Model-based clustering and co-clustering in high-dimensional scenarios

C. Biernacki

Research Summer School on Statistics for Data Science – S4D

June 15th-22th 2018, Caen (France)

Preamble

What is this course?

- Understand key locks in clustering due to large data scenarios
- Describe some clustering methods to overcome such locks

What is not this course?

- Not an exhaustive list of clustering methods (and related bibliography)
- Do not make specialists of clustering methods

This preamble is valid for both lessons:

1. Model-based clustering and co-clustering in high-dimensional scenarios
2. Model selection theory and considerations in large scale scenarios

Lectures

- General overview of data mining (contain some pretreatments before clustering):
  Gérard Govaert et al. (2009). Data Analysis. Wiley-ISTE, ISBN: 978-1-848-21098-1.
  `https://www.wiley.com/en-fr/Data+Analysis-p-9781848210981`
- More advanced material on clustering:
  - Christian Hennig, Marina Meila, Fionn Murtagh, Roberto Rocci (2015). Handbook of Cluster Analysis. Chapman and Hall/CRC, ISBN 9781466551886, Series: Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

    `https://www.crcpress.com/Handbook-of-Cluster-Analysis/Hennig-Meila-Murtagh-Rocci/p/book/9781466551886`
  - Christophe Biernacki. Mixture models. J-J. Droesbeke; G. Saporta; C. Thomas-Agnan. Choix de modèles et agrégation, Technip, 2017.

    `https://hal.inria.fr/hal-01252671/document`
  - Christophe Biernacki, Cathy Maugis. High-dimensional clustering. J-J. Droesbeke; G. Saporta; C. Thomas-Agnan. Choix de modèles et agrégation, Technip, 2017.

    `https://hal.archives-ouvertes.fr/hal-01252673v2/document`
- Advanced material on co-clustering:
  Gérard Govaert, Mohamed Nadif (2013). Co-Clustering: Models, Algorithms and Applications. Wiley-ISTE, ISBN-13: 978-1848214736.
  `https://www.wiley.com/en-fr/Co+Clustering:+Models,+Algorithms+and+Applications-p-9781848214736`
- Basic to more advanced R book: Pierre-Andre Cornillon, Arnaud Guyader, Francois Husson, Nicolas Jegou, Julie Josse, Maela Kloareg, Eric Matzner-Lober, Laurent Rouvière (2012). R for Statistics. Chapman and Hall/CRC, ISBN 9781439881453.
  `https://www.crcpress.com/R-for-Statistics/Cornillon-Guyader-Husson-Jegou-Josse-Kloareg-Matzner-Lober-Rouviere/p/book/9781439881453`

Keep-home message

High dimensional clustering is simple . . .
. . . in case of (essentially) relevant clustering variables

Outline

Everything begins from data!

Genesis of "Big Data"

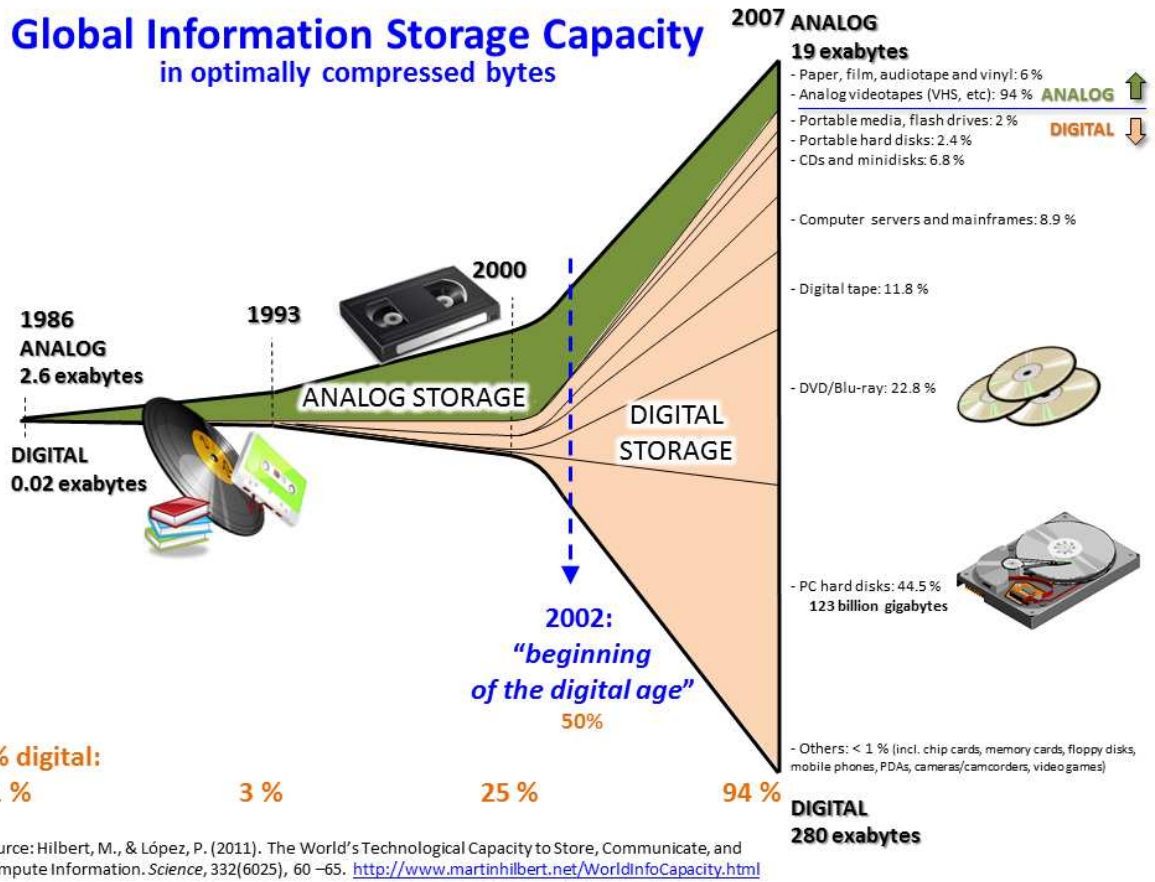
The Big Data phenomenon mainly originates in the increase of computer and digital resources at an ever lower cost

- Storage cost per MB: 700$ in 1981, 1$ in 1994, 0.01$ in 2013
  $\rightarrow$ price divided by 70,000 in thirty years
- Storage capacity of HDDs: $\approx$1.02 Go in 1982, $\approx$8 To today
  $\rightarrow$ capacity multiplied by 8,000 over the same period
- Computeur processing speed: 1 gigaFLOPS[1] in 1985, 33 petaFLOPS in 2013
  $\rightarrow$ speed multiplied by 33 million

[1] FLOP = FLoating-point Operations Per Second

Digital flow

- Digital in 1986: 1% of the stored information, 0.02 Eo[2]
- Digital in 2007: 94% of the stored information, 280 Eo (multiplied by 14,000)



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

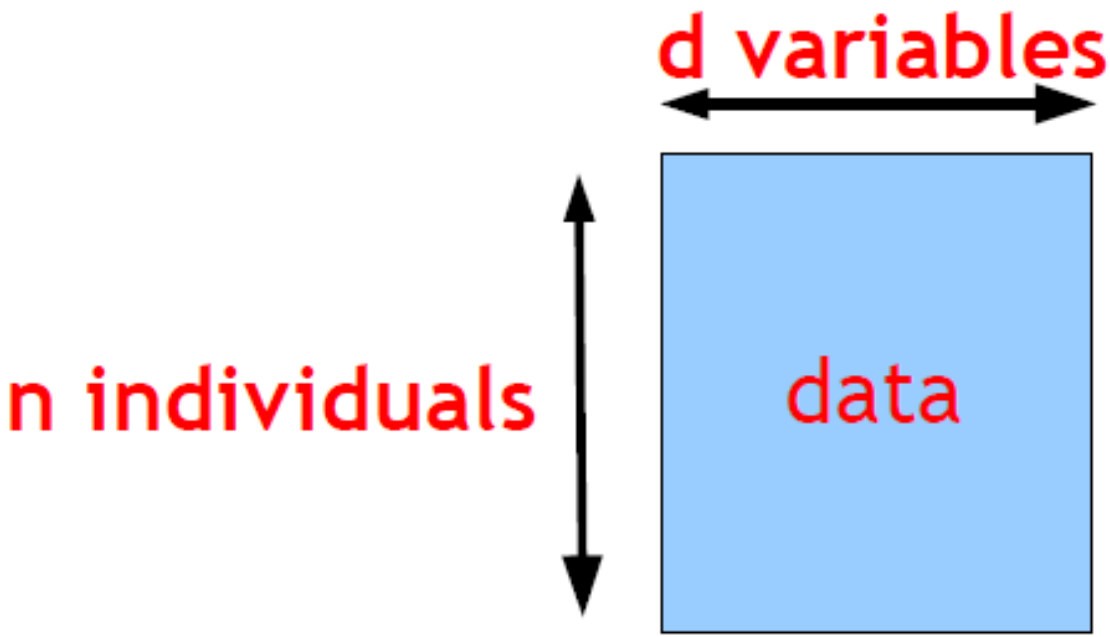
[2] Exabyte

Societal phenomenon

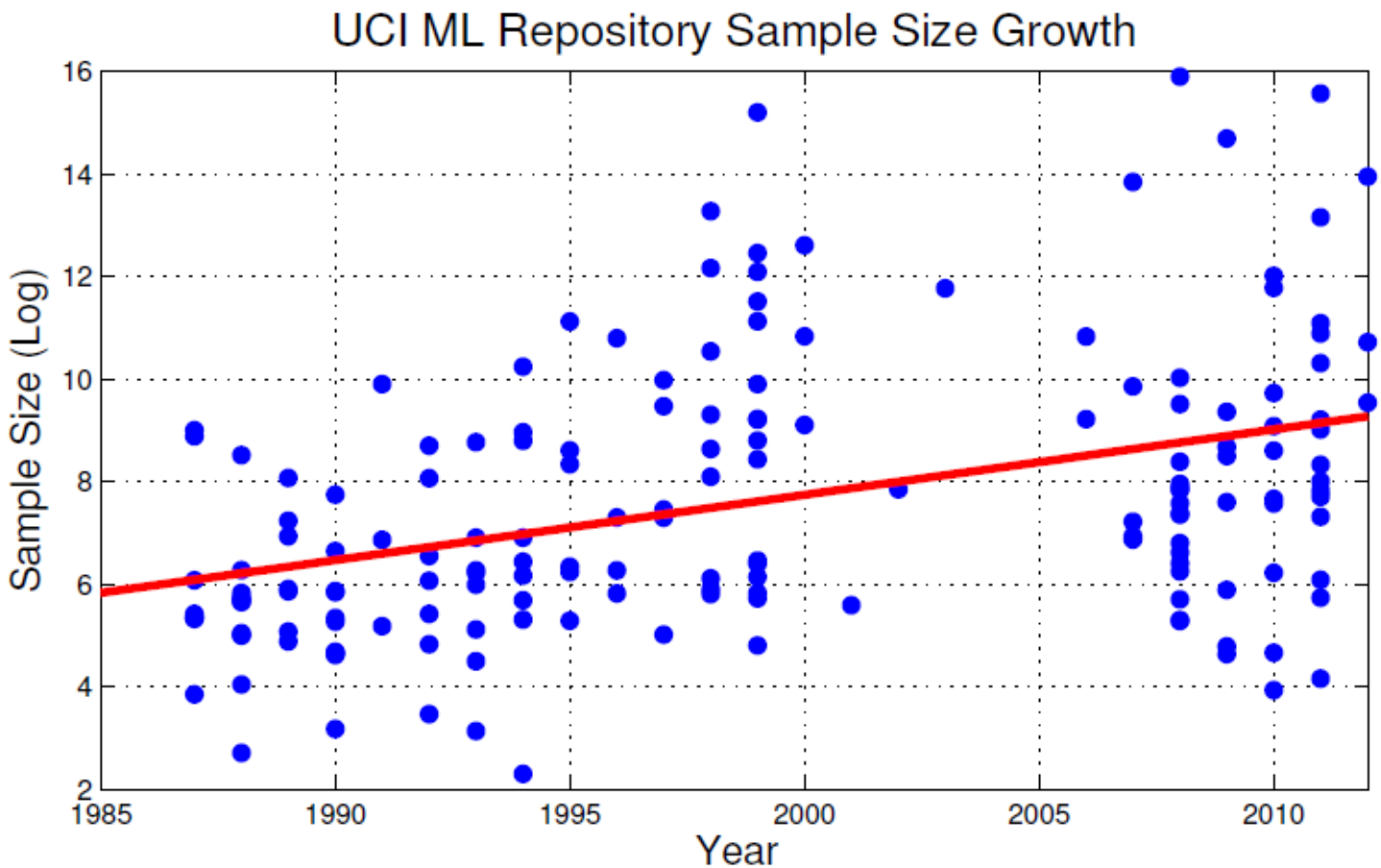
All human activities are impacted by data accumulation

- Trade and business: corporate reporting system , banks, commercial transactions, reservation systems. . .
- Governments and organizations: laws, regulations, standardizations , infrastructure. . .
- Entertainment: music, video, games, social networks. . .
- Sciences: astronomy, physics and energy, genome,. . .
- Health: medical record databases in the social security system. . .
- Environment: climate, sustainable development , pollution, power. . .
- Humanities and Social Sciences: digitization of knowledge , literature, history , art, architecture, archaeological data. . .

