# Clustering : une vision unifiée pour une utilisation éclairée 

# Partie 4 <br> Traitement de la grande dimension \& co-clustering (leçon plus avancée) 

C. Biernacki

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

CAIS
LL Université
de Lille

## Outline

1 HD clustering
2. Modeling

3 Estimating

4 Selecting

5 Experiments

6 To go further

## 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 ${ }^{1}$


[^0]
## HD data: definition (1/2)

An attempt in the non-parametric case
Dataset $\mathbf{x}=\left(\mathbf{x}_{1}, \ldots, \mathbf{x}_{n}\right), \mathbf{x}_{j}$ described by $d$ variables, where $n=o\left(e^{d}\right)$

Justifications:

- To approximate within error $\epsilon$ 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 $\mathbf{x}=\left(\mathbf{x}_{1}, \ldots, \mathbf{x}_{n}\right), \mathbf{x}_{j}$ described by $d$ variables and a model $\mathbf{m}$ with $\nu$ 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 $\kappa, A$ and $C$ such that

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

But $\nu$ 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:

$$
\pi_{1}=\pi_{2}=\frac{1}{2}, \quad \mathbf{X}_{1}\left|z_{11}=1 \sim \mathrm{~N}_{d}(\mathbf{0}, \mathbf{I}), \quad \mathbf{X}_{1}\right| z_{12}=1 \sim \mathrm{~N}_{d}(\mathbf{1}, \mathbf{I})
$$

Components are more and more separated when $d$ grows: $\left\|\mu_{2}-\mu_{1}\right\|_{I}=\sqrt{d} \ldots$

... but density estimation quality decreases with $d$

HD clustering: blessing (1/2)
A two-component $d$-variate Gaussian mixture:

$$
\pi_{1}=\pi_{2}=\frac{1}{2}, \quad \mathbf{X}_{1}\left|z_{11}=1 \sim \mathrm{~N}_{d}(\mathbf{0}, \mathbf{I}), \quad \mathbf{X}_{1}\right| z_{12}=1 \sim \mathrm{~N}_{d}(\mathbf{1}, \mathbf{I})
$$

Each variable provides equal and own separation information

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

... 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:

$$
\mathbf{X}_{1} \left\lvert\, z_{12}=1 \sim N_{d}\left(\left(\begin{array}{llll}
1 & 0 & \ldots & 0
\end{array}\right)^{\prime}, \mathbf{I}\right)\right.
$$

Groups are not separated more when $d$ grows: $\left\|\boldsymbol{\mu}_{2}-\boldsymbol{\mu}_{1}\right\|_{\mathbf{I}}=1 \ldots$

$\ldots$ thus theoretical error is constant $\left(=\Phi\left(-\frac{1}{2}\right)\right)$ 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: $\left\|\boldsymbol{\mu}_{2}-\boldsymbol{\mu}_{1}\right\|_{\boldsymbol{\Sigma}}=1 \ldots$

$\ldots$..thus err ${ }_{\text {theo }}$ is constant $\left(=\Phi\left(-\frac{1}{2}\right)\right)$ and empirical error increases (less) with $d$

## The trade-off bias/variance

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

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

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

## Bias/variance in HD: reduce variance, accept bias

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

$$
\pi_{1}=\pi_{2}=\frac{1}{2}, \quad \mathbf{X}_{1}\left|z_{11}=1 \sim \mathrm{~N}_{d}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \mathbf{X}_{1}\right| z_{12}=1 \sim \mathrm{~N}_{d}(\mathbf{1}, \boldsymbol{\Sigma})
$$

- Each variable provides equal and own separation information
- Theoretical error decreases when $d$ grows: $\operatorname{err}_{\text {theo }}=\Phi\left(-\left\|\boldsymbol{\mu}_{2}-\boldsymbol{\mu}_{1}\right\|_{\boldsymbol{\Sigma}^{-1}} / 2\right)$
- 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]

$$
\boldsymbol{\Sigma}_{k}=\mathbf{B}_{k} \mathbf{B}_{k}^{\prime}+\omega_{k} \boldsymbol{\Lambda}_{k}
$$

where

- $B_{k}$ is a loadings $d \times q$ non-square real matrix $\left(1 \leq q \leq q_{\max }, q_{\max }<d\right)$
- $\omega_{k}$ is a positive real number
- $\boldsymbol{\Lambda}_{k}$ is a $d \times d$ diagonal positive definite matrix such that $\left|\boldsymbol{\Lambda}_{k}\right|=1$

