = \text{nb. param.}_{\text{classic}} \times \frac{L}{d}$$

Other advantage: stay in the canonical space thus meaningful for the end-user

Binary illustration: easy interpretation

[Govaert, 2011]

	abcdefghij
y_1	1010001101
y_2	0101110011
y_3	1000001100
y_4	1010001100
y_5	0111001100
y_6	0101110101
y_7	0111110111
y_8	1100111011
y_9	0100110000
y_{10}	1010101101
y_{11}	1010001100
y_{12}	1010000100
y_{13}	1010001101
y_{14}	0010011100
y_{15}	0010010100
y_{16}	1111001100
y_{17}	0101110011
y_{18}	1010011101
y_{19}	1010001000
y_{20}	1100101100

Raw data

	a c g h	*b d e f i j*
y_2	0 0 0 0	1 1 1 1 1 1
y_6	0 0 0 1	1 1 1 1 0 1
y_7	0 1 0 1	1 1 1 1 1 1
y_8	1 0 1 0	1 0 1 1 1 1
y_9	0 0 0 0	1 0 1 1 0 0
y_{17}	0 0 0 0	1 1 1 1 1 1
y_1	1 1 1 1	0 0 0 0 0 1
y_3	1 0 1 1	0 0 0 0 0 0
y_4	1 1 1 1	0 0 0 0 0 0
y_5	0 1 1 1	1 1 0 0 0 0
y_{10}	1 1 1 1	0 0 1 0 0 1
y_{11}	1 1 1 1	0 0 0 0 0 0
y_{12}	1 1 0 1	0 0 0 0 0 0
y_{13}	1 1 1 1	0 0 0 0 0 1
y_{14}	0 1 1 1	0 0 0 1 0 0
y_{15}	0 1 0 1	0 0 0 1 0 0
y_{16}	1 1 1 1	1 1 0 0 0 0
y_{18}	1 1 1 1	0 0 0 1 0 1
y_{19}	1 1 1 0	0 0 0 0 0 0
y_{20}	1 0 1 1	1 0 1 0 0 0

Permuted data
(rows/columns)

mode

0	1
1	0

Summary

0.86	0.79
0.83	0.86

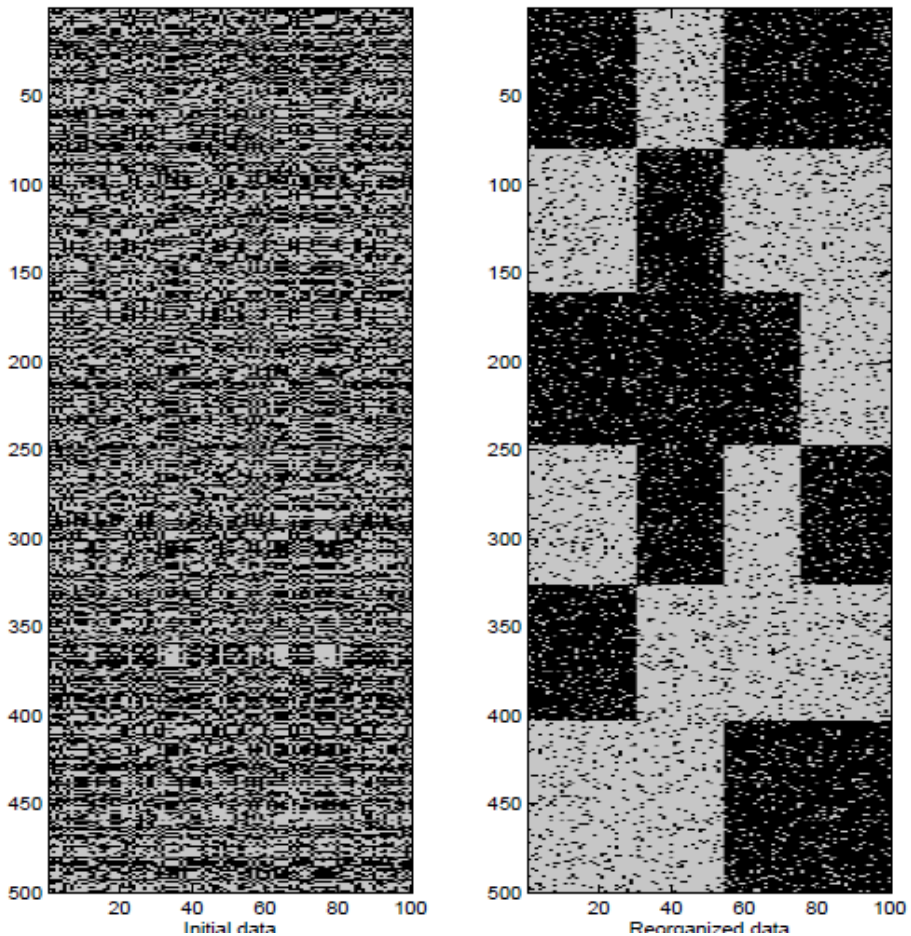
Homogeneity

proba=mode

iid Bin(0.83)

Binary illustration: user-friendly visualization

[Govaert, 2011]



$n = 500,\ d = 10,\ K = 6,\ L = 4$

Other kind of data: ordinal

[Jacques and Biernacki, 2018]