Data sets structure

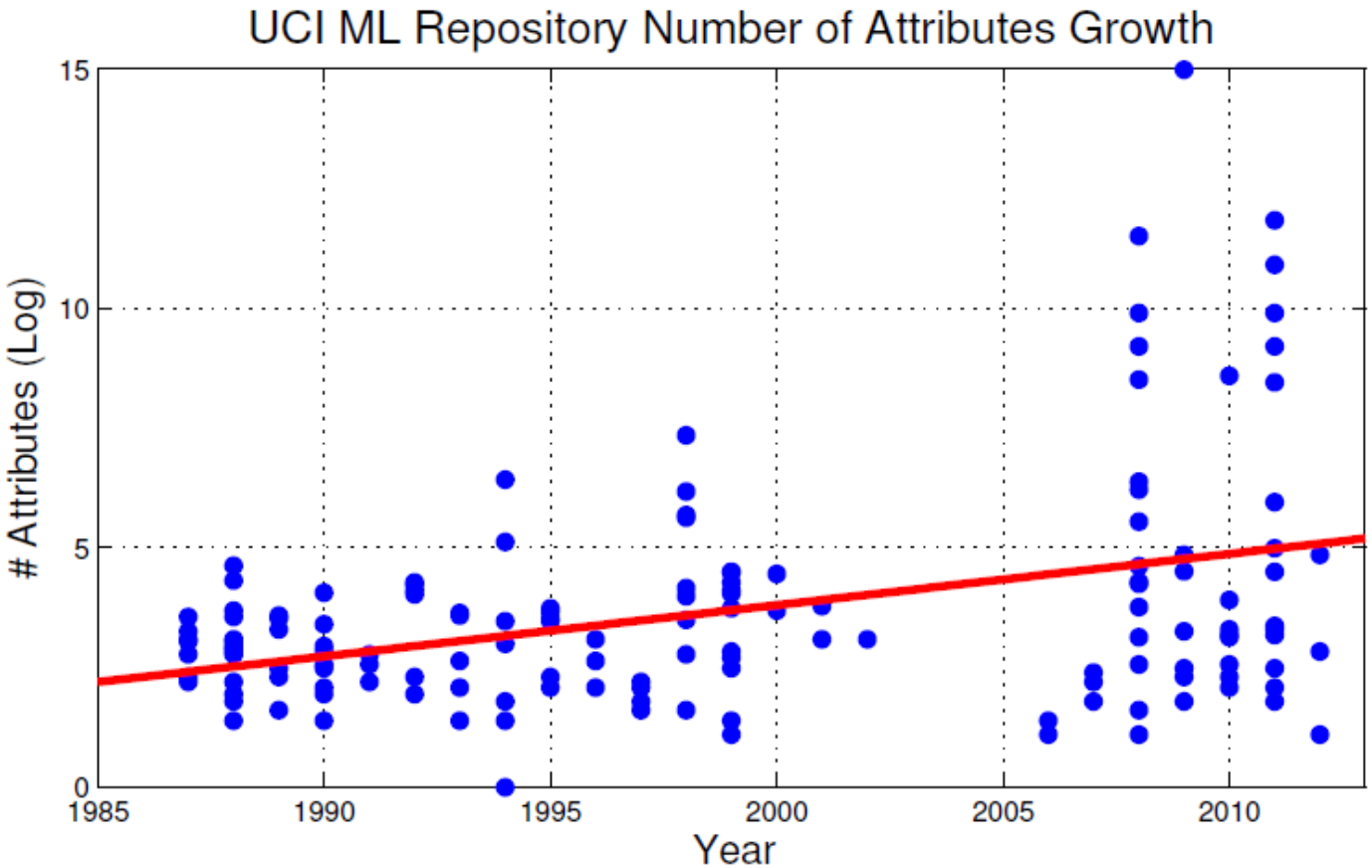


Large data sets $(n)^3$



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

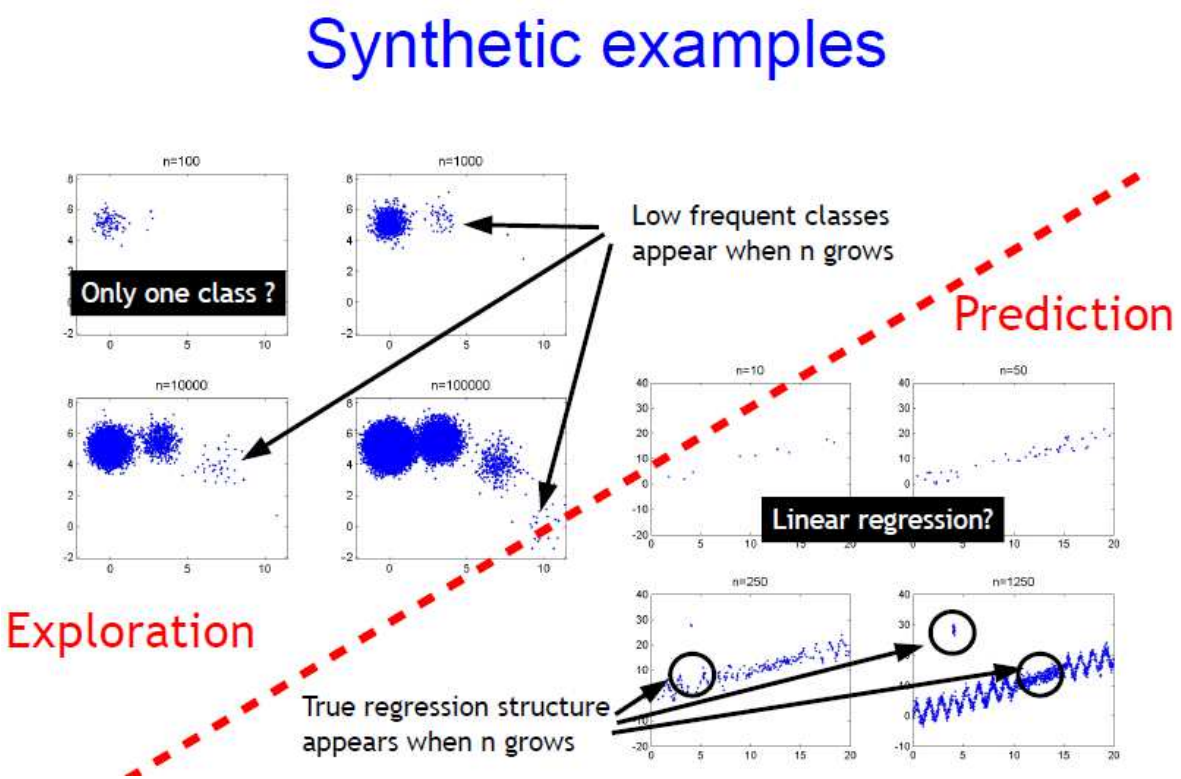
High-dimensional/HD data (d)[4]



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

More data for what?

Opportunity to improve accuracy of traditional questionings



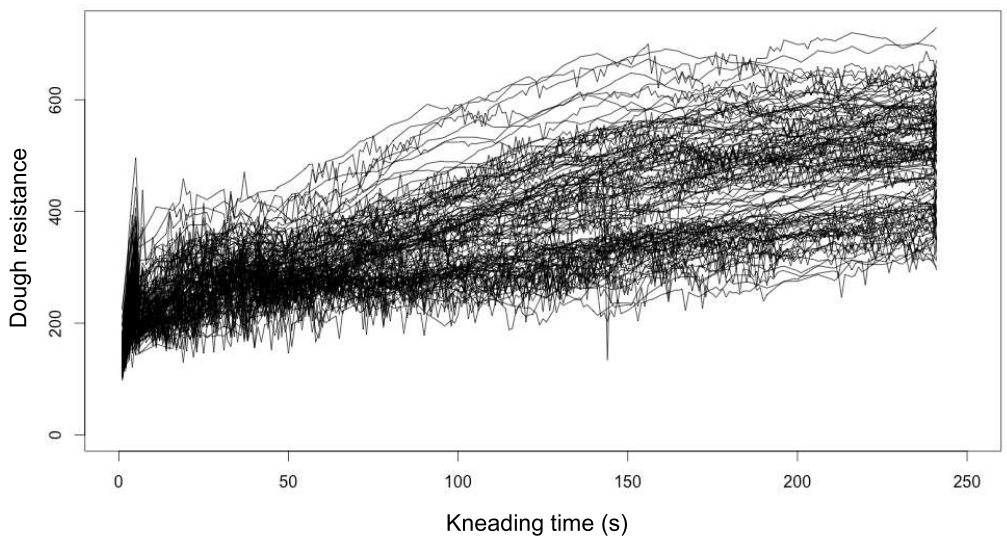
- Here is just illustrated the effect of n
- In a later section will be illustrated the effect of d (be patient)

HD data: domain dependency definition

- 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
- ...

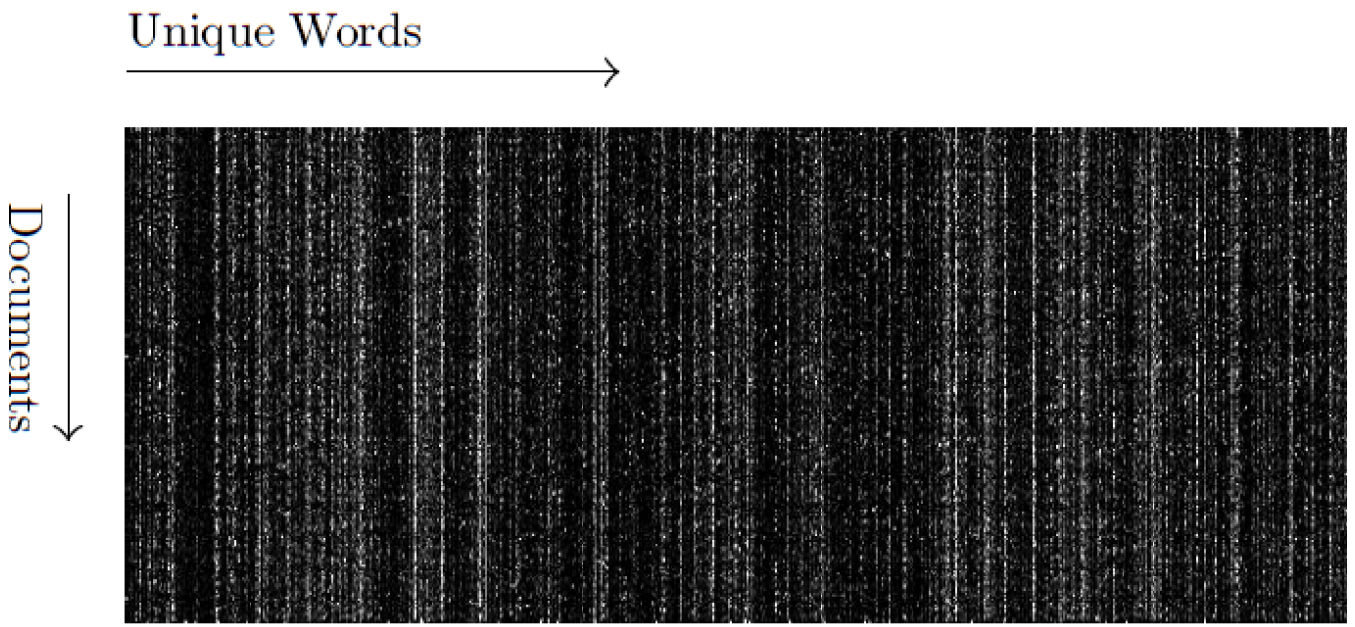
HD data: Curve "cookies" example

The Kneading dataset comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process [Lévéder *et al*, 04]. There are 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. One obtains 115 kneading curves observed at 241 equispaced instants of time in the interval [0; 480]. The 115 flours produce cookies of different quality: 50 of them have produced cookies of good quality, 25 produced medium quality and 40 low quality.



HD data: Medline example

$n = 2431$ documents described by the frequency of $d = 9275$ unique words



HD data: towards a theoretical definition (1/2)

An attempt in the non-parametric case

Dataset $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i$ 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: towards a theoretical definition (2/2)

An attempt in the parametric case

Dataset $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i$ 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 data: consequences on features

Since it is now easy to collect many features, it favors also

- data variety and/or mixed
- data missing
- data uncertainty (or interval data)

Mixed, missing, uncertain

?	0.5	?	5
0.3	0.1	green	3
0.3	0.6	{red,green}	3
0.9	[0.25 0.45]	red	?
↓	↓	↓	↓
continuous	continuous	categorical	integer

HD data: full mixed/missing

categorical
Marital status
maried

integer
Children
3

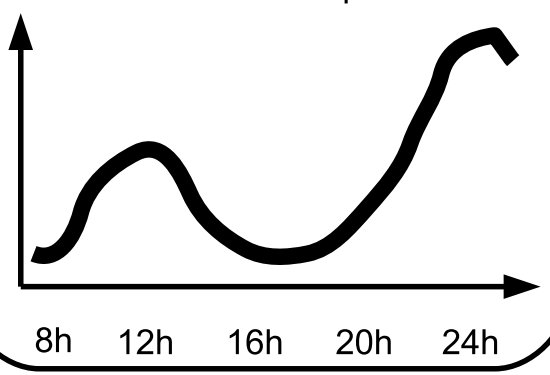
missing
Size (m)
?

rank
Drink preference
beer > soda > water

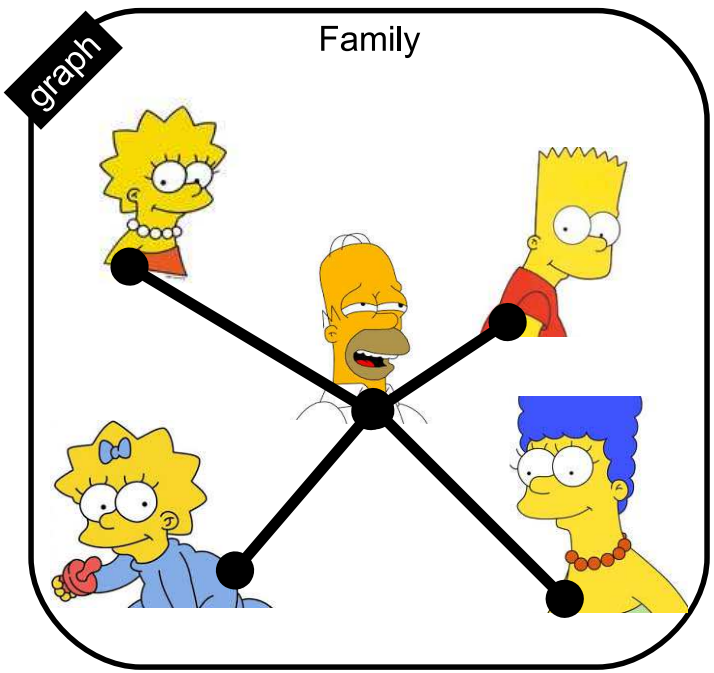
ordinal
Intelligence
low

continuous
Weight (kg)
119.5

functional



graph



And so on...

Coding for data $\boldsymbol{x}$

- A set of n individuals

$$\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$$

with $\boldsymbol{x}_i$ a set of (possibly non-scalar) d variables

$$\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{id}\}$$

where $\boldsymbol{x}_{ij} \in \mathcal{X}_j$

- A n-uplet of individuals

$$\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

with $\boldsymbol{x}_i$ a d-uplet of (possibly non-scalar) variables

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{id}) \in \mathcal{X}$$

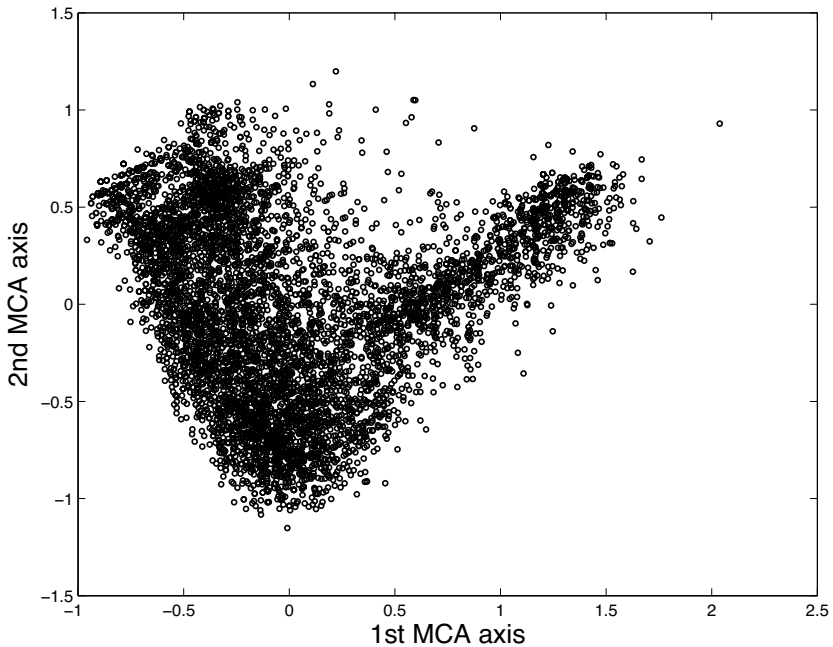
where $\mathcal{X} = \mathcal{X}_1 \times \ldots\ldots \mathcal{X}_d$

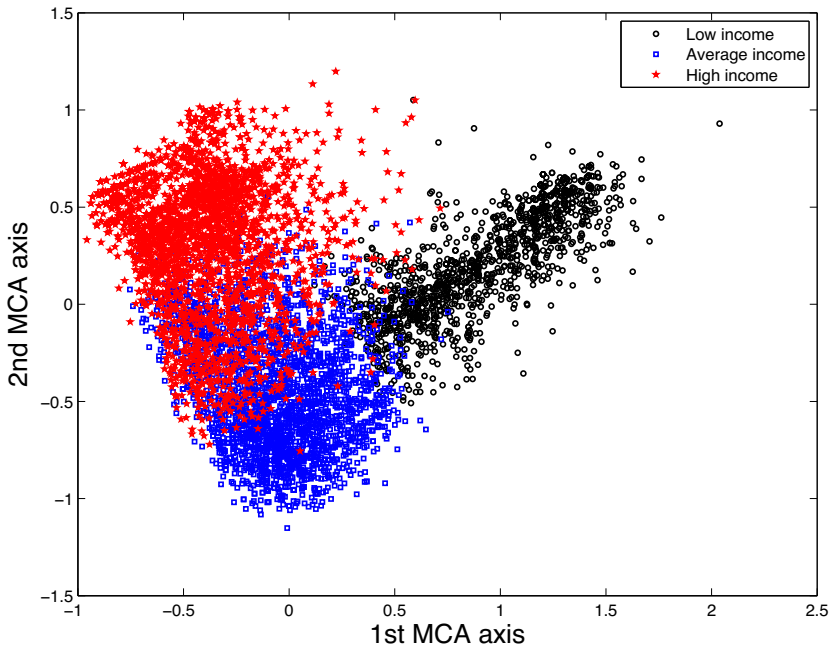
We will pass from a coding to another, depending of the practical utility (useful for some calculus to have matrices or vectors for instance)

Outline

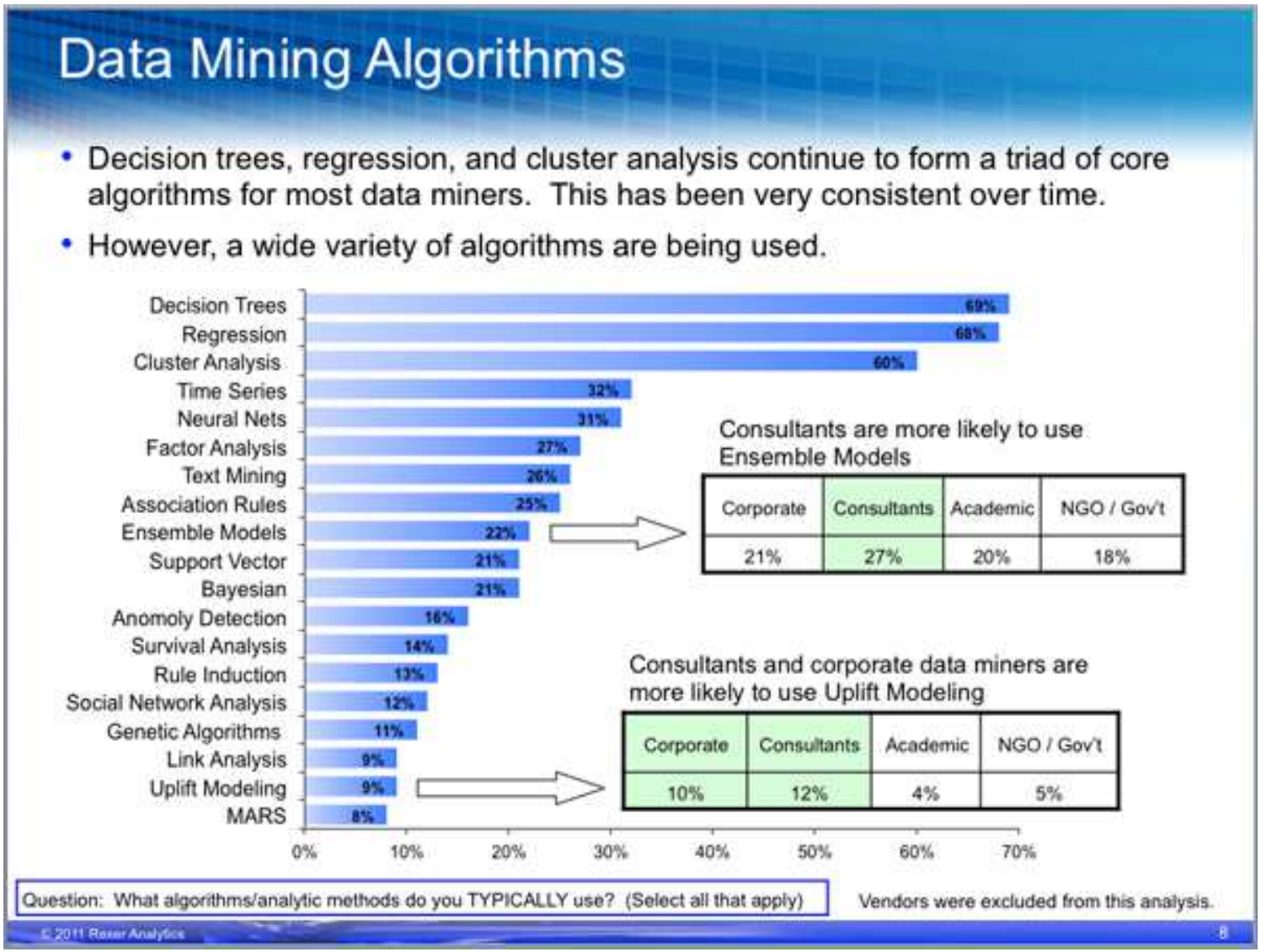
Clustering?