Interpretation $\mathbf{X}_{1} \in \mathbb{R}^{d}$ is generated by a latent variable $\mathbf{Y}_{1} \in \mathbb{R}^{q}$

$$
\mathbf{X}_{1 \mid \mathbf{Y}_{1}, z_{1 k}=1}=\mathbf{B}_{k} \underbrace{\mathbf{Y}_{1}}_{\text {factor }}+\boldsymbol{\mu}_{k}+\boldsymbol{\varepsilon}_{k}
$$

where $\mathbf{Y}_{1} \perp \varepsilon_{k}$ and

$$
\mathbf{Y}_{1} \sim \mathrm{~N}_{q}(\mathbf{0}, \mathbf{I}) \quad \text { and } \quad \varepsilon_{k} \sim \mathrm{~N}_{d}\left(\mathbf{0}, \omega_{k} \boldsymbol{\Lambda}_{k}\right)
$$

## Gaussian mixture of factor analysers (2/2)

Complexity

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

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]

$$
\boldsymbol{\Sigma}_{k}=\mathbf{D}_{k} \boldsymbol{\Delta}_{k} \mathbf{D}_{k}^{\prime}
$$

where

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

with $a_{k j} \geq b_{k}$, for $j=1, \ldots, \delta_{k}$ and $\delta_{k}<d$

Complexity

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

## HD Gaussian models (2/2)



Model selection
■ Models in competition: $\mathbf{m}=\left(\left(\delta_{1}, \ldots, \delta_{K}\right), K\right)+8$ 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]

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

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]

$$
\operatorname{crit}_{\lambda, \rho}=\ell(\boldsymbol{\theta} ; \overline{\mathbf{x}})-\lambda \sum_{k=1}^{K} \sum_{j=1}^{d}\left|\mu_{k j}\right|-\rho \sum_{k=1}^{K} \sum_{\left(j, j^{\prime}\right), j \neq j^{\prime}}^{d}\left|\left(\boldsymbol{\Sigma}_{k}^{-1}\right)_{j j^{\prime}}\right|
$$

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

biclustering



## 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 $\left(\mathbf{u}^{1}, \ldots, \mathbf{u}^{d}\right)$, ordered in decreasing value of the preserved variance, with respective coordinates $\left(a_{i}^{1}, \ldots, a_{i}^{d}\right)$ 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 $\mathbf{u}^{j}=\sum_{j^{\prime}=1}^{d} b_{j^{\prime}} \mathbf{e}^{j^{\prime}}$. 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:

$$
\begin{equation*}
\mathbf{x}_{i}=\sum_{j=1}^{d} x_{i}^{j} \mathbf{e}^{j}=\sum_{j=1}^{d} a_{i}^{j} \mathbf{u}^{j} \approx \sum_{j=1}^{J} a_{i}^{j} \mathbf{u}^{j} . \tag{7}
\end{equation*}
$$

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 $\varepsilon_{i}^{j} \sim \mathcal{N}\left(0, \sigma_{\tilde{w}_{j}}^{2}\right)$ 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 $\mathbf{v}^{\ell}=\sum_{\left\{j: \tilde{w}_{j}=\ell\right\}} \mathbf{e}^{j}$ and $\mathbf{r}_{i}=\sum_{j=1}^{d} \varepsilon_{i}^{j} \mathbf{e}^{j}$ :

$$
\begin{equation*}
\mathbf{x}_{i}=\sum_{j=1}^{d} x_{i}^{j} \mathbf{e}^{j}=\sum_{j=1}^{d}\left(\mu_{\tilde{w}_{j}}+\varepsilon_{i}^{j}\right) \mathbf{e}^{j}=\sum_{\ell=1}^{L} \mu_{\ell} \mathbf{v}^{\ell}+\mathbf{r}_{i} \approx \sum_{\ell=1}^{L} \mu_{\ell} \mathbf{v}^{\ell} . \tag{8}
\end{equation*}
$$