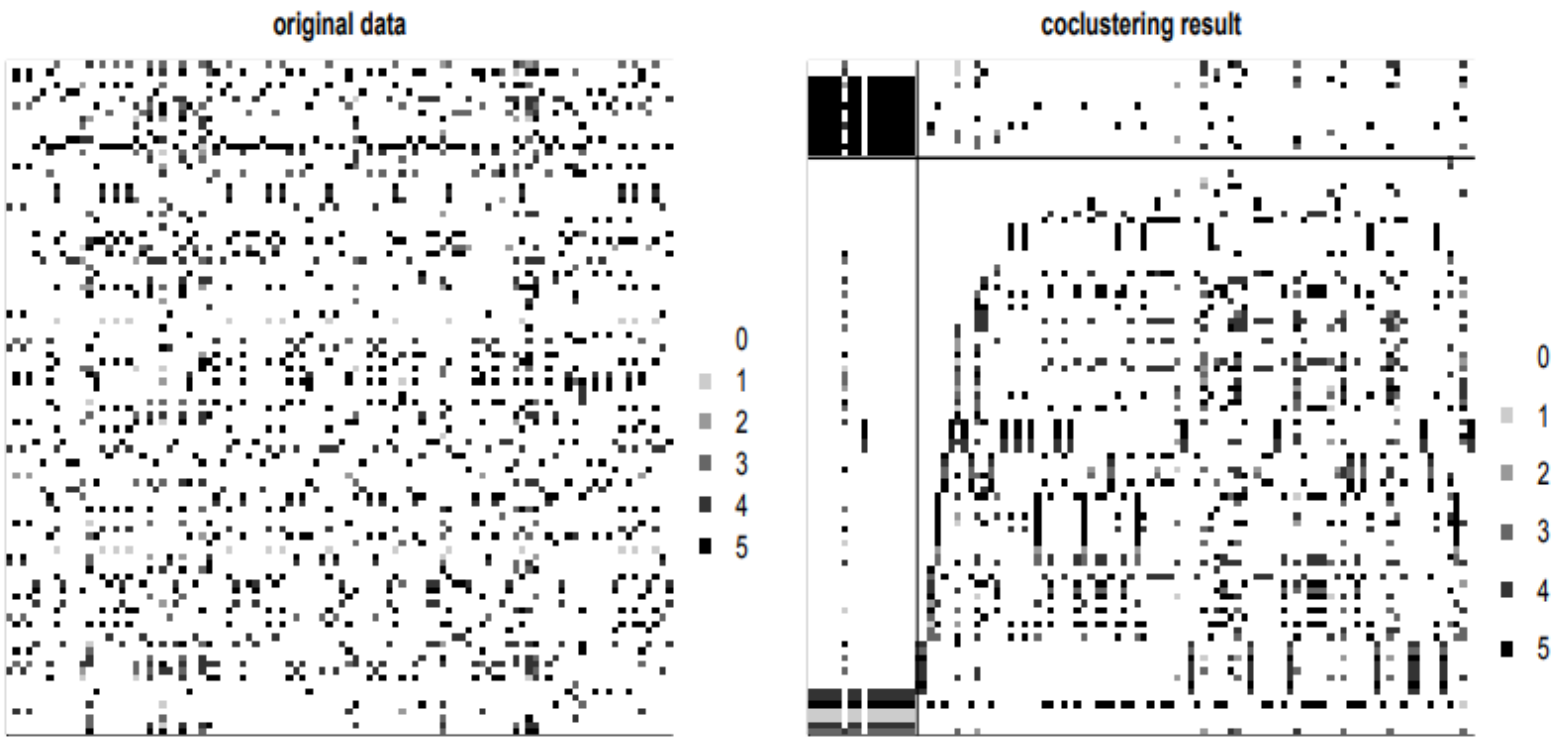
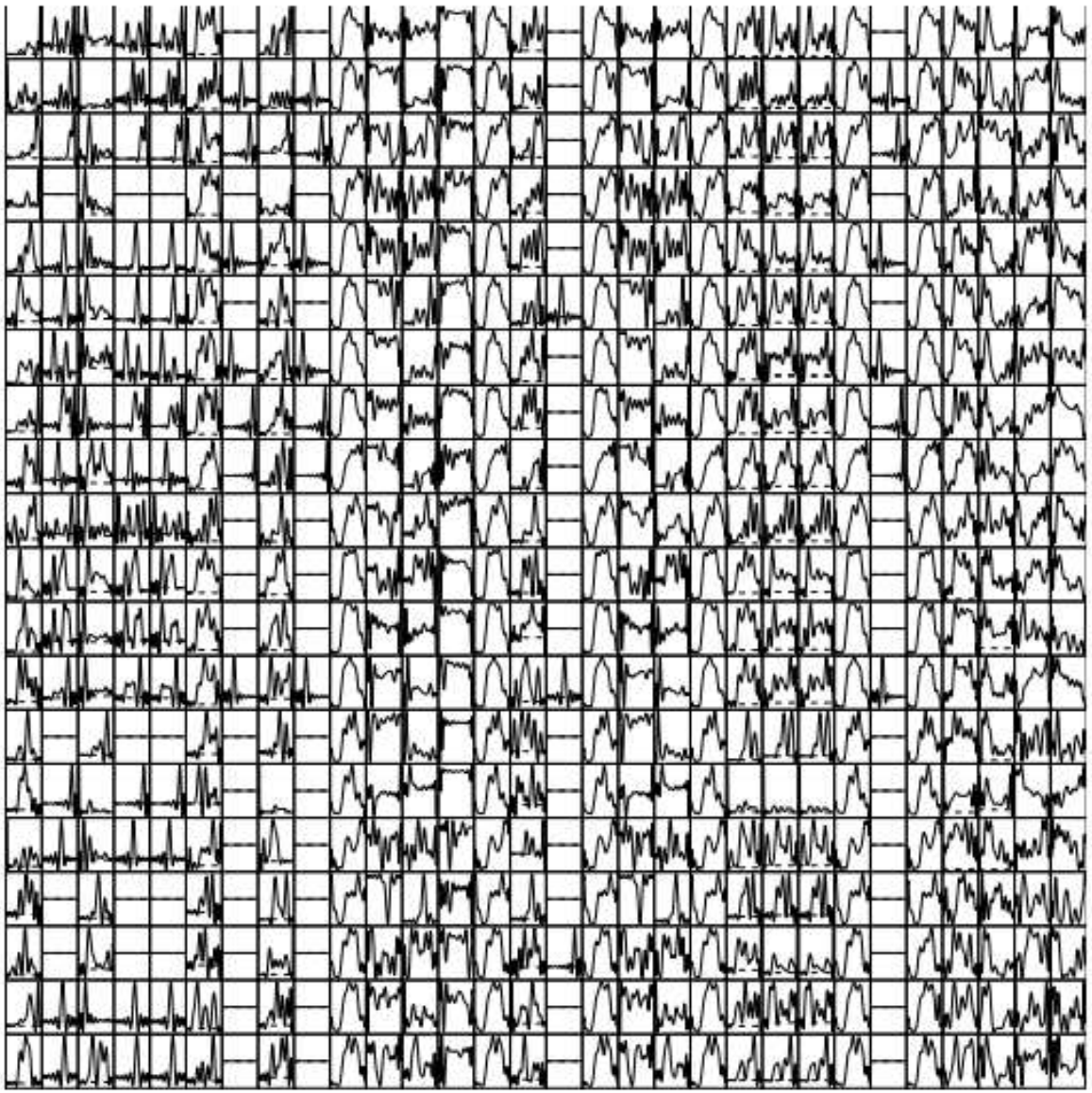


Figure 11: Top 100 Amazone Fine Food Review data (left) and co-clustering result (right).