Detect hidden structures in data sets





Clustering everywhere[5]



[5] Rexer Analytics's Annual Data Miner Survey is the largest survey of data mining, data science, and analytics professionals in the industry (survey of 2011)

Notations

- Data: n individuals: $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{x}^O, \mathbf{x}^M\}$ in a space $\mathcal{X}$ of dimension d
  - Observed individuals $\boldsymbol{x}^O$
  - Missing individuals $\boldsymbol{x}^M$
- Aim: estimation of the partition $\mathbf{z}$ and the number of clusters K
  Partition in K clusters $G_1, \ldots, G_K$: $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$

$$\mathbf{x}_i \in G_k \quad \Leftrightarrow \quad z_{ih} = \mathbb{I}_{\{h=k\}}$$

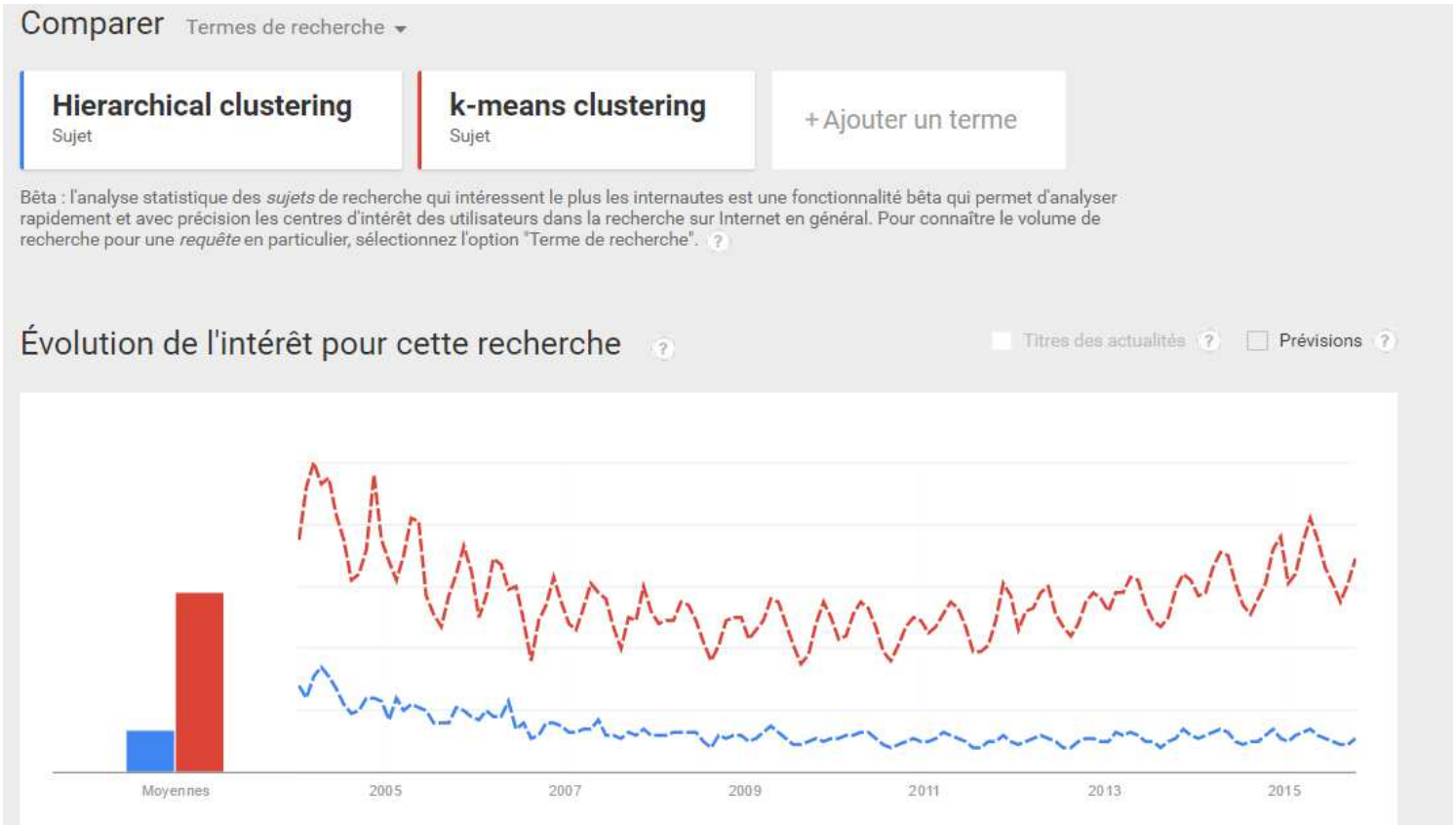
- Complex: mixed – missing – uncertain – n large – d large

Mixed, missing, uncertain

Individuals $\mathbf{x}$				Partition $\mathbf{z}$			$\Leftrightarrow$	Group
?	0.5	red	5	?	?	?	$\Leftrightarrow$	???
0.3	0.1	green	3	?	?	?	$\Leftrightarrow$	???
0.3	0.6	{red,green}	3	?	?	?	$\Leftrightarrow$	???
0.9	[0.25 0.45]	red	?	?	?	?	$\Leftrightarrow$	???
↓	↓	↓	↓					
continuous	continuous	categorical	integer					

Popularity of K-means and hierarchical clustering

Even K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering for several reasons: ease of implementation, simplicity, efficiency, empirical success. . .



K-means: within-cluster inertia criterion

Select the partition $\mathbf{z}$ minimizing the criterion

$$W_{\mathbf{M}}(\mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|_{\mathbf{M}}^2$$

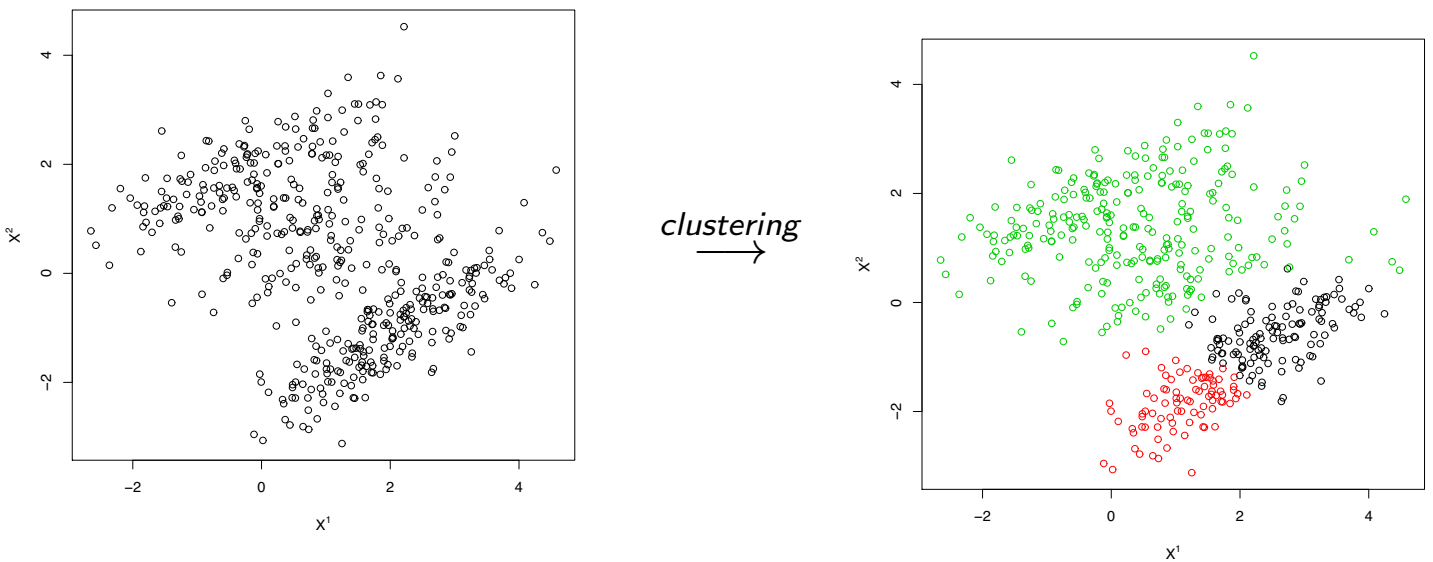
- Look for compact clusters (indiv. of the same cluster are close from each other)
- $\|\cdot\|_{\mathbf{M}}$ is the Euclidian distance with metric $\mathbf{M}$ in $\mathbb{R}^d$
- $\bar{\mathbf{x}}_k$ is the mean (or center) of the kth cluster

$$\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} \mathbf{x}_i$$

and $n_k = \sum_{k=1}^{n} z_{ik}$ indicates the number of individuals in cluster k

K-means: limitations

$\hat{\mathbf{z}} = \{\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_n\}$, $\hat{K}$ clusters



Clustering is an ill-posed problem

What is the precise definition of a cluster?

Reformulate K-means: the hidden Gaussian assumption

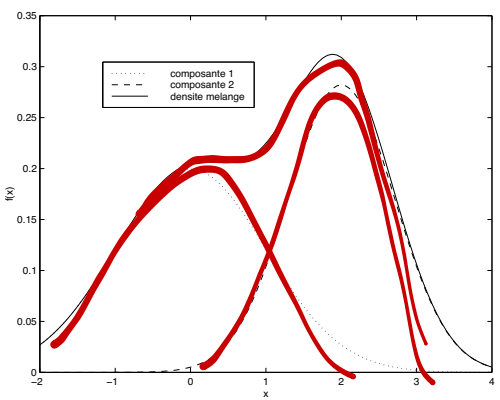
$$
\begin{aligned}
W_{\mathbf{I}}(\mathbf{z}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_{\mathbf{I}}^2 \\
&= -2\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln\left[\underbrace{\frac{1}{K}}_{!} \underbrace{\frac{1}{(2\pi)^{d/2}|\mathbf{I}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\mathbf{I}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right)}_{\mathrm{N}_d(\boldsymbol{\mu}_k,\mathbf{I})}\right] + \text{cst}
\end{aligned}
$$

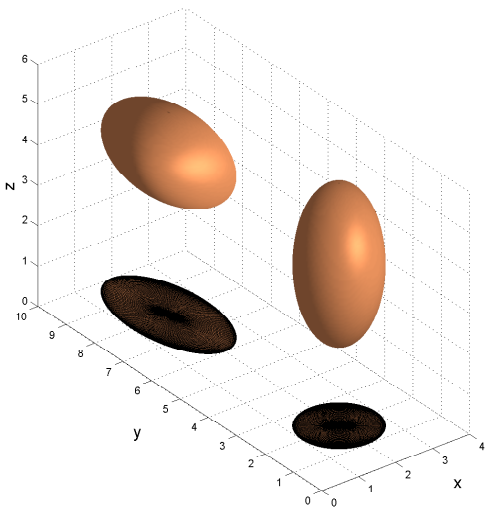
Model

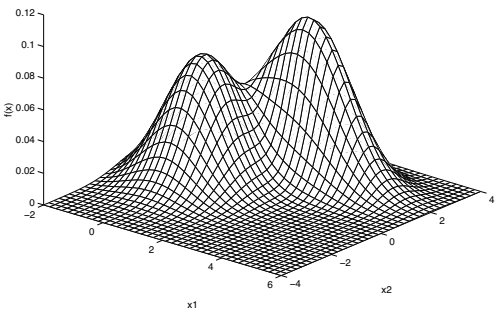
d-variate Gaussian with variance matrix $\mathbf{I}$ and same cluster sample size (see later)

Gaussian mixture model

$$p(\cdot;\alpha_k) = N_d(\mu_k, \Sigma_k) \quad \text{where} \quad \alpha_k = (\underbrace{\mu_k}_{\text{center}}, \underbrace{\Sigma_k}_{\text{dispersion}})$$







Parametric mixture model

- Parametric assumption:

$$p_k(\mathbf{x}_1) = p(\mathbf{x}_1; \boldsymbol{\alpha}_k)$$

thus

$$p(\mathbf{x}_1) = p(\mathbf{x}_1; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_1; \boldsymbol{\alpha}_k)$$

- Mixture parameter:

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha}) \text{ with } \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$$

- Model: it includes both the family $p(\cdot; \boldsymbol{\alpha}_k)$ and the number of groups K

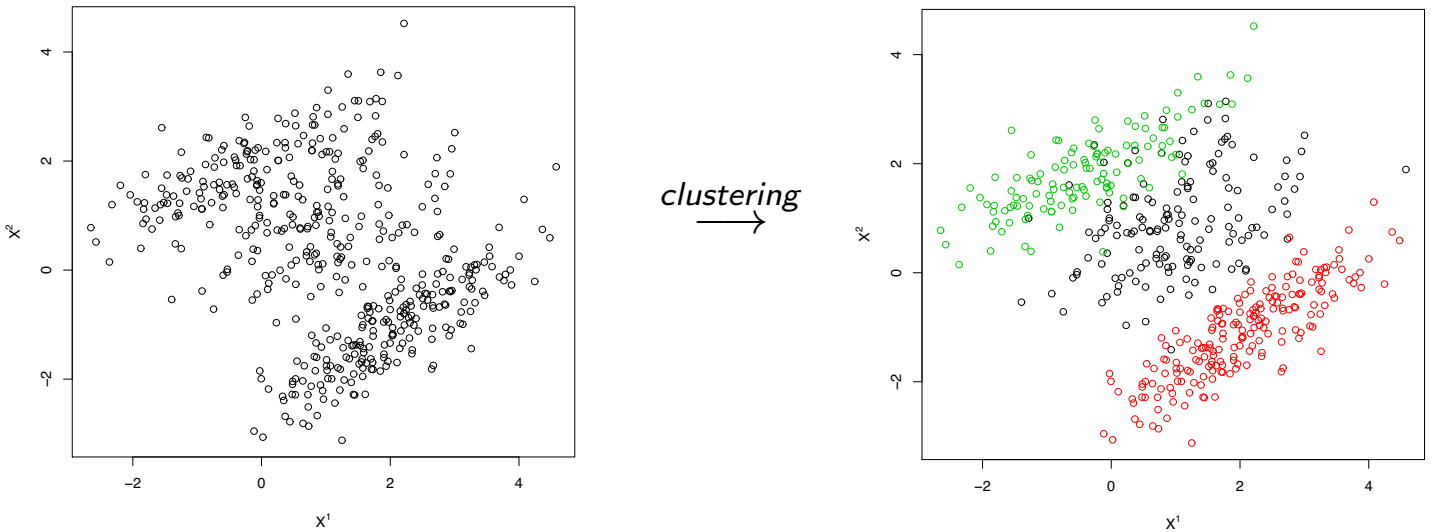
$$\mathbf{m} = \{p(\mathbf{x}_1; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

The number of free *continuous* parameters is given by

$$\nu = \dim(\Theta)$$

Mixture models: a probabilistic view of K-means

$\mathbf{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ $\quad\quad$ $\hat{\mathbf{z}} = \{\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_n\}$, $\hat{K}$ clusters