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

## Outline

1 HD clustering

2 Modeling

3 Estimating

4 Selecting

5 Experiments

6 To go further

## Notations

- $\boldsymbol{z}_{i}$ : the cluster of the row $i$
- $\boldsymbol{w}_{j}$ : the cluster of the column $j$
- $\left(\boldsymbol{z}_{i}, \mathbf{w}_{j}\right)$ : the block of the element $\boldsymbol{x}_{i j}$ (row $i$, column $j$ )

■ $\mathbf{z}=\left(z_{1}, \ldots, z_{n}\right)$ : partition of individuals in $K$ custers of rows
■ $\mathbf{w}=\left(\boldsymbol{w}_{1}, \ldots, \boldsymbol{w}_{d}\right)$ : partition of variables in $L$ clusters of columns

- ( $\mathbf{z}, \mathbf{w}$ ): bi-partition of the whole data set $\boldsymbol{x}$
- 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

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

with $\mu_{k l}$ the center of the block $(k, I)$

$$
\mu_{k l}=\frac{1}{n_{k l}} \sum_{i, j} z_{i k} w_{j l} x_{i j}
$$

where $n_{k l}=\sum_{i j} z_{i k} w_{j l}$ 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 $x_{i j}$ is assumed to be independent once $z_{i}$ and $w_{j}$ are fixed $\left(\boldsymbol{\alpha}=\left(\boldsymbol{\alpha}_{k l}\right)\right)$ :

$$
\mathrm{p}(\mathbf{x} \mid \mathbf{z}, \mathbf{w} ; \boldsymbol{\alpha})=\prod_{i, j} \mathrm{p}\left(x_{i j} ; \boldsymbol{\alpha}_{z_{i} w_{j}}\right)
$$

■ $\boldsymbol{\pi}=\left(\pi_{k}\right)$ : vectors of proba. $\pi_{k}$ that a row belongs to the $k$ th row cluster

- $\boldsymbol{\rho}=\left(\rho_{k}\right)$ : vectors of proba. $\rho_{k}$ 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 $\left(\boldsymbol{\alpha}=\left(\boldsymbol{\alpha}_{k l}\right)\right)$ :

$$
\mathrm{p}(\mathrm{x} ; \boldsymbol{\theta})=\sum_{(\mathrm{z}, \mathrm{w}) \in \mathcal{Z} \times \mathcal{W}} \prod_{i, j} \pi_{z_{i}} \rho_{w_{j}} \mathrm{p}\left(x_{i j} ; \boldsymbol{\alpha}_{z_{i} w_{j}}\right)
$$

## Distribution for different kinds of data

[Govaert and Nadif, 2014] The pdf $\mathrm{p}\left(\cdot ; \boldsymbol{\alpha}_{z_{i} w_{j}}\right)$ depends on the kind of data $x_{i j}$ :

- Binary data: $x_{i j} \in\{0,1\}, \mathrm{p}\left(\cdot ; \boldsymbol{\alpha}_{k l}\right)=\mathcal{B}\left(\alpha_{k l}\right)$
- Categorical data with $m$ levels:
$\mathbf{x}_{i j}=\left\{x_{i j h}\right\} \in\{0,1\}^{m}$ with $\sum_{h=1}^{m} x_{i j h}=1$ and $\mathrm{p}\left(\cdot ; \boldsymbol{\alpha}_{k l}\right)=\mathcal{M}\left(\boldsymbol{\alpha}_{k l}\right)$ with $\boldsymbol{\alpha}_{k l}=\left\{\alpha_{k j h}\right\}$
- Count data: $x_{i}^{j} \in \mathbb{N}, \mathrm{p}\left(\cdot ; \boldsymbol{\alpha}_{k l}\right)=\mathcal{P}\left(\mu_{k} \nu_{l} \gamma_{k l}\right)^{3}$

■ Continuous data: $x_{i}^{j} \in \mathbb{R}, \mathrm{p}\left(\cdot ; \boldsymbol{\alpha}_{k l}\right)=\mathcal{N}\left(\mu_{k l}, \sigma_{k l}^{2}\right)$