Other kind of data: functional

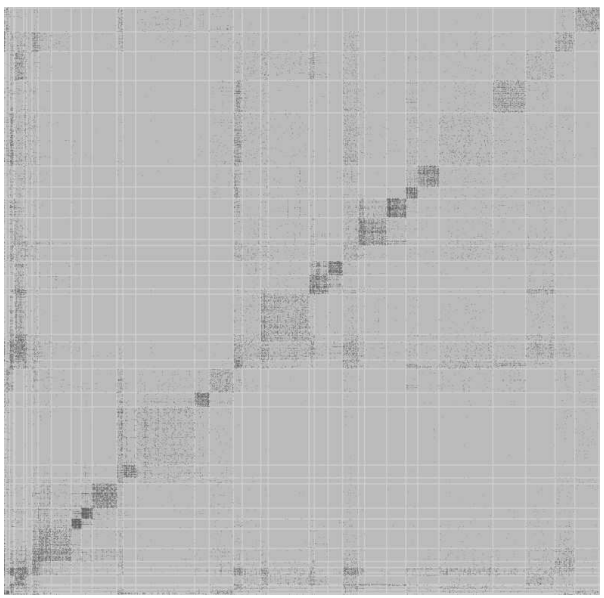
[Jacques, 2016]



Other kind of data: image

Particular case: graph clustering

Stochastic Block Model (SBM): adjacency matrix with $n = d$ and $K = L$



Outline

MLE estimation: log-likelihood(s)

- Remember Lesson 3: first estimate $\boldsymbol{\theta}$, then deduce estimate of $(\boldsymbol{z}, \boldsymbol{w})$
- Observed log-likelihood: $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \ln \mathrm{p}(\boldsymbol{x}; \boldsymbol{\theta})$
- MLE:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{x})$$

- Complete log-likelihood:

$$\begin{aligned} \ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}, \mathbf{w}) &= \ln \mathrm{p}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}; \boldsymbol{\theta}) \\ &= \sum_{i,k} z_{ik} \log \pi_k + \sum_{k,l} w_{jl} \log \rho_l + \sum_{i,j,k,l} z_{ik} w_{jl} \log \mathrm{p}(x_i^j; \boldsymbol{\alpha}_{kl}) \end{aligned}$$

Be careful with asymptotics. . .

If $\ln(d)/n \to 0$, $\ln(n)/d \to 0$ when $n \to \infty$ and $d \to \infty$, then the MLE is consistent [Brault *et al.*, 2017]

MLE estimation: EM algorithm

- E-step of EM (iteration q):

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) &= E[\ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}, \mathbf{w}) | \boldsymbol{x}; \boldsymbol{\theta}^{(q)}] \\ &= \sum_{i,k} \underbrace{p(z_i = k | \mathbf{x}; \boldsymbol{\theta}^{(q)})}_{t_{ik}^{(q)}} \ln \pi_k + \sum_{j,l} \underbrace{p(w_i = l | \mathbf{x}; \boldsymbol{\theta}^{(q)})}_{s_{jl}^{(q)}} \ln \rho_l \\ &\quad + \sum_{i,j,k,l} \underbrace{p(z_i = k, w_j = l | \mathbf{x}; \boldsymbol{\theta}^{(q)})}_{e_{ijkl}^{(q)}} \ln \mathrm{p}(x_{ij}; \boldsymbol{\alpha}_{kl}) \end{aligned}$$

- M-step of EM (iteration q): classical. For instance, for the Bernoulli case, it gives

$$\pi_k^{(q+1)} = \frac{\sum_i t_{ik}^{(q)}}{n}, \quad \rho_l^{(q+1)} = \frac{\sum_j s_{jl}^{(q)}}{d}, \quad \alpha_{kl}^{(q+1)} = \frac{\sum_{i,j} e_{ijkl}^{(q)} x_{ij}}{\sum_{i,j} e_{ijkl}^{(q)}}$$

MLE: intractable E step

$e_{ijkl}^{(q)}$ is usually intractable...

- Consequence of dependency between $\boldsymbol{x}_{ij}$s (link between rows and columns)
- Involve $K^n L^d$ calculus (number of possible blocks)
- Example: if $n = d = 20$ and $K = L = 2$ then 10^{12} blocks
- Example (cont'd): 33 years with a computer calculating 100,000 blocks/second

Alternatives to EM

- Variational EM (numerical approx.): conditional independence assumption

$$\mathrm{p}(\mathbf{z}, \mathbf{w}|\mathbf{x}; \boldsymbol{\theta}) \approx \mathrm{p}(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})\mathrm{p}(\mathbf{w}|\mathbf{x}; \boldsymbol{\theta})$$

- SEM-Gibbs (stochastic approx.): replace E-step by a S-step approx. by Gibbs

$$\mathbf{z}|\mathbf{x}, \mathbf{w}; \boldsymbol{\theta} \quad \text{and} \quad \mathbf{w}|\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}$$

MLE: variational EM (1/2)

- Use a general variational result from [Hathaway, 1985]
- Maximizing $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ on $\boldsymbol{\theta}$ is equivalent to maximize $\tilde{\ell}_c(\boldsymbol{\theta}; \boldsymbol{x}, \mathbf{e})$ on $(\boldsymbol{\theta}, \mathbf{e})$

$$\tilde{\ell}_c(\boldsymbol{\theta}; \boldsymbol{x}, \mathbf{e}) = \sum_{i,k} t_{ik} \ln \pi_k + \sum_{j,l} s_{jl} \ln \rho_l + \sum_{i,j,k,l} e_{ijkl} \ln \mathrm{p}(x_{ij}; \boldsymbol{\alpha}_{kl})$$

  where $\mathbf{e} = (e_{ijkl})$, $e_{ijkl} \in \{0, 1\}$, $\sum_{k,l} e_{ijkl} = 1$, $t_{ik} = \sum_{j,l} e_{ijkl}$, $s_{jl} = \sum_{i,k} e_{ijkl}$

- Of course maximizing $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ or $\tilde{\ell}_c(\boldsymbol{\theta}; \boldsymbol{x}, \mathbf{e})$ are both intractable
- Idea: restriction on $\mathbf{e}$ to obtain tractability $\mathbf{e}_{ijkl} = t_{ik} s_{jl}$
- New variables are thus now $\mathbf{t} = (t_{ik})$ and $\mathbf{s} = (s_{jl})$
- As a consequence, it is a maximization of a lower bound of the max. likelihood

$$\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{x}) \geq \max_{\boldsymbol{\theta}, \mathbf{t}, \mathbf{s}} \tilde{\ell}_c(\boldsymbol{\theta}; \boldsymbol{x}, \mathbf{e})$$

MLE: variational EM (2/2)

Approximated E-step

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) \approx \sum_{i,k} t_{ik}^{(q)} \ln \pi_k + \sum_{j,l} s_{jl}^{(q)} \ln \rho_l + \sum_{i,j,k,l} t_{ik}^{(q)} s_{jl}^{(q)} \ln \mathrm{p}(x_{ij}; \boldsymbol{\alpha}_{kl})$$

- We called it now VEM
- Also known as mean field approximation
- Consistency of the variational estimate [Brault *et al.*, 2017]

MLE: local maxima

- More local maxima than in classical mixture models
- It is a consequence of many more latent variables (blocks)
- Thus: either many VEM runs, or use the SEM-Gibbs algorithm

MLE: SEM-Gibbs

- We have already seen the SEM algorithm in Lesson 3 (thus we do not detail more)
- It limits dependency to starting point, so it limits local maxima
- The S-step: a draw $(\boldsymbol{z}^{(q)}, \boldsymbol{w}^{(q)}) \sim \mathrm{p}(\boldsymbol{z}, \boldsymbol{w}|\boldsymbol{x}; \boldsymbol{\theta}^{(q)})$ instead an expectation
- But it is still intractable, thus use a Gibbs algorithm to approx. this draw