Clustering becomes a well-posed problem

$$p(\mathbf{x}|K;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|K;\boldsymbol{\alpha}_k) \quad \text{can be used for} \quad \begin{cases} \mathbf{x} \to \hat{\boldsymbol{\theta}} \to \boxed{p(\mathbf{z}|\mathbf{x}, K; \hat{\boldsymbol{\theta}})} \to \hat{\mathbf{z}} \\ \mathbf{x} \to \boxed{\hat{p}(K|\mathbf{x})} \to \hat{K} \\ \ldots \end{cases}$$

with $\boldsymbol{\theta} = (\pi_k, (\boldsymbol{\alpha}_k))$

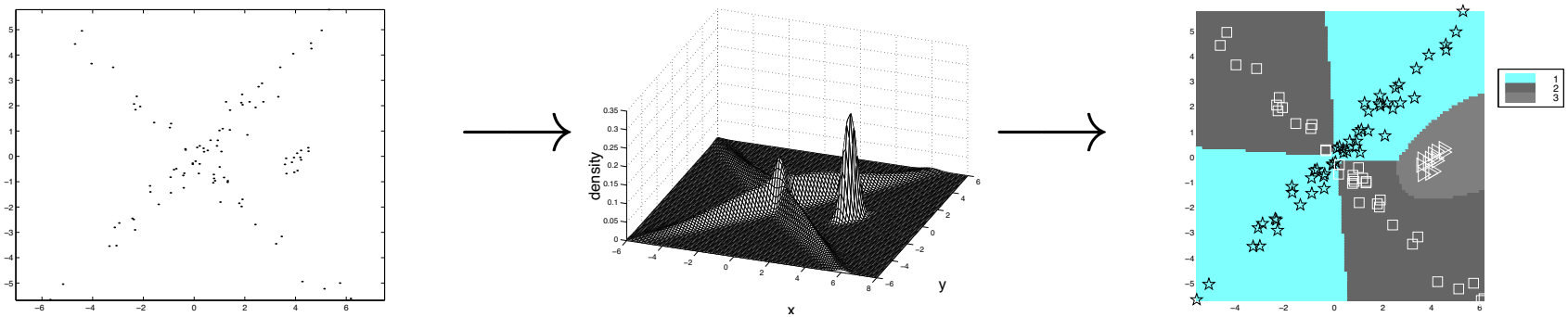
The clustering process in mixtures

1. Estimation of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$
2. Estimation of the conditional probability that $\mathbf{x}_i \in G_k$

$$t_{ik}(\hat{\boldsymbol{\theta}}) = p(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i; \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_k p(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)}{p(\mathbf{x}_i; \hat{\boldsymbol{\theta}})}$$

3. Estimation of $\mathbf{z}_i$ by *maximum a posteriori* (MAP)

$$\hat{z}_{ik} = \mathbb{I}_{\{k = \arg\max_{h=1,\ldots,K} t_{ih}(\hat{\boldsymbol{\theta}})\}}$$



Estimation of $\boldsymbol{\theta}$ by *observe* likelihood

Maximize the observe log-likelihood on $\boldsymbol{\theta}$

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i; \boldsymbol{\theta})$$

- Convergence of $\hat{\boldsymbol{\theta}}$, asymptotic efficiency, asymptotically unbiased
- General algorithm for missing data: EM
- Interpretation: it is a kind of fuzzy clustering

Principle of EM

- Initialization: $\boldsymbol{\theta}^0$
- Iteration n^o q:
  - Step E: estimate probabilities $\mathbf{t}^q = \{t_{ik}(\boldsymbol{\theta}^q)\}$
  - Step M: maximize $\boldsymbol{\theta}^{q+1} = \arg\max_{\boldsymbol{\theta}} \ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}^q)$[6]
- Stopping rule: iteration number or criterion stability

Properties

- $\oplus$: simplicity, monotony, low memory requirement
- $\ominus$: local maxima (depends on $\boldsymbol{\theta}^0$), linear convergence

[6] It is the so-called *complete* log-likelihood: $\ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \{\pi_k p(\mathbf{x}_i; \boldsymbol{\alpha}_k)\}$

Gaussian M-step

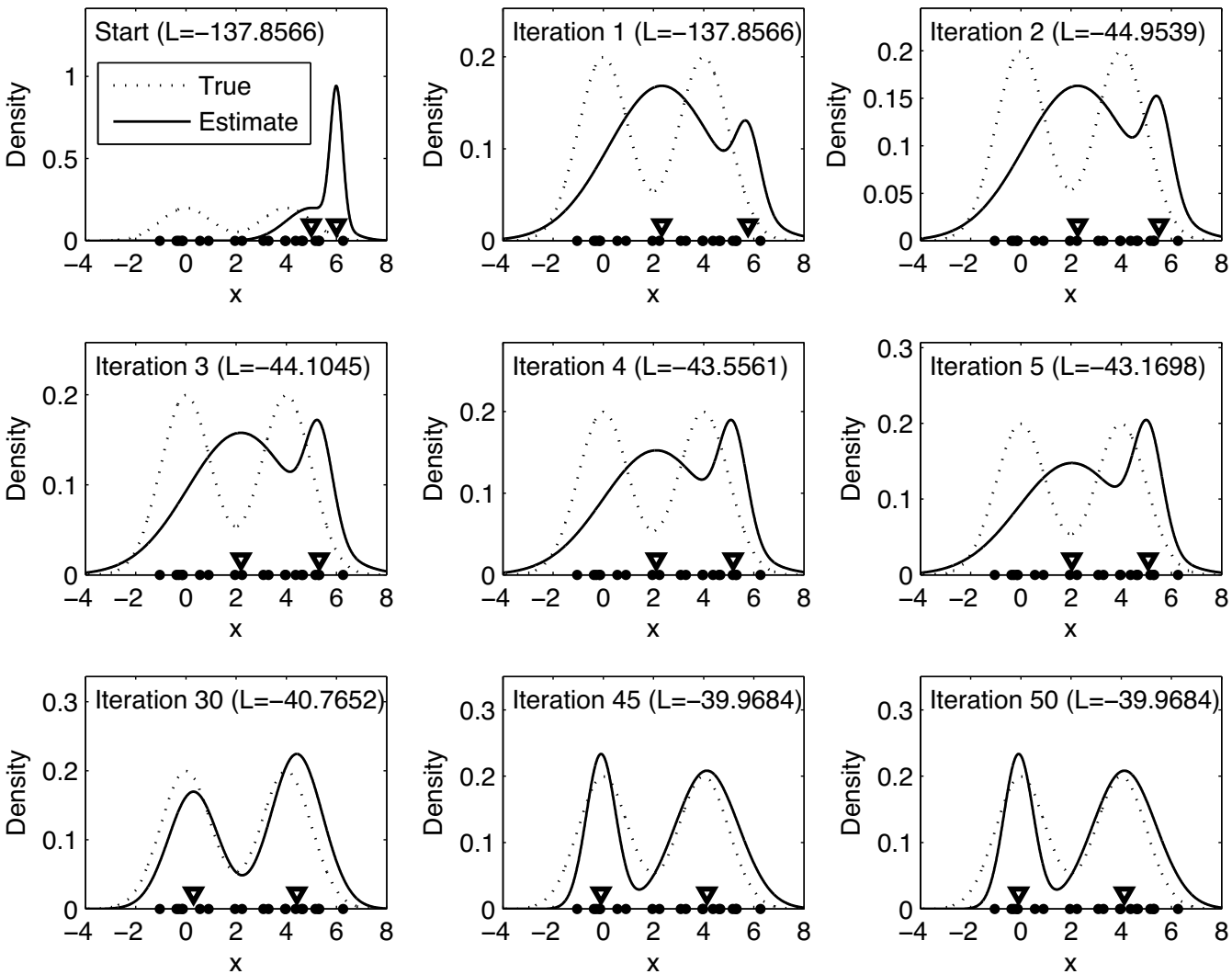
$$n_k^{(q)} = \sum_{i=n^l+1}^{n} t_{ik}(\boldsymbol{\theta}^{(q)})$$

$$\pi_k^{(q+1)} = \frac{n_k^{(q)}}{n}$$

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{n_k^{(q)}} \left(\sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{(q)}) \mathbf{x}_i \right)$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{n_k^{(q)}} \left(\sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{(q)}) (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})' \right)$$

Example of EM in the univariate case



Note : low at the beginning but increase of the log-likelihood

Categorical variables: latent class model

- Categorical variables: d variables with m_j modalities each, $\mathbf{x}_i^j \in \{0,1\}^{m_j}$ and

$$\mathbf{x}_i^{jh} = 1 \quad \Leftrightarrow \quad \text{variable } j \text{ of } \mathbf{x}_i \text{ takes level } h$$

- Conditional independence:

$$\mathrm{p}(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}$$

and

$$\alpha_k^{jh} = \mathrm{p}(\boldsymbol{x}_i^{jh} = 1 | z_{ik} = 1)$$

with $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$

Integer: Poisson mixture model

- integer variables: d variables $\mathbf{x}_i^j \in \mathbb{N}$
- Intra conditional independence:

$$p(\mathbf{x}_i^{int}; \boldsymbol{\alpha}_k^{int}) = \prod_{j=1}^{d} \frac{(\alpha_k^j)^{x_i^j}}{\alpha_k^j!} e^{-\alpha_k^j}$$

SPAM E-mail Database[8]

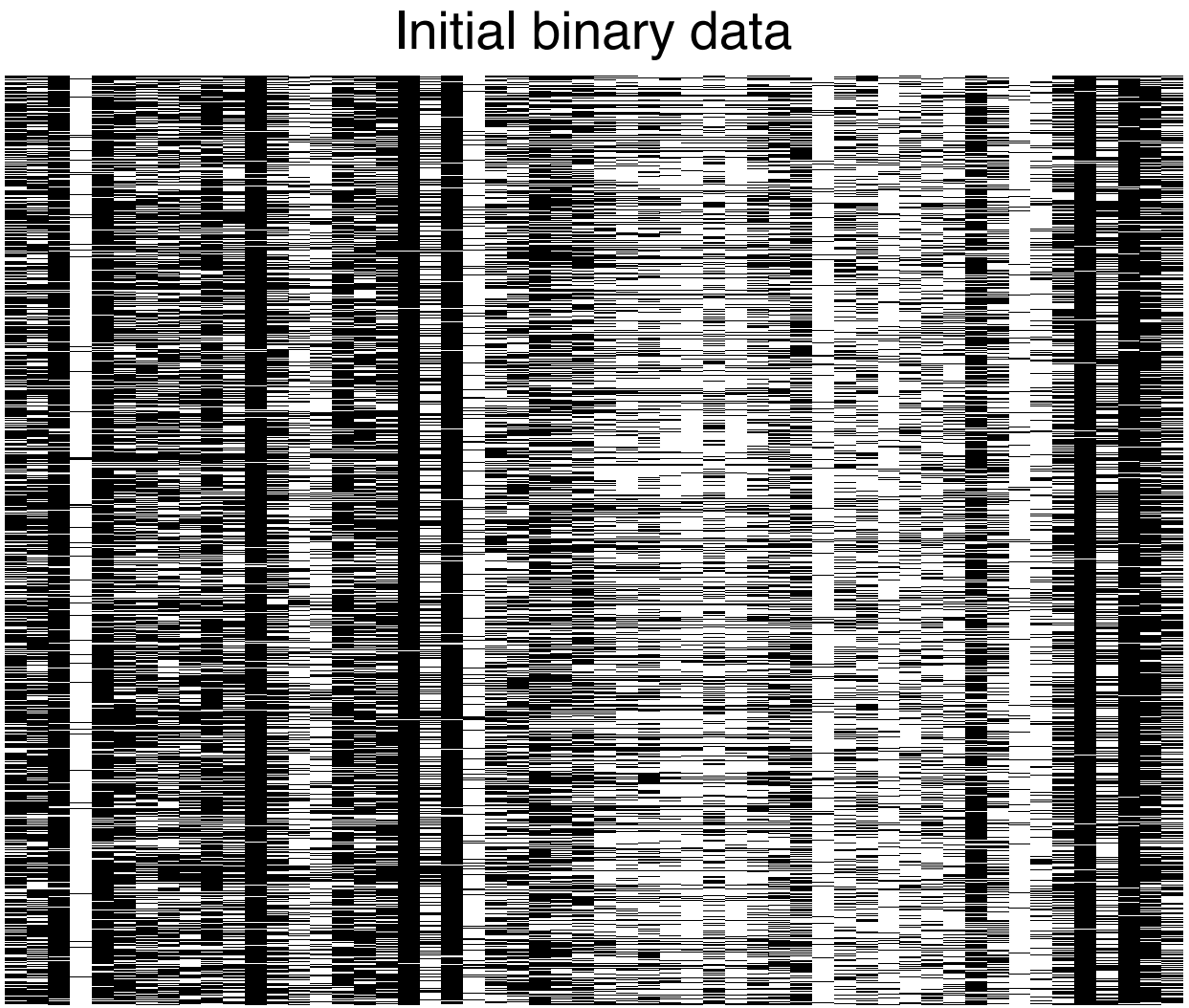
- $n = 4601$ e-mails composed by 1813 "spams" and 2788 "good e-mails"
- $d = 48 + 6 = 54$ continuous descriptors[7]
  - 48 percentages that a given word appears in an e-mail ("make", "you"...)
  - 6 percentages that a given char appears in an e-mail (";", "$"...)
- Transformation of continuous descriptors into binary descriptors

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

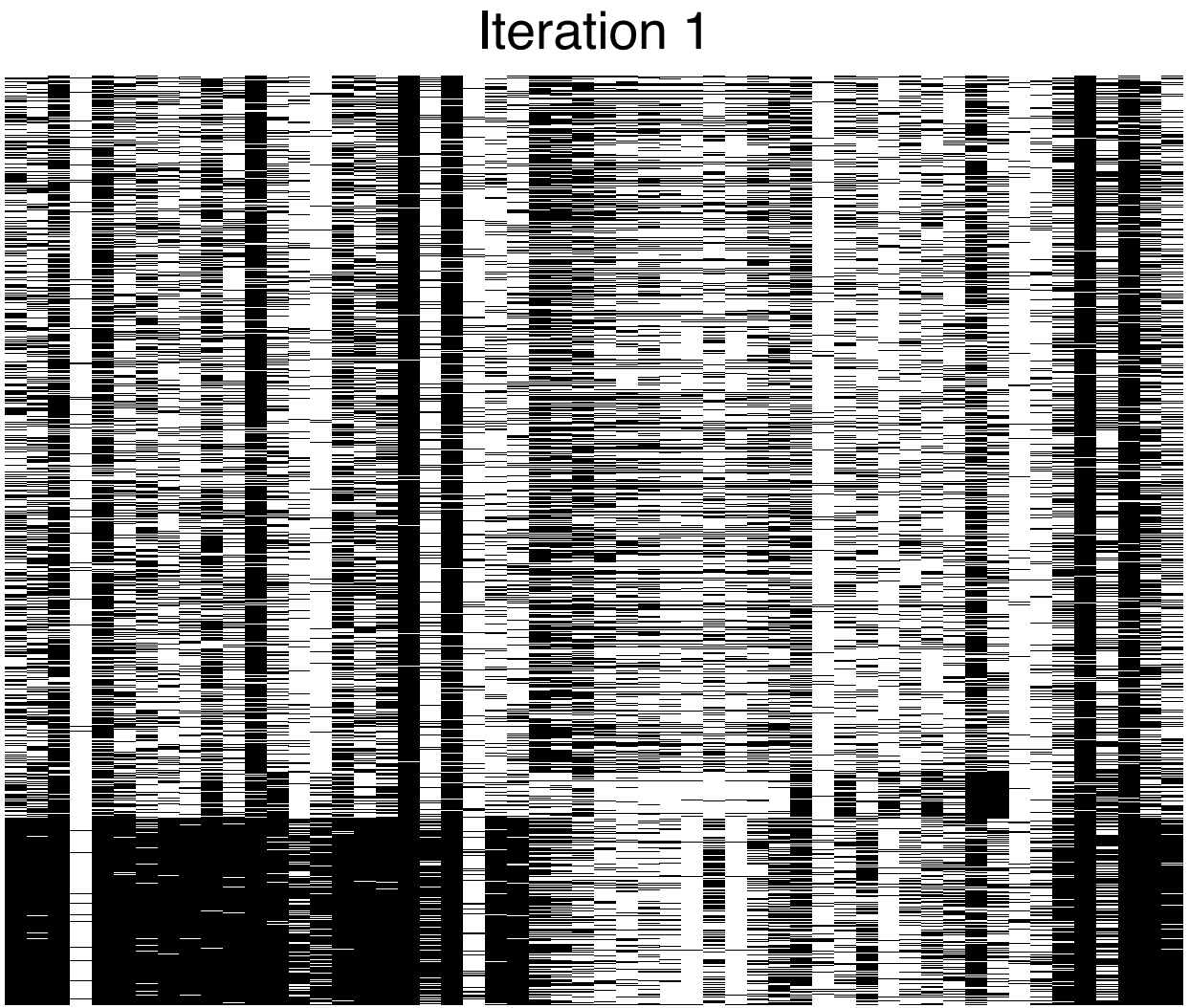
[7] There are 3 other continuous descriptors we do not use