[^1]
## Extreme parsimony ability

| Model | Number of parameters |
| :---: | :---: |
| Binary | $\operatorname{dim}(\boldsymbol{\pi})+\operatorname{dim}(\boldsymbol{\rho})+K L$ |
| Categorical | $\operatorname{dim}(\boldsymbol{\pi})+\operatorname{dim}(\boldsymbol{\rho})+K L(m-1)$ |
| Contingency | $\operatorname{dim}(\boldsymbol{\pi})+\operatorname{dim}(\boldsymbol{\rho})+K L$ |
| Continuous | $\operatorname{dim}(\boldsymbol{\pi})+\operatorname{dim}(\boldsymbol{\rho})+2 K L$ |

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

|  | $a$ | $c$ | $g$ | $h$ | $b$ |  |  | $d$ | $e$ | $f$ | $i$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |$]$

Permuted data (rows/columns)
mode


Summary


Homogeneity
proba=mode

## Binary illustration: user-friendly visualization

[Govaert, 2011]


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

## Other kind of data: ordinal



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

## Original Data

Co-Clustered Data


## Particular case: graph clustering

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


## Outline

1 HD clustering

2 Modeling

3 Estimating

4 Selecting

5 Experiments

6 To go further

## 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:

$$
\begin{aligned}
\ell_{c}(\boldsymbol{\theta} ; \mathbf{x}, \mathbf{z}, \mathbf{w}) & =\ln \mathrm{p}(\mathbf{x}, \mathbf{z}, \boldsymbol{w} ; \boldsymbol{\theta}) \\
& =\sum_{i, k} z_{i k} \log \pi_{k}+\sum_{k, l} w_{j l} \log \rho_{l}+\sum_{i, j, k, l} z_{i k} w_{j l} \log \mathrm{p}\left(x_{i}^{j} ; \boldsymbol{\alpha}_{k l}\right)
\end{aligned}
$$

Be careful with asymptotics. .
If $\ln (d) / n \rightarrow 0, \ln (n) / d \rightarrow 0$ when $n \rightarrow \infty$ and $d \rightarrow \infty$, then the MLE is consistent [Brault et al., 2017]

## MLE estimation: EM algorithm

- E-step of EM (iteration q):

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}\right)= & E\left[\ell_{c}(\boldsymbol{\theta} ; \mathbf{x}, \mathbf{z}, \mathbf{w}) \mid \mathbf{x} ; \boldsymbol{\theta}^{(q)}\right] \\
= & \sum_{i, k} \underbrace{p\left(z_{i}=k \mid \mathbf{x} ; \boldsymbol{\theta}^{(q)}\right)}_{t_{i k}^{(q)}} \ln \pi_{k}+\sum_{j, l} \underbrace{p\left(w_{i}=I \mid \mathbf{x} ; \boldsymbol{\theta}^{(q)}\right)}_{s_{j l}^{(q)}} \ln \rho_{I} \\
& +\sum_{i, j, k, l} \underbrace{p\left(z_{i}=k, w_{j}=I \mid \mathbf{x} ; \boldsymbol{\theta}^{(q)}\right)}_{e_{i j k l}^{(q)}} \ln \mathrm{p}\left(x_{i j} ; \boldsymbol{\alpha}_{k l}\right)
\end{aligned}
$$

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

$$
\pi_{k}^{(q+1)}=\frac{\sum_{i} t_{i k}^{(q)}}{n}, \quad \rho_{l}^{(q+1)}=\frac{\sum_{j} s_{j l}^{(q)}}{d}, \quad \alpha_{k l}^{(q+1)}=\frac{\sum_{i, j} e_{i j k l}^{(q)} x_{i j}}{\sum_{i, j} e_{i j k l}^{(q)}}
$$

## MLE: intractable E step

$e_{i j k l}^{(q)}$ is usually intractable...