Approximated S-step

Two easy draws

$$\boldsymbol{z}^{(q)} \sim \mathrm{p}(\boldsymbol{z}|\boldsymbol{w}^{(q-1)}, \boldsymbol{x}; \boldsymbol{\theta}^{(q)})$$

and

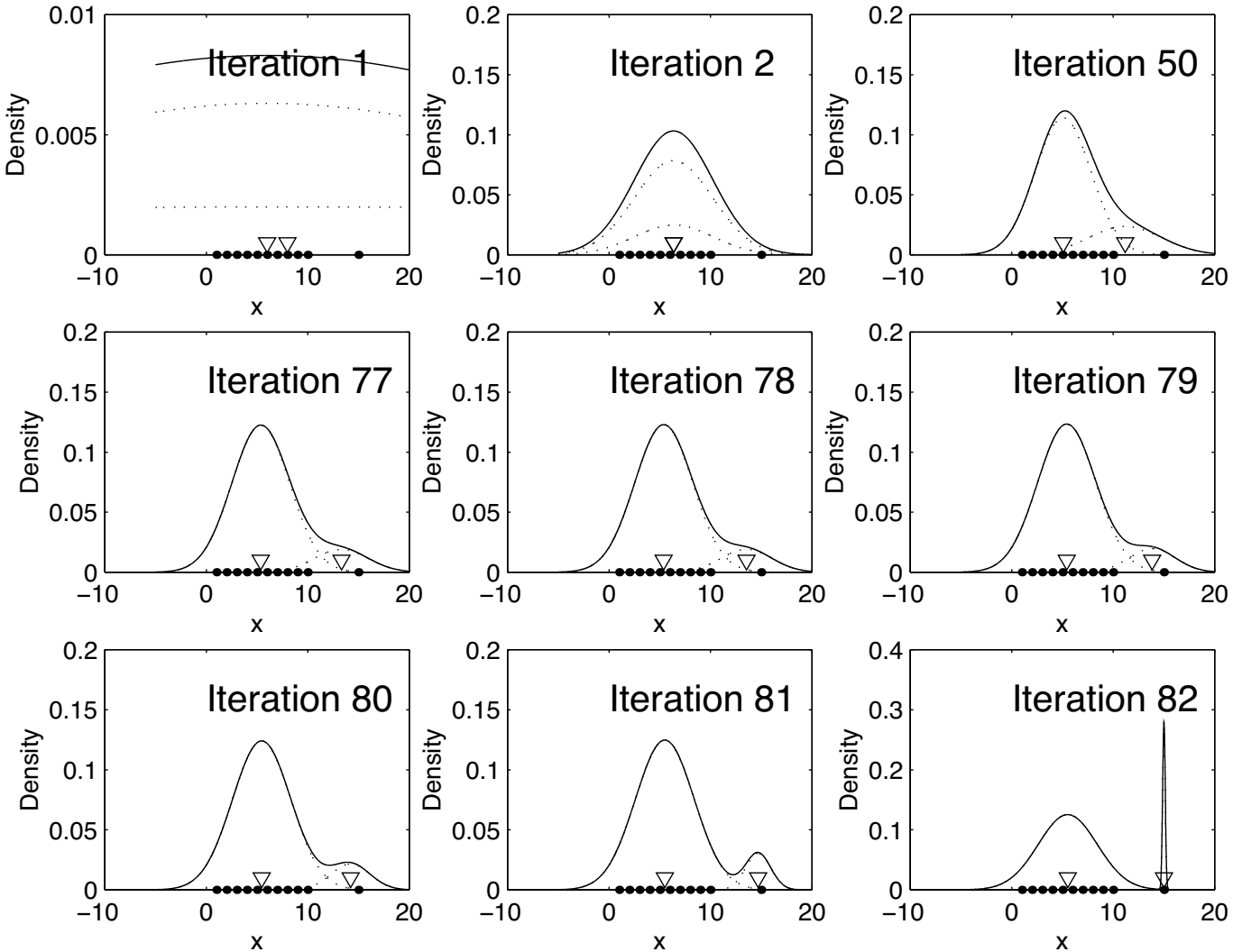
$$\boldsymbol{w}^{(q)} \sim \mathrm{p}(\boldsymbol{w}|\boldsymbol{z}^{(q)}, \boldsymbol{x}; \boldsymbol{\theta}^{(q)})$$

- Rigorously speaking, many draws within the S-step should be performed
- Indeed, Gibbs has to reach a stochastic convergence
- In practice it works well while saving computation time

MLE: degeneracy

- More degenerate situations than in classical mixture models
- It is again a consequence of many more latent variables (blocks)
- The Bayesian regularization (instead MLE) can be an answer

Illustration of a degenerate situation



Bayesian estimation: pitch

- Everything passes by the posterior distribution of $\boldsymbol{\theta}$

$$\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) \propto \underbrace{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})}_{\text{log-likelihood}} \underbrace{\mathrm{p}(\boldsymbol{\theta})}_{\text{prior}}$$

- Then, take (for instance) the MAP as a $\boldsymbol{\theta}$ estimate (use a VEM like algo. . .)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$$

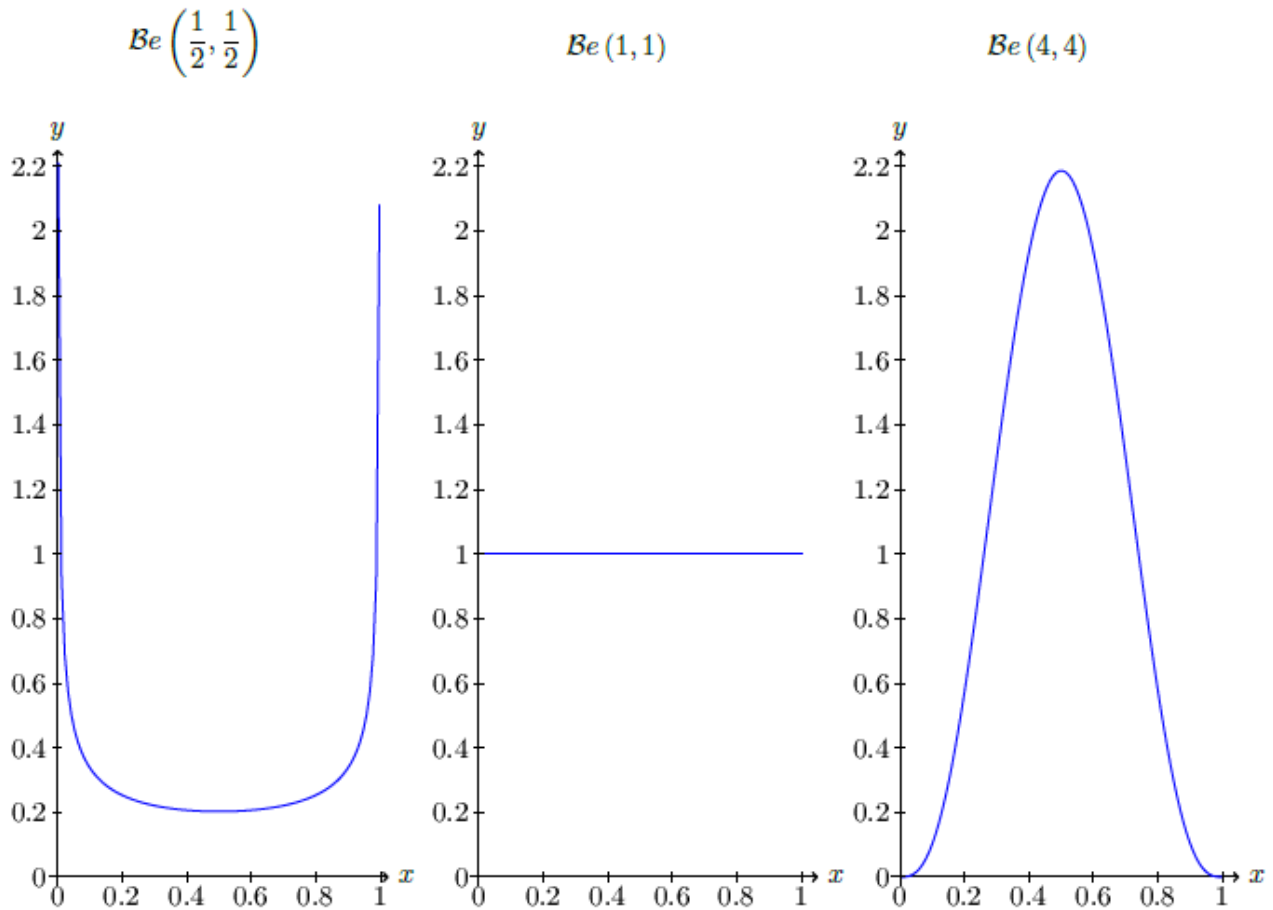
Bayesian estimation: limiting degeneracy