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

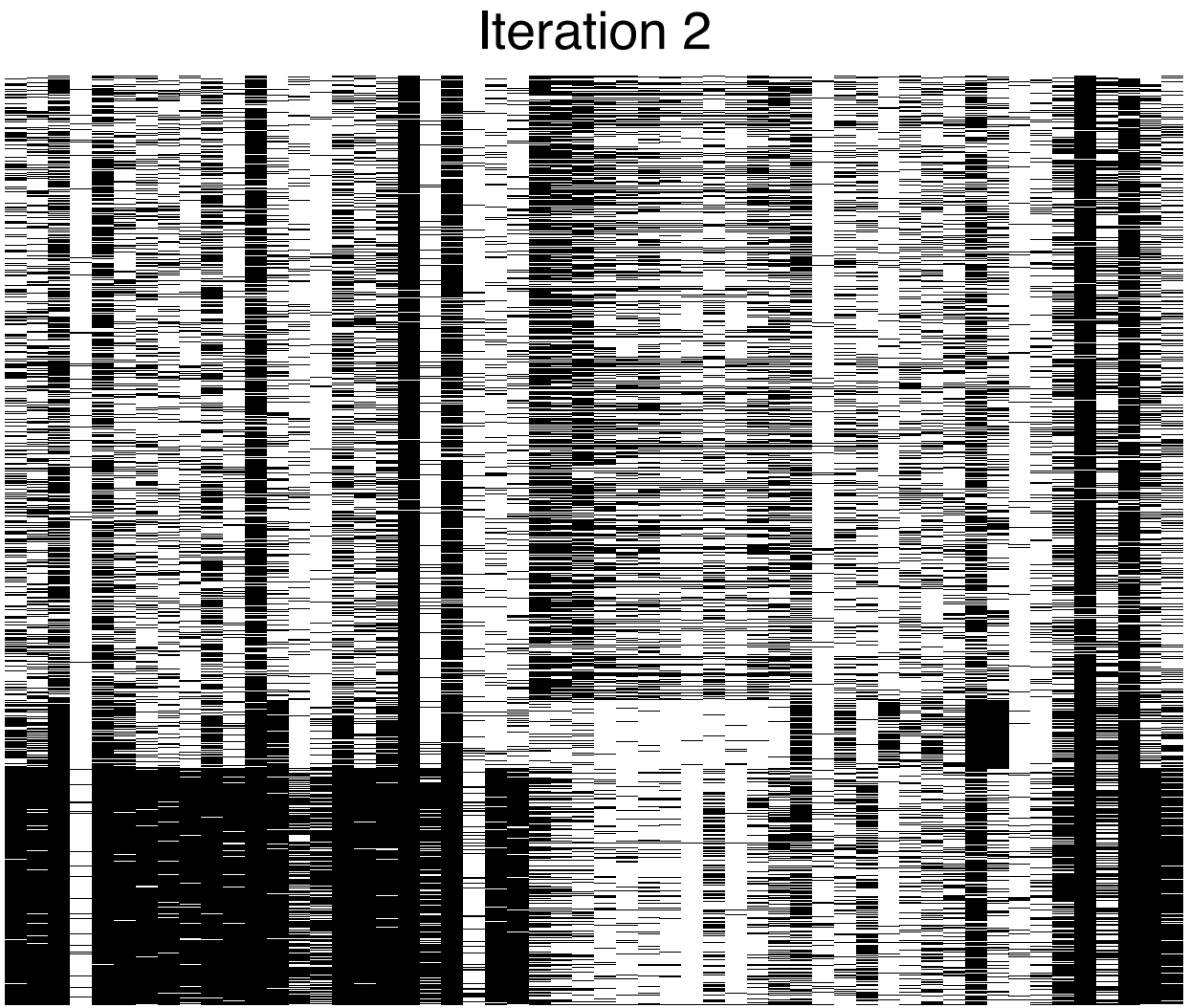
An EM run with a binary data set



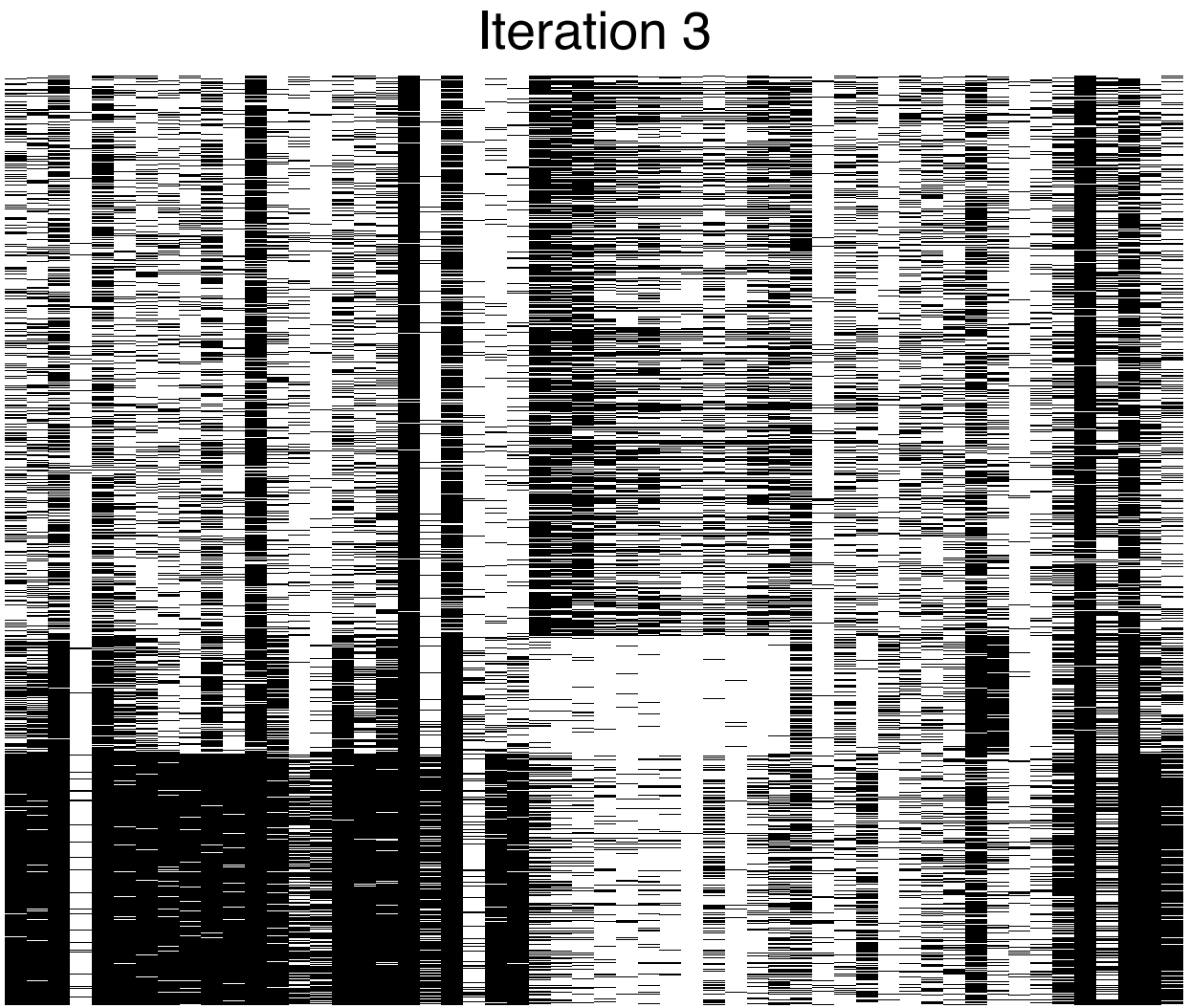
An EM run with a binary data set



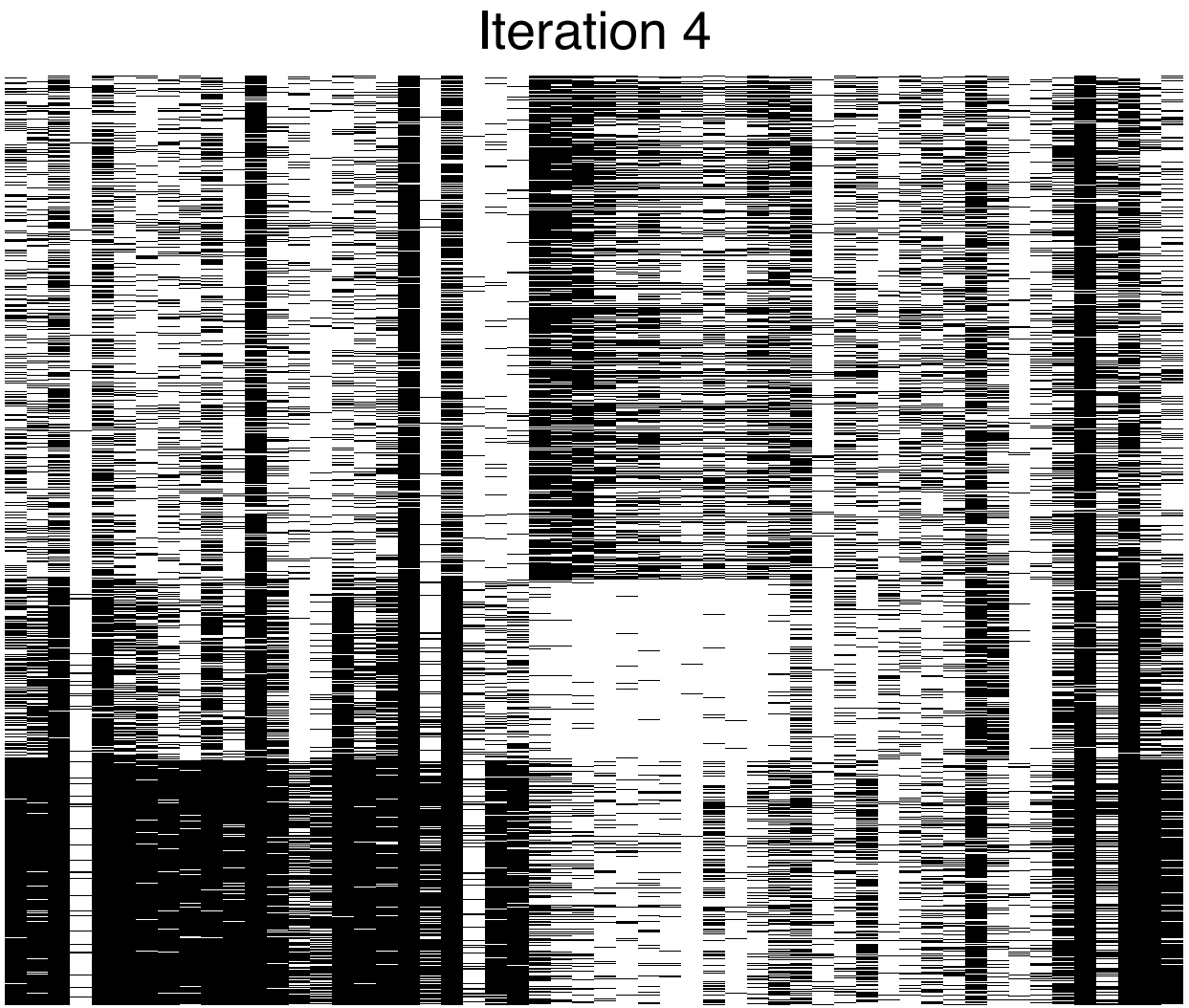
An EM run with a binary data set



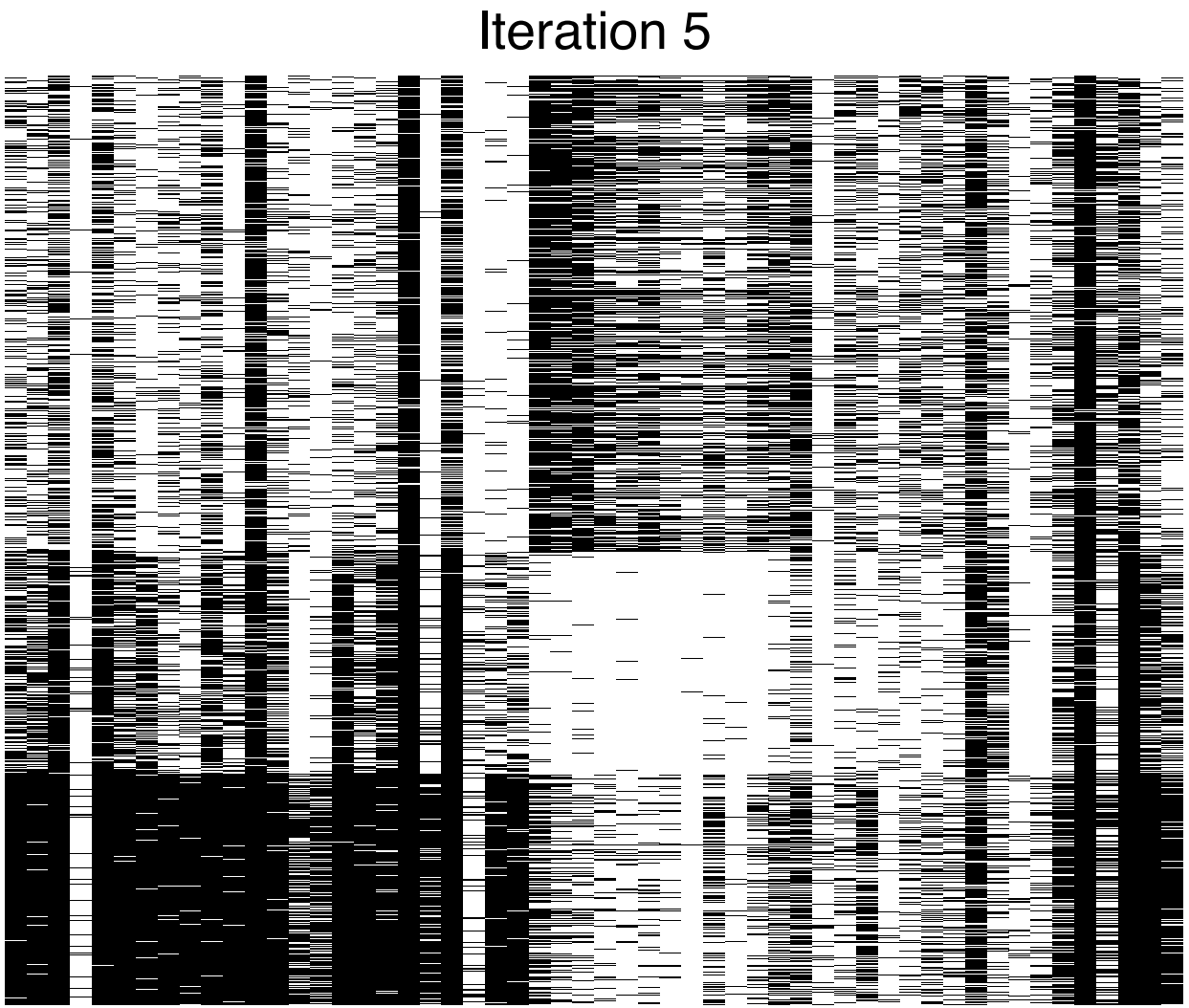
An EM run with a binary data set