- Consequence of dependency between $\boldsymbol{x}_{i j} \mathrm{~s}$ (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

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

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

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

## 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 $(\boldsymbol{\theta}, \mathbf{e})$

$$
\tilde{\ell}_{c}(\boldsymbol{\theta} ; \boldsymbol{x}, \mathbf{e})=\sum_{i, k} t_{i k} \ln \pi_{k}+\sum_{j, l} s_{j l} \ln \rho_{l}+\sum_{i, j, k, l} e_{i j k l} \ln \mathrm{p}\left(x_{i j} ; \boldsymbol{\alpha}_{k l}\right)
$$

where $\mathbf{e}=\left(e_{i j k l}\right), e_{i j k l} \in\{0,1\}, \sum_{k, l} e_{i j k l}=1, t_{i k}=\sum_{j, l} e_{i j k l}, s_{j l}=\sum_{i, k} e_{i j k l}$
■ 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 $\mathbf{e}_{i j k l}=t_{i k} s_{j l}$
- New variables are thus now $\mathbf{t}=\left(t_{i k}\right)$ and $\mathbf{s}=\left(s_{j l}\right)$
- 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

$$
Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}\right) \approx \sum_{i, k} t_{i k}^{(q)} \ln \pi_{k}+\sum_{j, l} s_{j l}^{(q)} \ln \rho_{l}+\sum_{i, j, k, l} t_{i k}^{(q)} s_{j l}^{(q)} \ln \mathrm{p}\left(x_{i j} ; \boldsymbol{\alpha}_{k l}\right)
$$

- 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 $\left(\boldsymbol{z}^{(q)}, \boldsymbol{w}^{(q)}\right) \sim \mathrm{p}\left(\boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{x} ; \boldsymbol{\theta}^{(q)}\right)$ instead an expectation

■ But it is still intractable, thus use a Gibbs algorithm to approx. this draw

## Approximated S-step

Two easy draws

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

and

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

- 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 $\boldsymbol{\theta}$

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

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

$$
\hat{\boldsymbol{\theta}}=\arg \max _{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{x})
$$

## 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 $\rho$ (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


$\mathcal{B e}\left(\frac{1}{2}, \frac{1}{2}\right)$
$\mathcal{B e}(1,1)$
$\mathcal{B} e(4,4)$

## 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

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

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

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

## Block estimation: evaluation

- Empirical error rate between blocks:

$$
\operatorname{err}_{\text {blocks }}(\underbrace{(\boldsymbol{z}, \boldsymbol{w})}_{\text {"True" blocks }}, \underbrace{(\hat{\boldsymbol{z}}, \hat{\boldsymbol{w}})}_{\text {Estimated blocks }})=\operatorname{err}(\boldsymbol{z}, \hat{\boldsymbol{z}})+\operatorname{err}(\boldsymbol{w}, \hat{\boldsymbol{w}})-\operatorname{err}(\boldsymbol{z}, \hat{\boldsymbol{z}}) \times \operatorname{err}(\boldsymbol{w}, \hat{\boldsymbol{w}})
$$

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

## Block estimation: consistency

[Mariadassou and Matias, 12]

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

Thus we retrieve the HD clustering blessing. .

Block estimation: non asymptotic properties (1/2)
Binary case: marginals seems so simple mixtures! [Brault, 14]

Matrice initiale

Histogrammes
des sommes

Séparations



## Block estimation: non asymptotic properties (2/2)

## [Brault, 14]

- Probability of $x_{i j}$ with no regard to the column membership is Bernoulli

$$
\mathrm{p}\left(x_{i j}=1 \mid z_{i k}=1\right)=\tau_{k}=\sum_{l=1}^{L} \alpha_{k l} \rho_{l}
$$

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

$$
\left(\sum_{j} x_{i j}\right) \mid z_{i k}=1 \sim \mathrm{~B}\left(d, \tau_{k}\right)
$$

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

$$
\mathrm{p}\left(\hat{\mathbf{z}}^{m i x} \neq \mathbf{z}^{*}\right) \leq 2 n \exp \{-\frac{1}{8} d[\underbrace{\min _{k \neq k^{\prime}}\left|\tau_{k}-\tau_{k^{\prime}}\right|}_{\text {overlap }}]\}+K\left(1-\min _{k} \pi_{k}\right)^{n}
$$

- 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 $\times$ words
- Poisson model