- Interest for avoiding degeneracy is the prior: it acts as a penalization term
- Typical choices are Dirichlet for $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ (with independence between $\boldsymbol{\pi}$, $\boldsymbol{\rho}$, $\boldsymbol{\alpha}$)

$$\mathrm{p}(\boldsymbol{\theta}) = \underbrace{\mathrm{p}(\boldsymbol{\pi})}_{D_K(a,\ldots,a)} \times \underbrace{\mathrm{p}(\boldsymbol{\rho})}_{D_L(a,\ldots,a)} \times \underbrace{\mathrm{p}(\boldsymbol{\alpha})}_{\text{model dependent}}$$

- The Dirichlet distribution is conjugate, thus easy calculus
- Control degeneracy frequency with the a value:
  - $a = 1$: uniform prior, so $\hat{\boldsymbol{\theta}}$ is strictly the MLE (no regularisation)
  - $a = 1/2$: Jeffreys prior, classical (no informative prior) but may favor degeneracy
  - $a = 4$: a rule of thumb working well for limiting degeneracy frequency

Bayesian estimation: prior overview



Block estimation: estimate

- Once we have a parameter estimate $\hat{\boldsymbol{\theta}}$, we need to have an block estimate $(\hat{\boldsymbol{z}}, \hat{\boldsymbol{w}})$
- But MAP not directly available because of the following maximization difficulty

$$(\hat{\boldsymbol{z}}, \hat{\boldsymbol{w}}) = \arg\max_{(\boldsymbol{z},\boldsymbol{w})} \underbrace{\mathrm{p}(\boldsymbol{z}, \boldsymbol{w}|\boldsymbol{x}; \hat{\boldsymbol{\theta}})}_{\text{intractable}}$$

- Instead the following (easily, as classical mixtures) estimates are usually retained

$$\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \mathrm{p}(\boldsymbol{z}|\boldsymbol{x}; \hat{\boldsymbol{\theta}}) \quad \text{and} \quad \hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \mathrm{p}(\boldsymbol{w}|\boldsymbol{x}; \hat{\boldsymbol{\theta}})$$

Block estimation: evaluation

- Empirical error rate between blocks:

$$\mathrm{err}_{\mathrm{blocks}}\Big(\underbrace{(\boldsymbol{z}, \boldsymbol{w})}_{\text{“True” blocks}} , \underbrace{(\hat{\boldsymbol{z}}, \hat{\boldsymbol{w}})}_{\text{Estimated blocks}} \Big) = \mathrm{err}(\boldsymbol{z}, \hat{\boldsymbol{z}}) + \mathrm{err}(\boldsymbol{w}, \hat{\boldsymbol{w}}) - \mathrm{err}(\boldsymbol{z}, \hat{\boldsymbol{z}}) \times \mathrm{err}(\boldsymbol{w}, \hat{\boldsymbol{w}})$$

- Rand index between blocks: it exists also a recent definition...

Block estimation: consistency

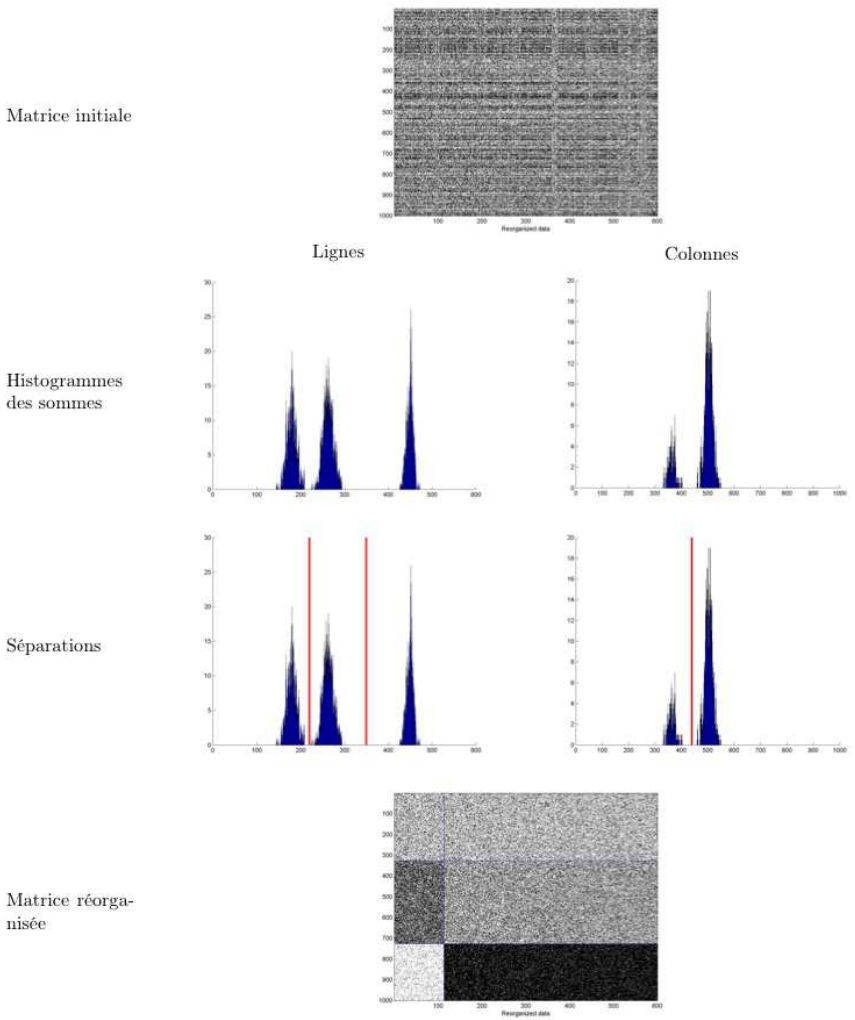
[Mariadassou and Matias, 12]