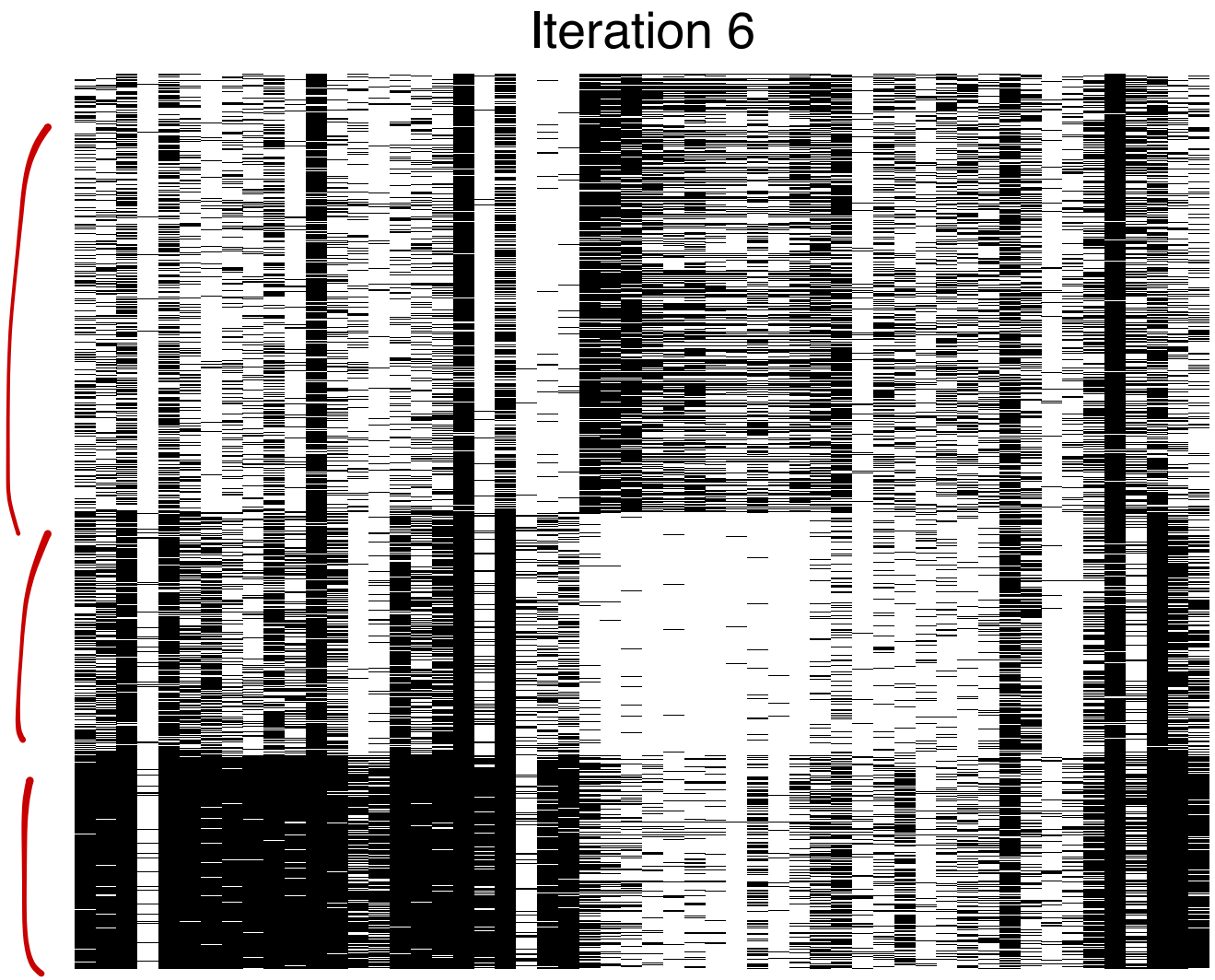
An EM run with a binary data set



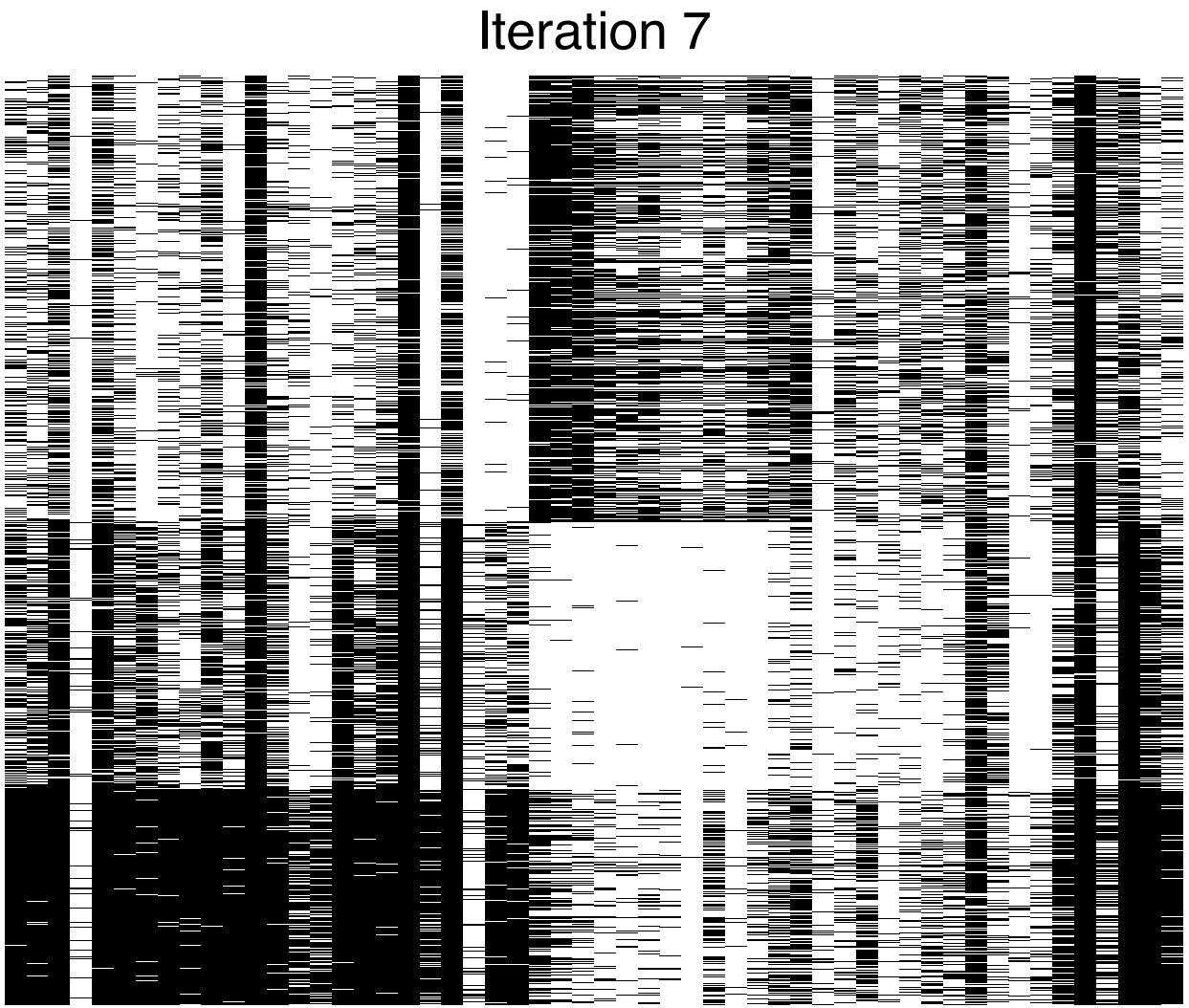
An EM run with a binary data set



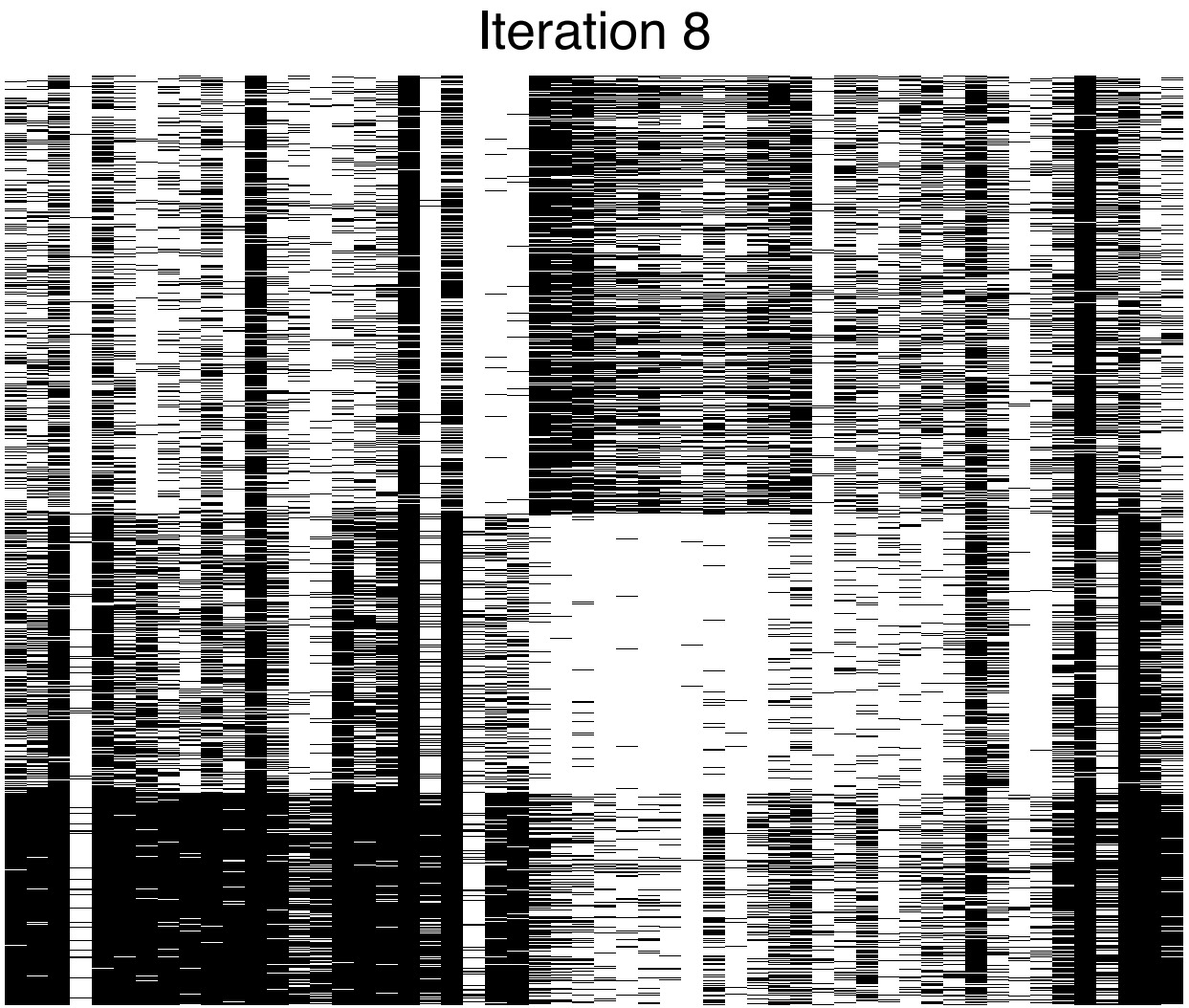
An EM run with a binary data set



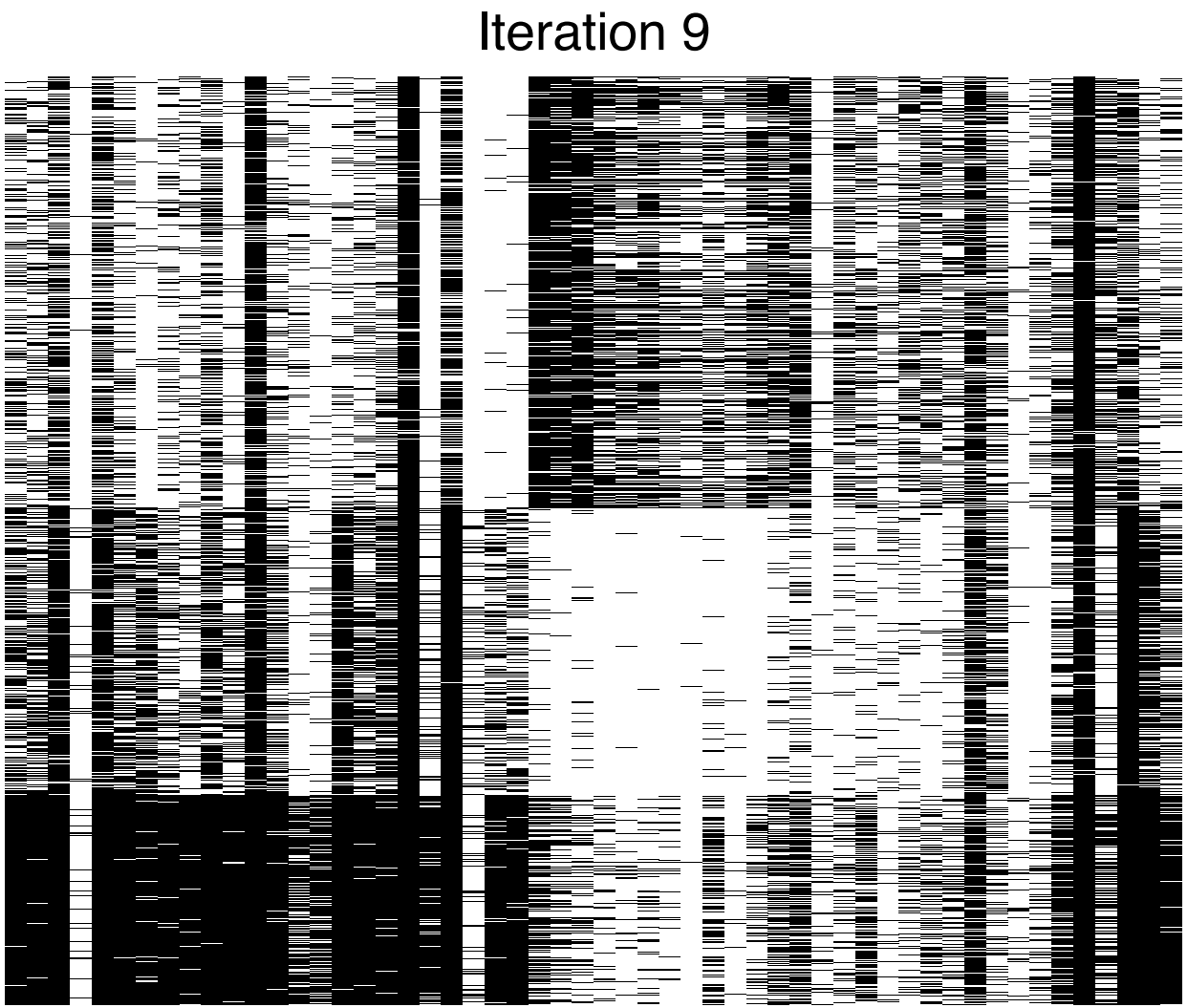
An EM run with a binary data set



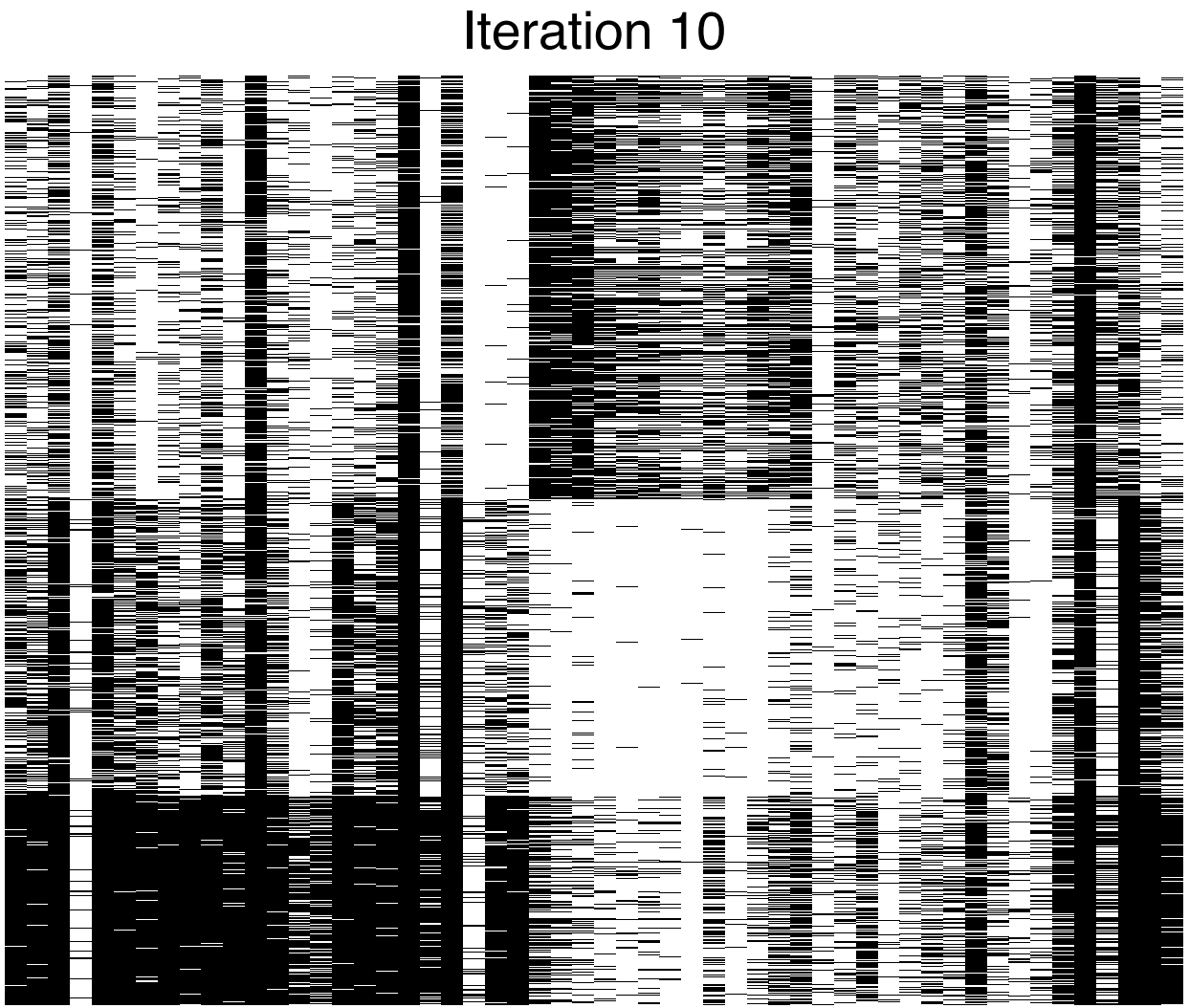