## Illustration: document clustering (2/2)



Results with $2 \times 2$ blocks

|  | Medline | Cranfield |
| :---: | :---: | :---: |
| Medline | 1033 | 0 |
| Cranfield | 0 | 1398 |

Experiment illustrates previous theory: HD clustering is blessing

# Outline 

1 HD clustering

2 Modeling

3 Estimating

4 Selecting

5 Experiments

6 To go further

## 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

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

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

## ICL criterion: overcome both difficulties

- ICL uses complete likelihood thus no intractability

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

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

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

- We can make a conjecture for the general case

$$
\text { ICLbic }=\ell_{c}(\hat{\boldsymbol{\theta}} ; \mathbf{x}, \hat{\mathbf{z}}, \hat{\mathbf{w}})-\frac{K-1}{2} \ln (n)-\frac{L-1}{2} \ln (d)-\frac{K L \nu_{\boldsymbol{\alpha}_{k l}}}{2} \ln (n d)
$$

## ICL criterion: consistency

- We can obtain a BIC expression from ICLbic

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

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

$$
" \ell(\hat{\boldsymbol{\theta}} ; \mathbf{x})=\ell_{c}(\hat{\boldsymbol{\theta}} ; \mathbf{x}, \hat{\mathbf{z}}, \hat{\boldsymbol{w}})^{\prime \prime}
$$

- 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

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

[^2]
## Illustration: discuss the dimension (2/2)

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

- Perform clustering ${ }^{6}$ with $K=2$ : $\mathrm{ICL}=-89,433$, err $=0.1837$

Thus use preferably co-clustering in the HD setting

[^3]
# Outline 

1 HD clustering

2 Modeling

3 Estimating

4 Selecting

5 Experiments

6 To go further

## 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.natrix(read.table("blockcluater-example.csv", sap $={ }^{\circ}$, ") beed(data)
v1 v2 v3 v4 v5 ve v7 ve v9 v10 .... v91 v92 v93 v94 v95 v96 v97 v98 v99 v100 A matric $6 \times 100$
fore

| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | $\cdots$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | $\cdots$ | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | $\cdots$ | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | $\cdots$ | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | $\cdots$ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | $\cdots$ | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |

Clustering with Blockcluster
Launch the BlockCluster package.
In 141: Library(bleckeluster)
wosing required package: rekore
toading required package: Rcpp
Attaching package1 'rtkore.

The following object is masked from 'pockage: Rcpp' i
Ldr2aga
blackeluater version 4.4.3 loaded

Copyright (C) CMODAL team EIMRIA, Lille 6 O.M.R. C.A.R.S. 6599 Bendiasyc, uTc:
please post questions and bugs at: <https://qforge-inria. fr/forun/forun.phplforum_id-11190sgroup_id-3679:

Define the strategy.
Specity the number of groups/Clusters to create in rows and coloumns.

```
    ronclusters <-2
```

    roncluaters <- 2
    Run the coclustering.

## 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]:
reseICLvalue
$-45557.0741834614$
Cluster Plot
Draw the original data matrix, and the matrix obtained after performing co-clustering

In [8]: plot(res)


# Outline 

1 HD clustering

2 Modeling

3 Estimating

4 Selecting

5 Experiments

6 To go further

## 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


[^0]:    ${ }^{1}$ S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. Data Clustering: Algorithms and Applications, 29

[^1]:    ${ }^{3}$ The Poisson parameter is here split into $\mu_{k}$ and $\nu_{l}$ the effects of the row $k$ and the column $/$ respectively and $\gamma_{k l}$ the effect of the block $k l$. Unfortunately, this parameterization is not identifiable. It is therefore not possible to estimate simultaneously $\mu_{k}, \nu_{l}$ and $\gamma_{k l}$ without imposing further constraints. Constraints $\sum_{k} \pi_{k} \gamma_{k l}=\sum_{l} \rho_{l} \gamma_{k l}=1$ and $\sum_{k} \mu_{k}=1, \sum_{l} \nu_{l}=1$ are a possibility.

[^2]:    ${ }^{4}$ https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/
    ${ }^{5}$ There are 3 other continuous descriptors we do not use

[^3]:    ${ }^{6}$ Equivalent to co-clustering with $L=54$