$$\underbrace{\hat{\boldsymbol{\theta}} \overset{n,d\to\infty}{\longrightarrow} \boldsymbol{\theta}^*}_{\text{we have seen that...}} \quad \Rightarrow \quad \underbrace{\mathrm{p}(\hat{\mathbf{z}} = \mathbf{z}^*, \hat{\mathbf{w}} = \mathbf{w}^* | \mathbf{x}; \hat{\boldsymbol{\theta}}) \overset{n,d\to\infty}{\longrightarrow} 1}_{\text{exact bi-partition retrieval!}}$$

Thus we retrieve the HD clustering blessing...

Block estimation: non asymptotic properties (1/2)

Binary case: marginals seems so simple mixtures! [Brault, 14]



Block estimation: non asymptotic properties (2/2)

[Brault, 14]

- Probability of x_{ij} with no regard to the column membership is Bernoulli

$$\mathrm{p}(x_{ij} = 1 | z_{ik} = 1) = \tau_k = \sum_{l=1}^{L} \alpha_{kl} \rho_l$$

- Thus marginal distribution of x_{ij} is a mixture (indep. of x_{ij} cond. $z_{ik} = 1$)

$$\left(\sum_j x_{ij} \right) | z_{ik} = 1 \sim \mathrm{B}(d, \tau_k)$$

- Control of error on this partition mixture estimate $\hat{\mathbf{z}}^{mix}$ of binomial distributions

$$\mathrm{p}(\hat{\mathbf{z}}^{mix} \neq \mathbf{z}^*) \leq 2n \exp\left\{ -\frac{1}{8} d \Big[\underbrace{\min_{k \neq k'} |\tau_k - \tau_{k'}|}_{\text{overlap}} \Big] \right\} + K(1 - \min_k \pi_k)^n$$

- We retrieve also consistency for very high dimension with constraint

$$\ln(n) = o(d)$$

Illustration: document clustering (1/2)

- Mixture of 1033 medical summaries and 1398 aeronautics summaries
- Lines: 2431 documents
- Columns: present words (except stop), thus 9275 unique words
- Data matrix: cross counting document×words
- Poisson model

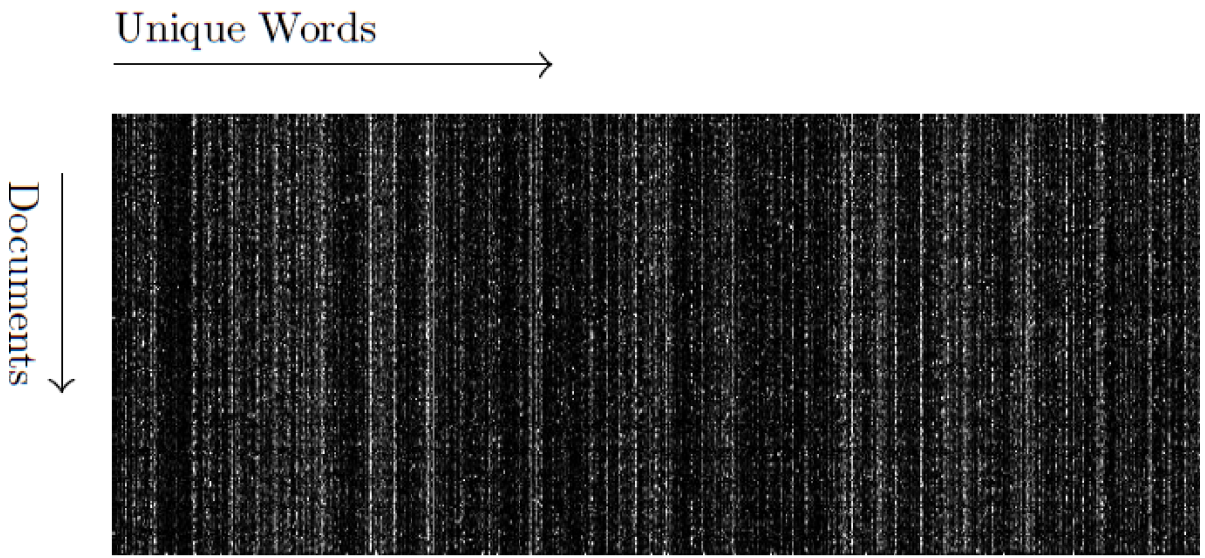
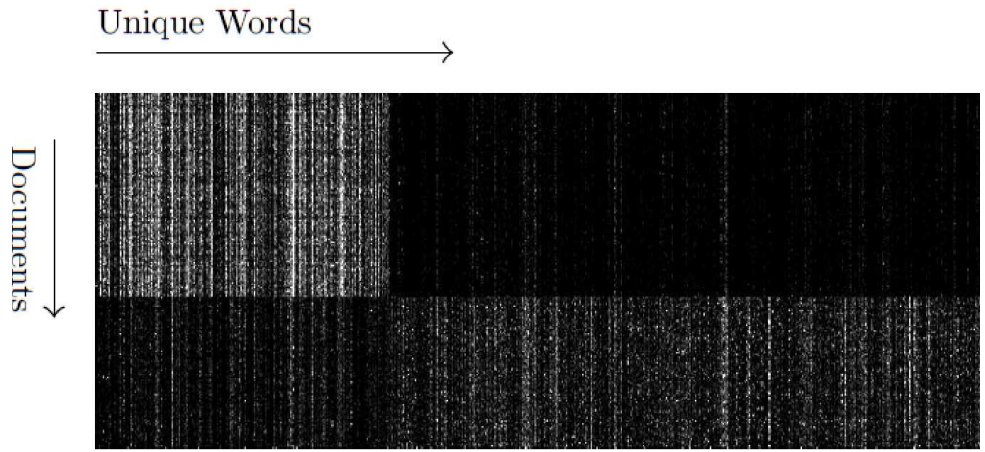


Illustration: document clustering (2/2)



Results with 2×2 blocks

	Medline	Cranfield
Medline	1033	0
Cranfield	0	1398

Experiment illustrates previous theory: HD clustering is blessing

Outline

Models in competition

$\mathbf{m} = (K, L)$ typically, but not restricted to

BIC criterion: two difficulties

- Difficult 1: which BIC definition because of the double asymptotic on n and d?
- Difficult 2: the observed log-likelihood value is intractable