An EM run with a binary data set



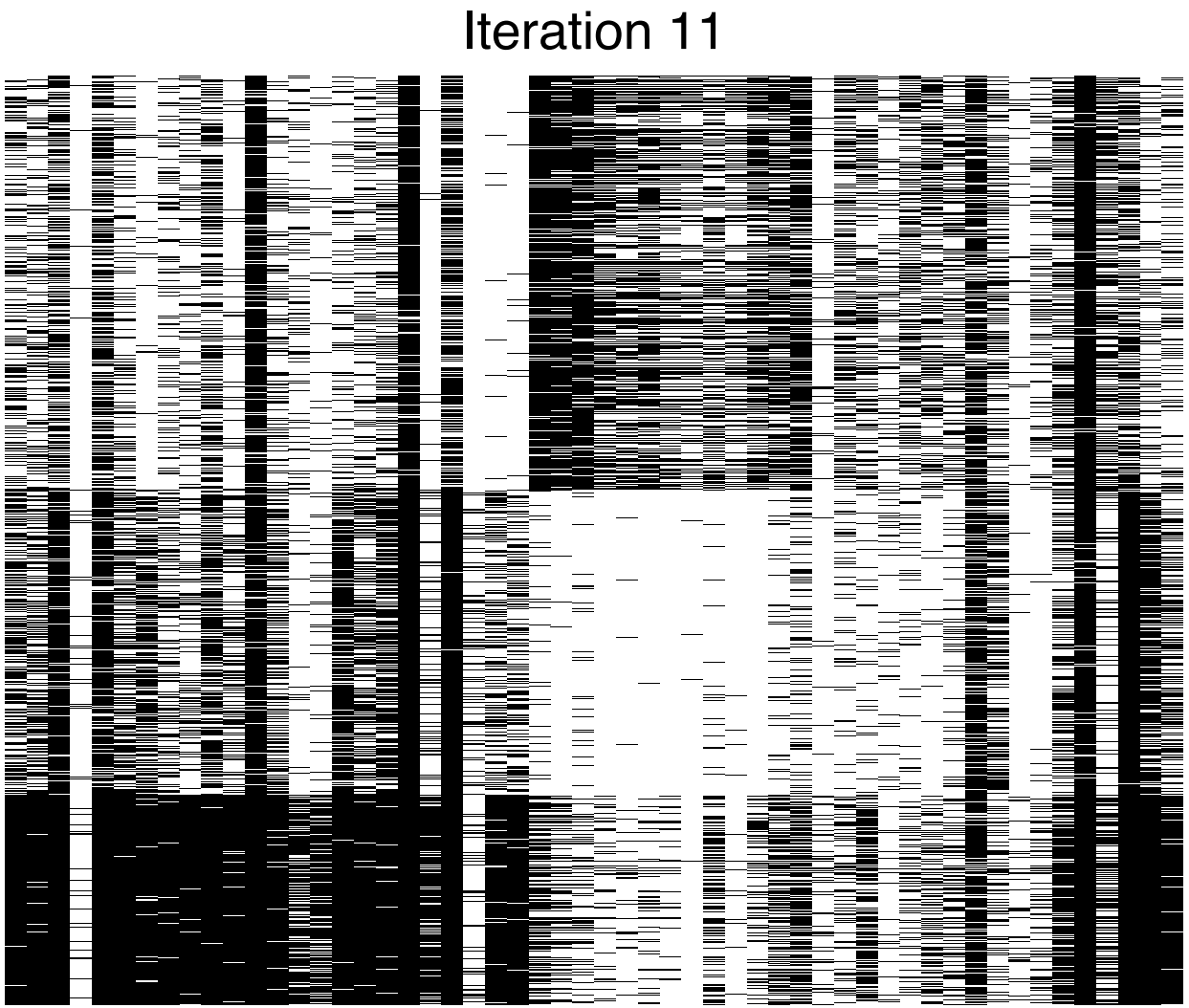
An EM run with a binary data set



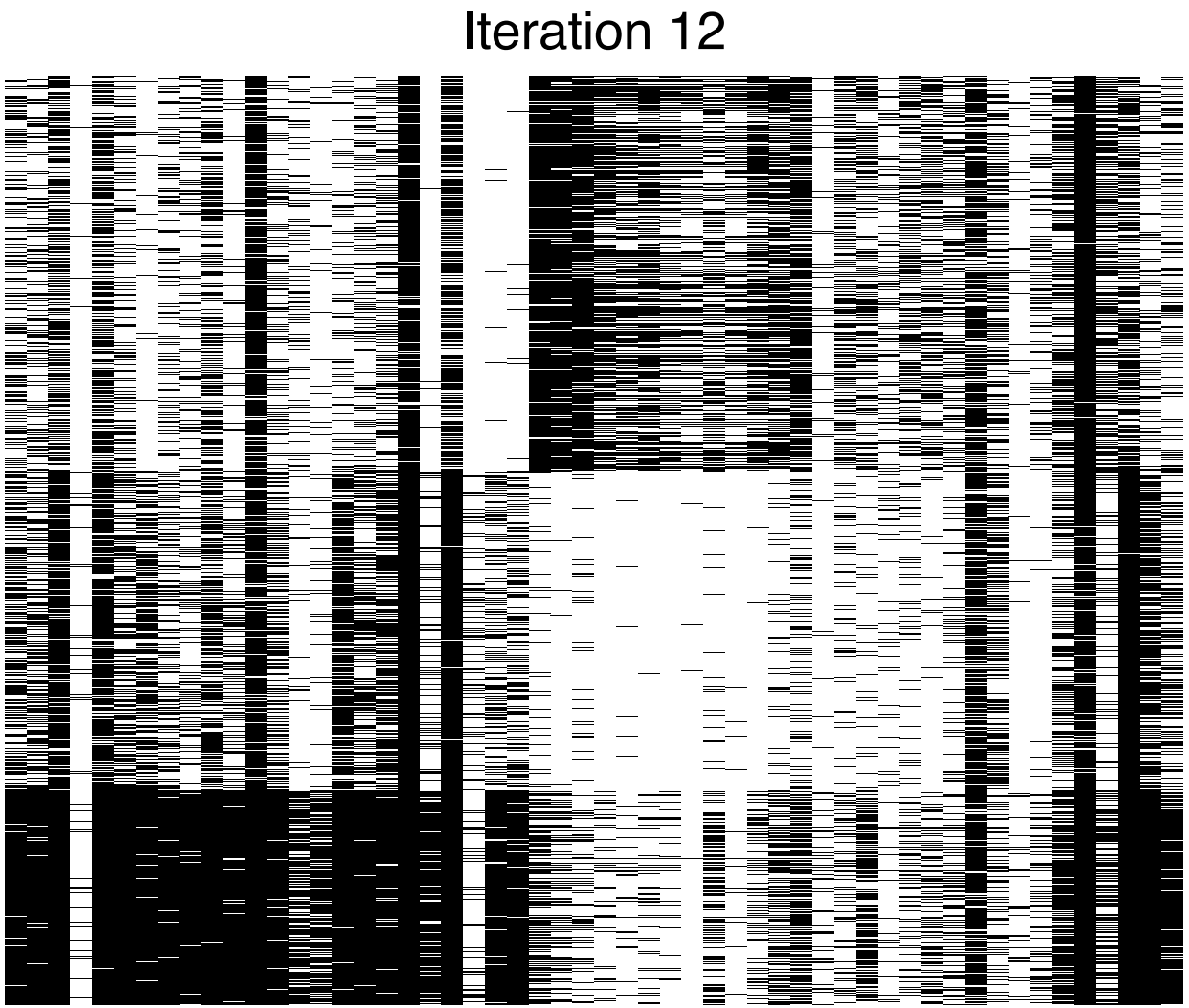
An EM run with a binary data set



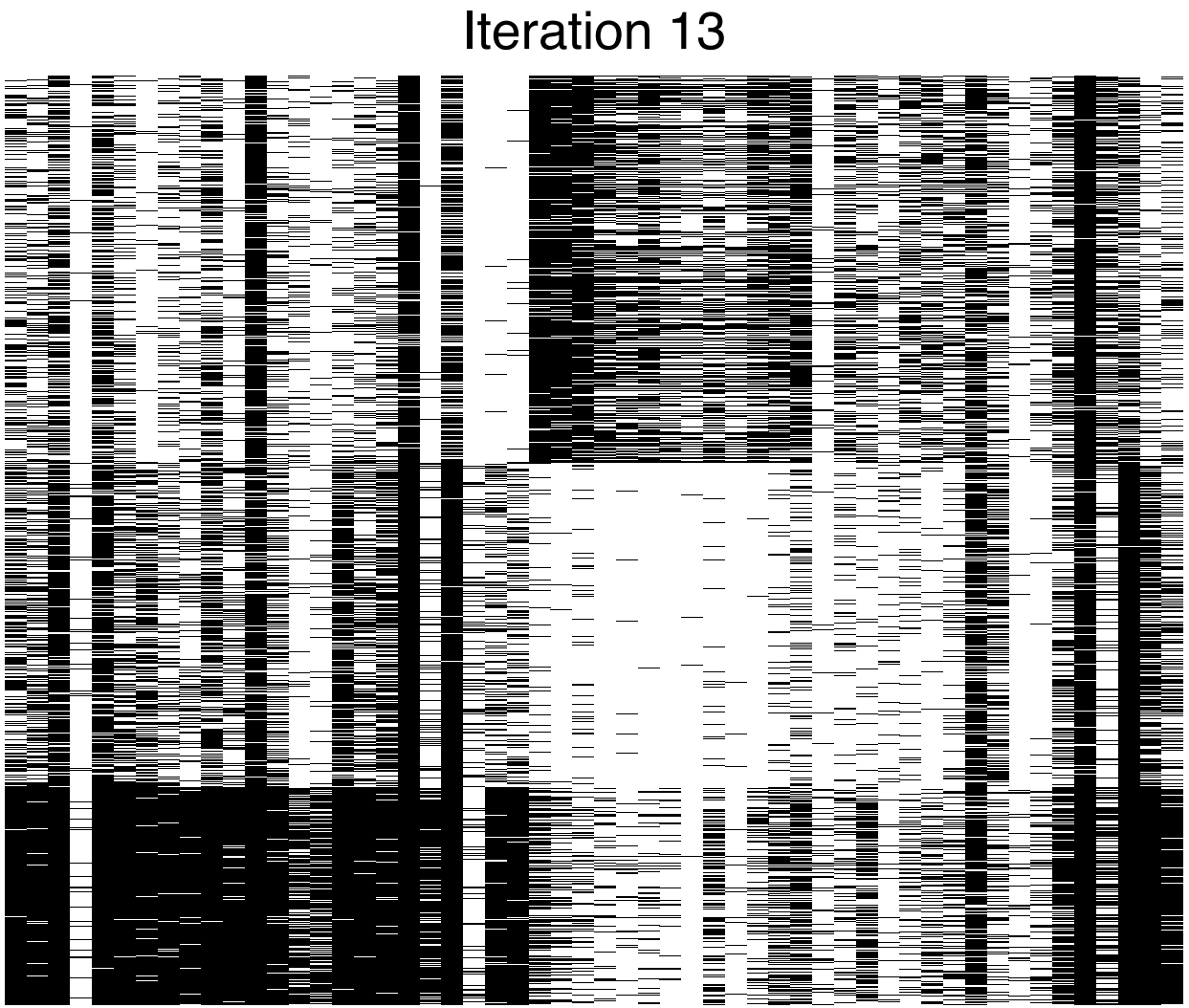
An EM run with a binary data set



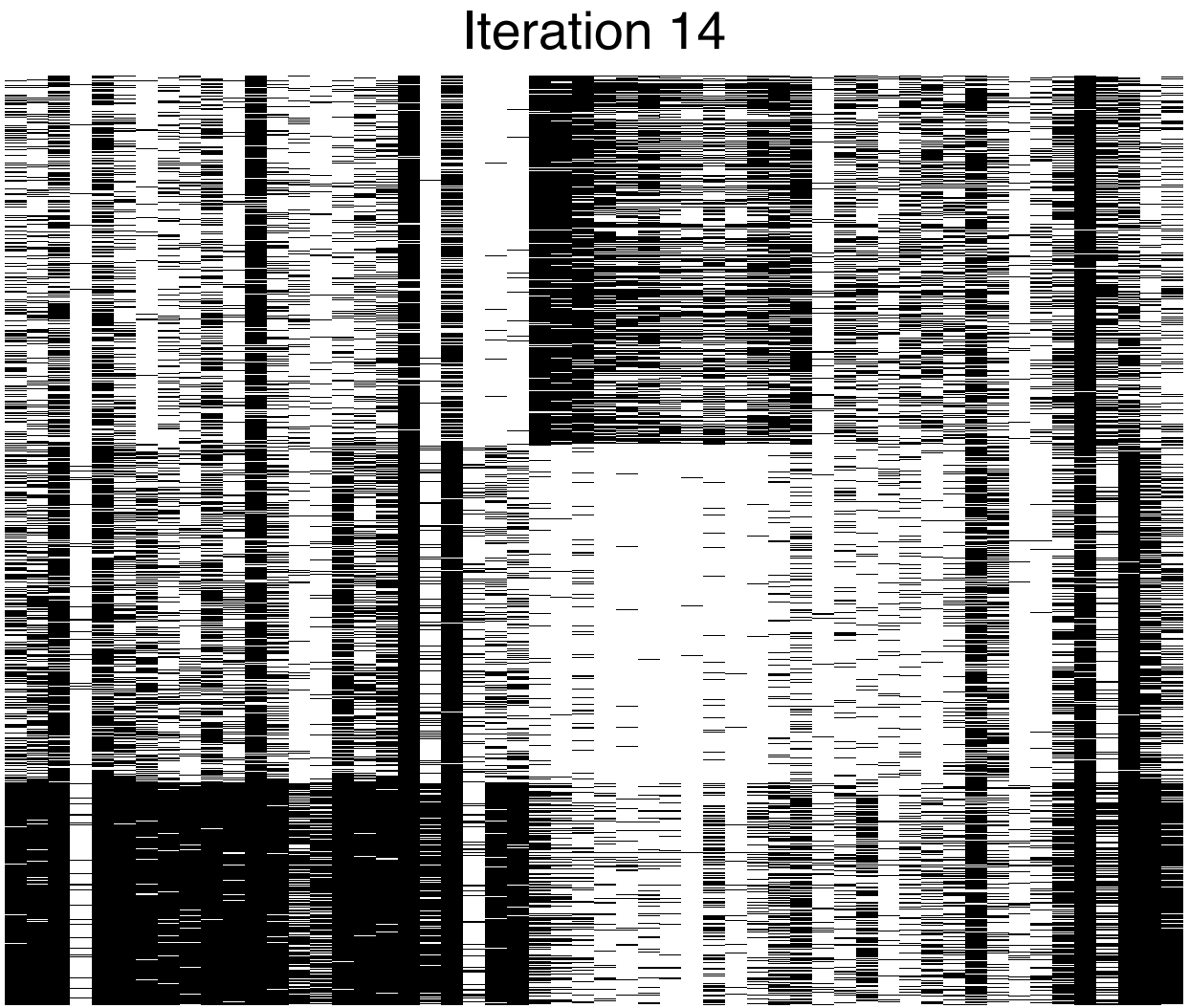
An EM run with a binary data set



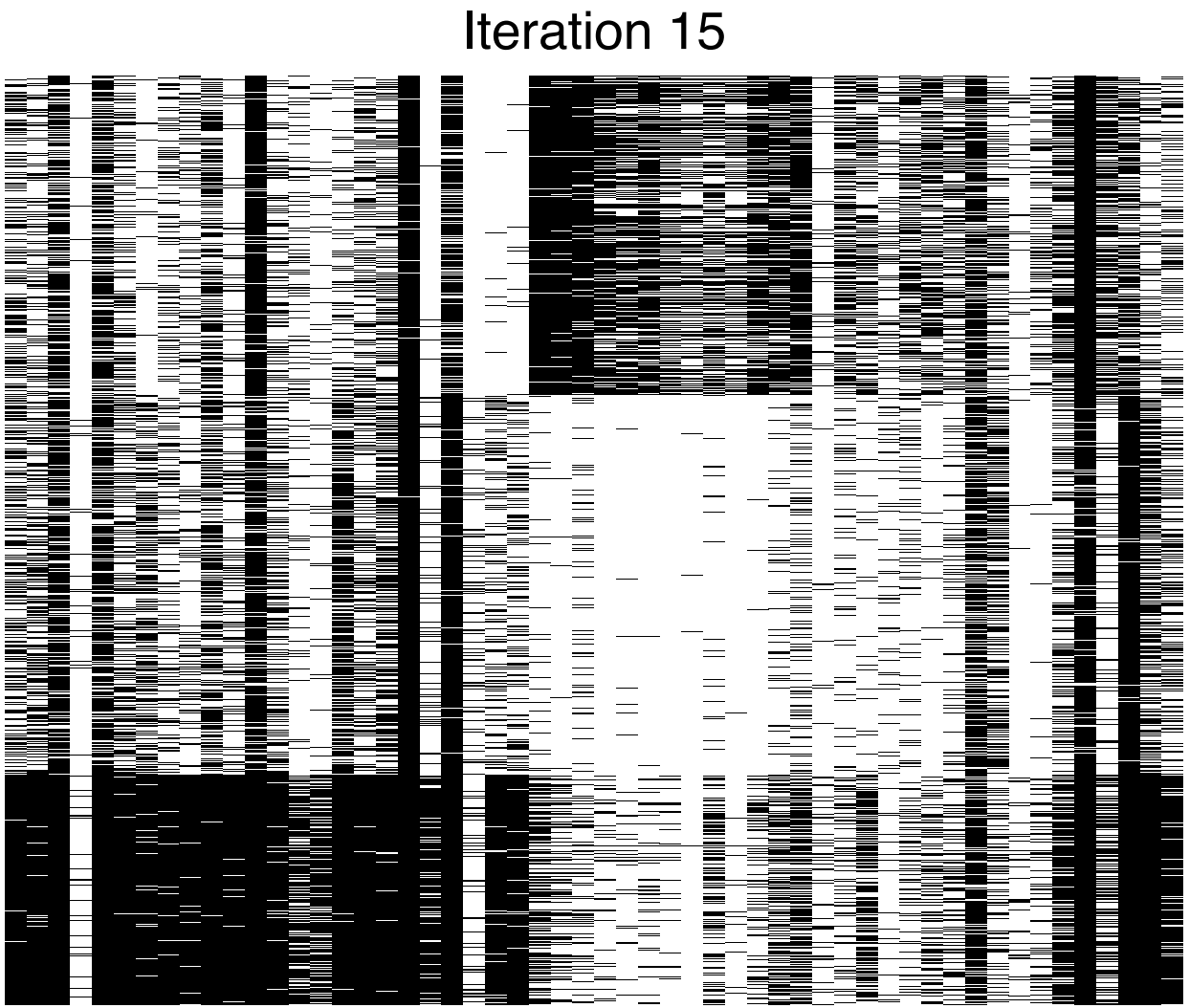