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{(\boldsymbol{z}, \boldsymbol{w}) \in \mathcal{Z} \times \mathcal{W}} \mathrm{p}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}; \boldsymbol{\theta})$$

Could be estimated by harmonic mean but time consuming and high variance

ICL criterion: overcome both difficulties

- ICL uses complete likelihood thus no intractability

$$\text{ICL} = \ln \mathrm{p}(x, \hat{\mathbf{z}}, \hat{\mathbf{w}}) = \ln \mathrm{p}(\mathbf{x}|\hat{\mathbf{z}}, \hat{\mathbf{w}}) + \ln \mathrm{p}(\hat{\mathbf{z}}) + \ln \mathrm{p}(\hat{\mathbf{w}})$$

- Multinomial case (r levels): [Keribin *et al., 2014*]
  - Derive an exact (non-asymptotic) ICL version
  - Deduce an asymptotic approximation of ICL

$$\text{ICLbic} = \ell_c(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\mathbf{z}}, \hat{\mathbf{w}}) - \frac{K-1}{2}\ln(n) - \frac{L-1}{2}\ln(d) - \frac{KL(r-1)}{2}\ln(nd)$$

- We can make a conjecture for the general case

$$\text{ICLbic} = \ell_c(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\mathbf{z}}, \hat{\mathbf{w}}) - \frac{K-1}{2}\ln(n) - \frac{L-1}{2}\ln(d) - \frac{KL\nu_{\boldsymbol{\alpha}_{kl}}}{2}\ln(nd)$$

ICL criterion: consistency

- We can obtain a BIC expression from ICLbic

$$\begin{aligned} \text{BIC} &= \text{ICLbic} - \ln \mathrm{p}(\hat{\mathbf{z}}, \hat{\mathbf{w}} | \mathbf{x}; \hat{\boldsymbol{\theta}}) \\ &= \underbrace{\ell(\hat{\boldsymbol{\theta}}; \mathbf{x})}_{\text{difficult}} - \frac{K-1}{2} \ln(n) - \frac{L-1}{2} \ln(d) - \frac{KL(m-1)}{2} \ln(nd) \end{aligned}$$

- [Brault *et al.*, 2017] establish that asymptotically on n and d

$$\text{“}\ell(\hat{\boldsymbol{\theta}}; \mathbf{x}) = \ell_c(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\boldsymbol{z}}, \hat{\boldsymbol{w}})''$$

- Thus, since BIC is consistent, ICL is also consistent

Again the HD clustering blessing is here!

Strategy to smart browsing of (K, L)

[Robert, 2017] Algorithm Bi-KM1

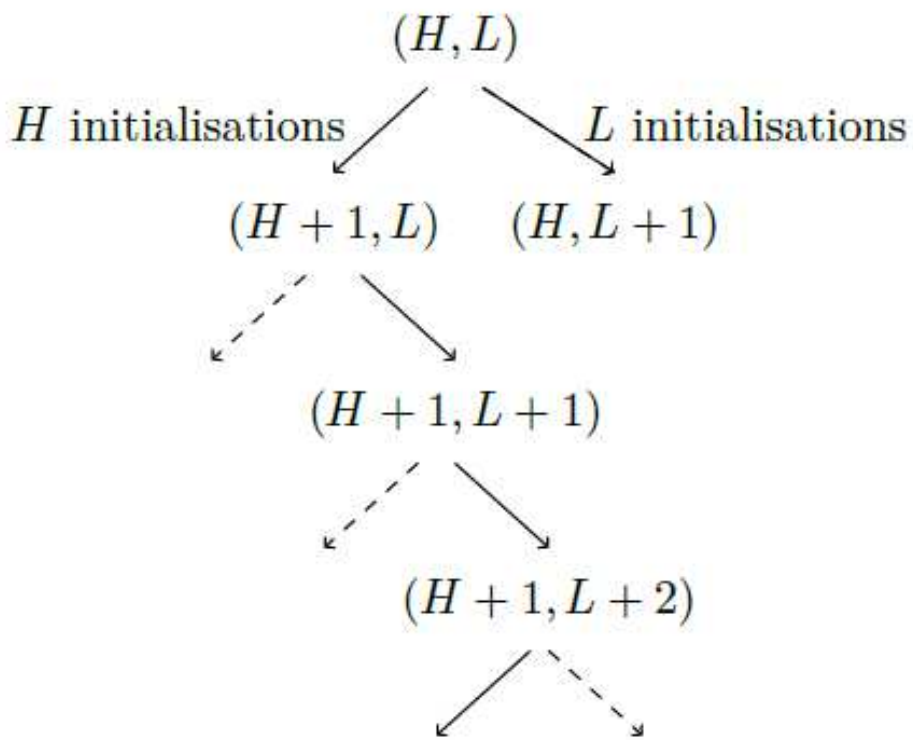


Illustration: discuss the dimension (1/2)

- SPAM E-mail Database[4]
- $n = 4601$ e-mails composed by 1813 "spams" and 2788 "good e-mails"
- $d = 48 + 6 = 54$ continuous descriptors[5]
  - 48 percentages that a given word appears in an e-mail ("make", "you'...)
  - 6 percentages that a given char appears in an e-mail (";", "$"...)
- Transformation of continuous descriptors into binary descriptors

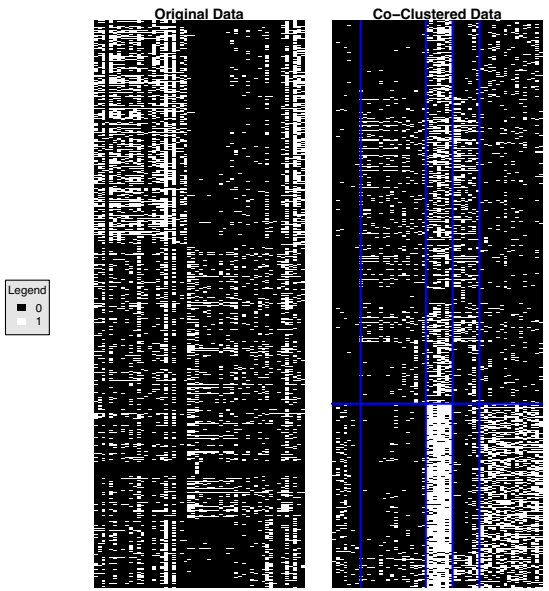
$$x_{ij} = \begin{cases} 1 & \text{if word/char } j \text{ appears in e-mail } i \\ 0 & \text{otherwise} \end{cases}$$

[4] https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/

[5] There are 3 other continuous descriptors we do not use

Illustration: discuss the dimension (2/2)

- Perform co-clustering with $K = 2$ and $L = 5$: ICL=-92,682, err=0.1984



- Perform clustering[6] with $K = 2$: ICL=-89,433, err=0.1837

Thus use preferably co-clustering in the HD setting

[6] Equivalent to co-clustering with $L = 54$

Outline

The Blockcluster package

BlockCluster

launch binder *Visit notebooks repo*

BlockCluster can estimate the parameters of co-clustering models for binary, contingency and continuous data. Simply put, when considering a set of data as rows and columns, BlockCluster will make simultaneous permutations of rows and columns in order to organise the data into homogeneous blocks.

Here is a related publication for this software. Here is the related package on the CRAN.

Two ways to be used

- On the CRAN:
  `https://cran.r-project.org/web/packages/blockcluster/index.html`
- As a note book (nothing to install):
  `https://team.inria.fr/modal/software/notebooks/`

A synthetic example

Blockcluster basic example

Data

Load the CSV data file as dataframe.

In [3]:
```
data <- as.matrix(read.table("blockcluster-example.csv", sep = ","))
head(data)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	⋯	V91	V92	V93	V94	V95	V96	V97	V98	V99	V100
A matrix: 6 × 100 of type int																					
	1	1	1	0	1	1	1	1	1	0	⋯	0	0	0	0	0	0	1	0	1	1
	0	0	1	0	0	0	0	0	0	1	⋯	0	1	0	0	1	1	1	1	0	0
	0	1	1	0	1	1	1	1	1	0	⋯	0	0	1	1	1	0	1	0	0	0
	0	0	0	0	1	0	1	1	0	1	⋯	1	1	1	0	0	1	1	0	0	0
	0	0	0	1	1	1	1	1	1	0	⋯	0	0	0	0	1	1	0	0	0	1
	1	1	1	1	0	0	1	1	0	1	⋯	1	1	0	0	1	1	1	0	0	0

Clustering with Blockcluster

Launch the BlockCluster package.

In [4]:
```
library(blockcluster)
```

```
Loading required package: rtkore

Loading required package: Rcpp


Attaching package: 'rtkore'


The following object is masked from 'package:Rcpp':

    LdFlags


blockcluster version 4.4.3 loaded

----------------
Copyright (C)  <MODAL team @INRIA,Lille & U.M.R. C.N.R.S. 6599 Heudiasyc, UTC>
Please post questions and bugs at: <https://gforge.inria.fr/forum/forum.php?forum_id=11190&group_id=3679>
```

Define the strategy.

Specify the number of groups/clusters to create in rows and coloumns.

In [5]:
```
rowClusters <- 2
columnClusters <- 3
```

Run the coclustering.

In [6]:
```
res <- cocluster(data, datatype = "binary", nbcocluster = c(rowClusters, columnClusters))
```

```
Co-Clustering successfully terminated!
```

A synthetic example

Output's Analysis

Model criterion

This chart represents the criterion value for each model that was built. The lower the value (close to 0) the better the model.

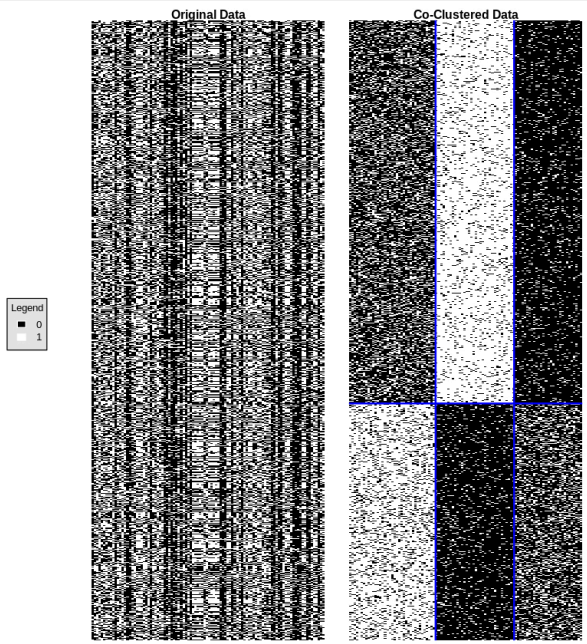
In [7]:
```
res@ICLvalue
```

-45557.0741834614

Cluster Plot

Draw the original data matrix, and the matrix obtained after performing co-clustering

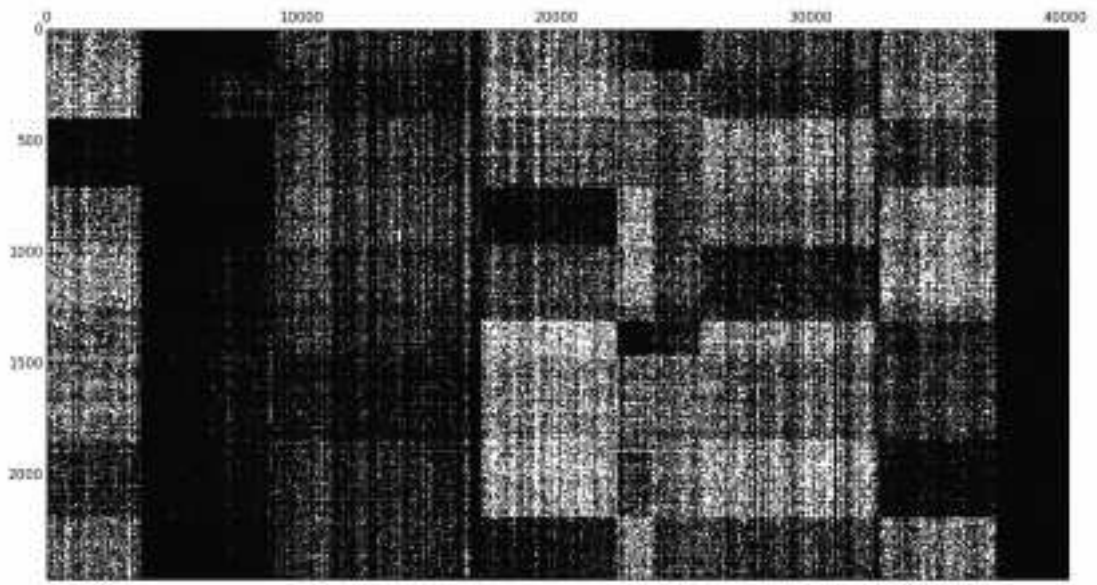
In [8]:
```
plot(res)
```



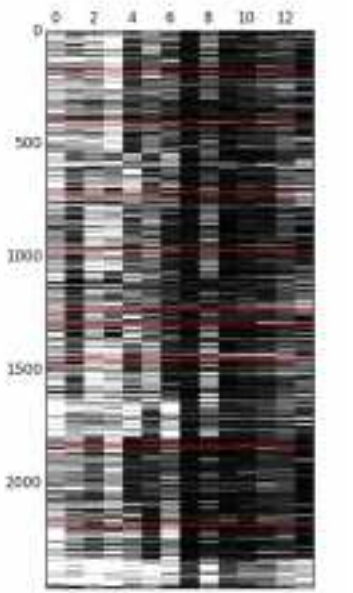
Outline

Co-clustering of mixed data

- Same partitions in lines, disjoint partitions in columns
- Example: data set TED talks, with talks $\times$ (terms,scores)



Poisson



Gaussian