An EM run with a binary data set



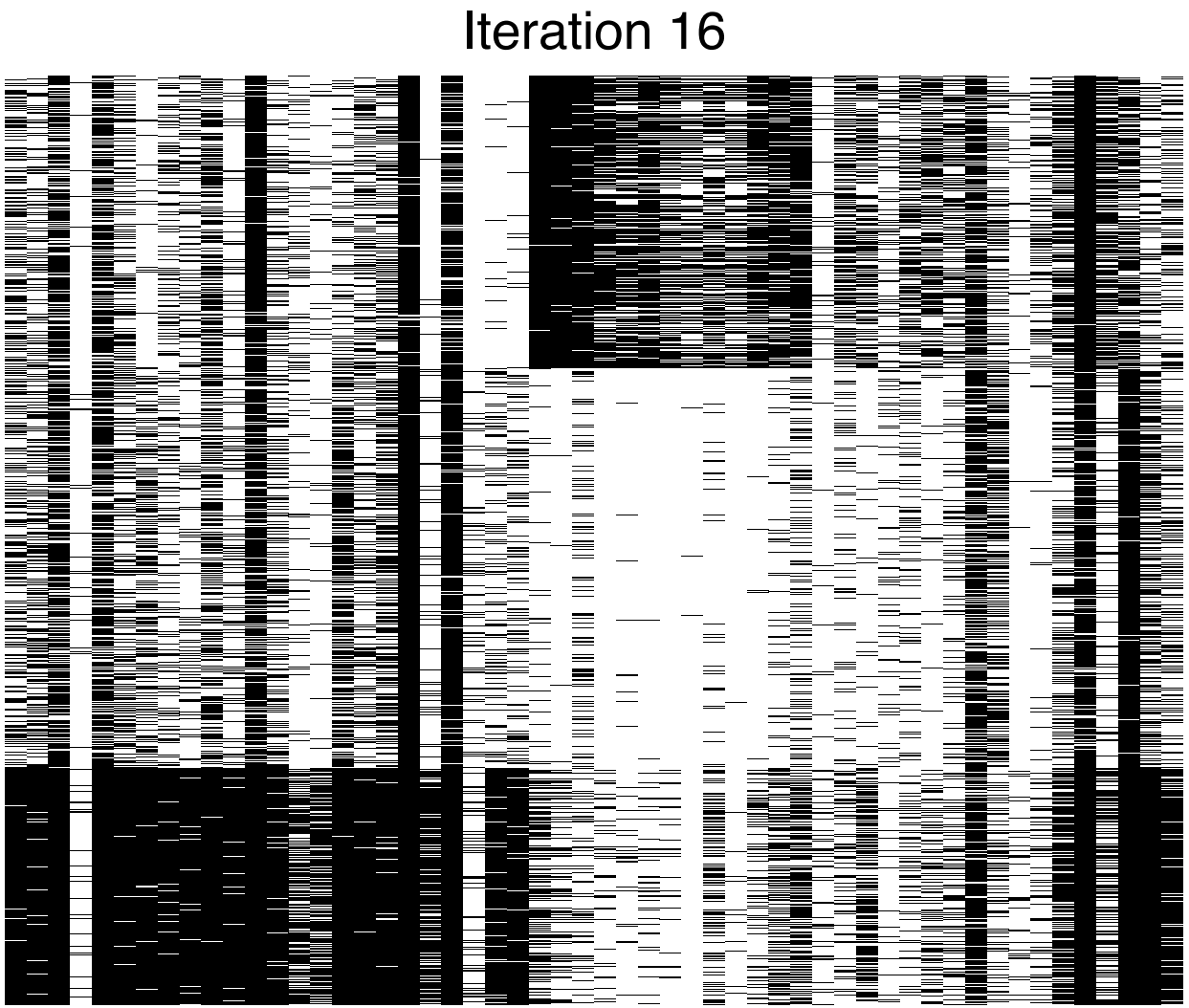
An EM run with a binary data set



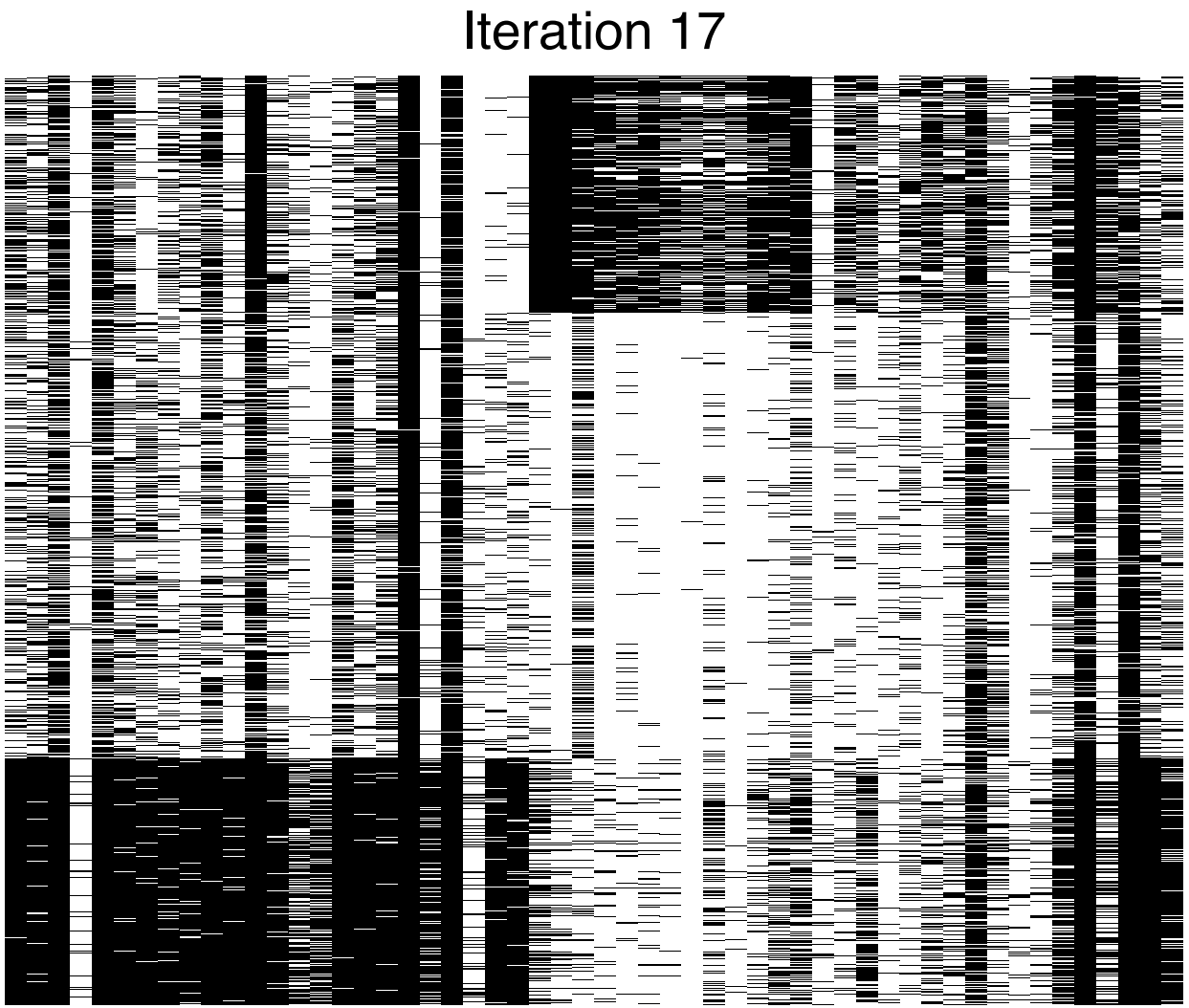
An EM run with a binary data set



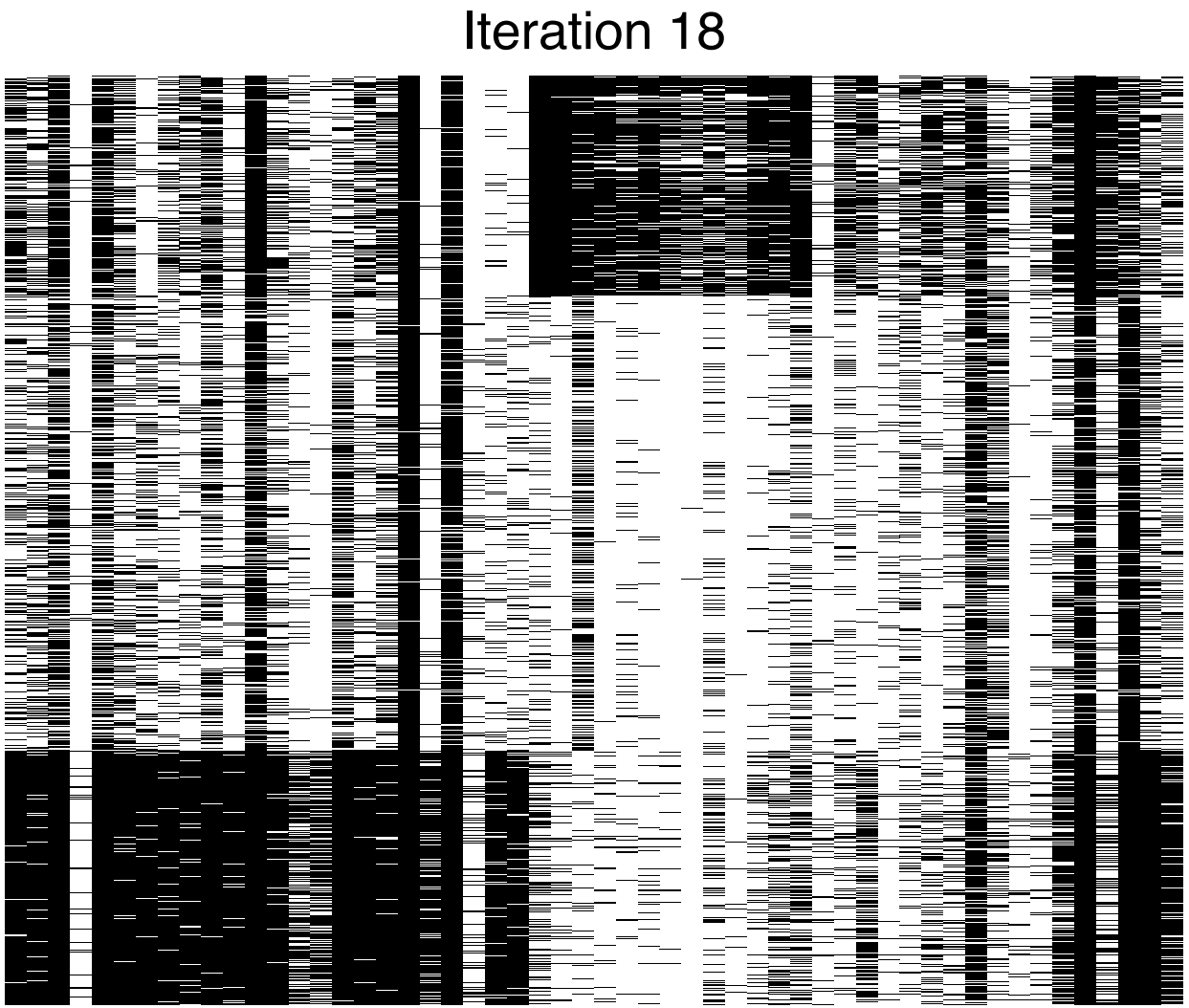
An EM run with a binary data set



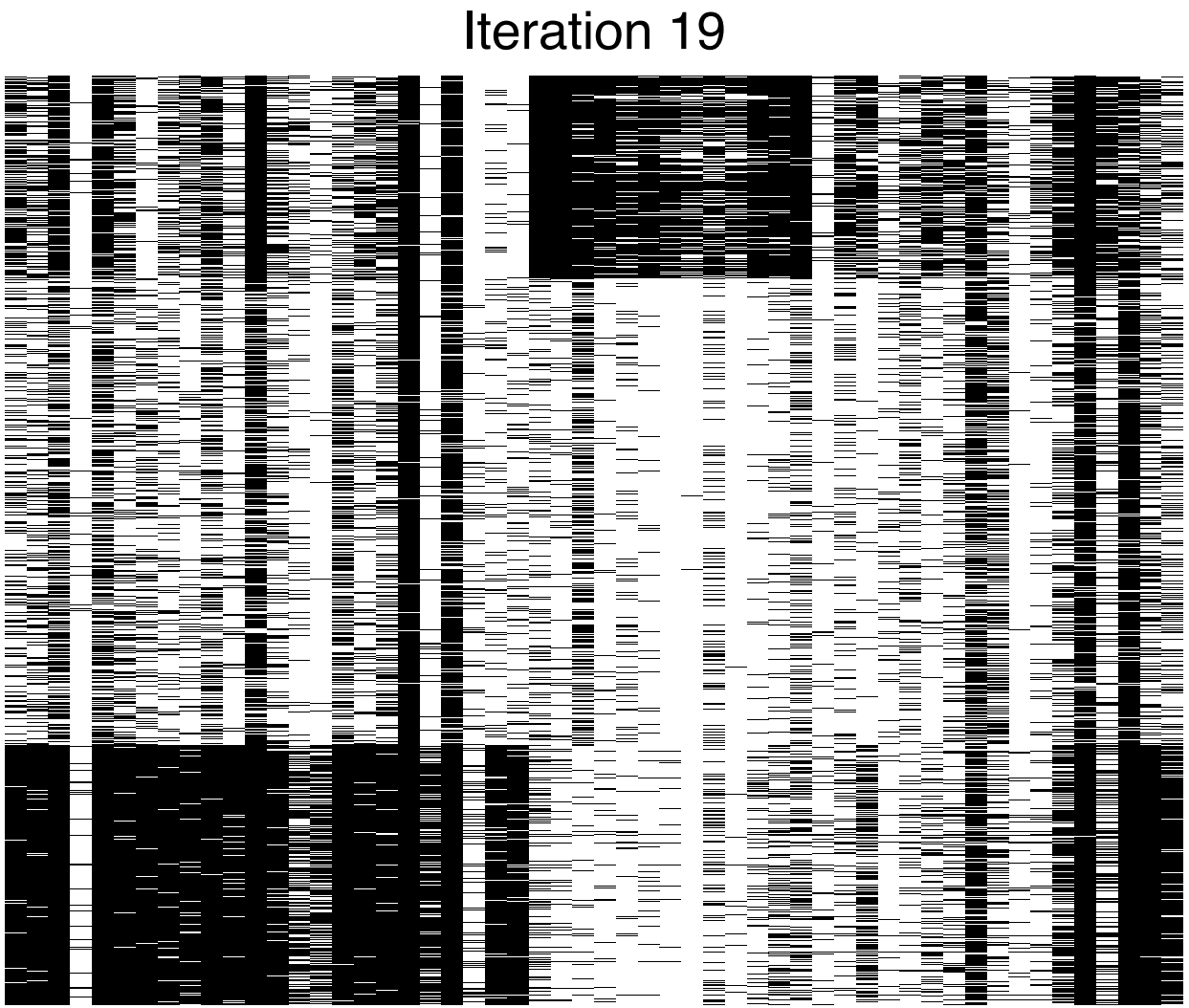
An EM run with a binary data set



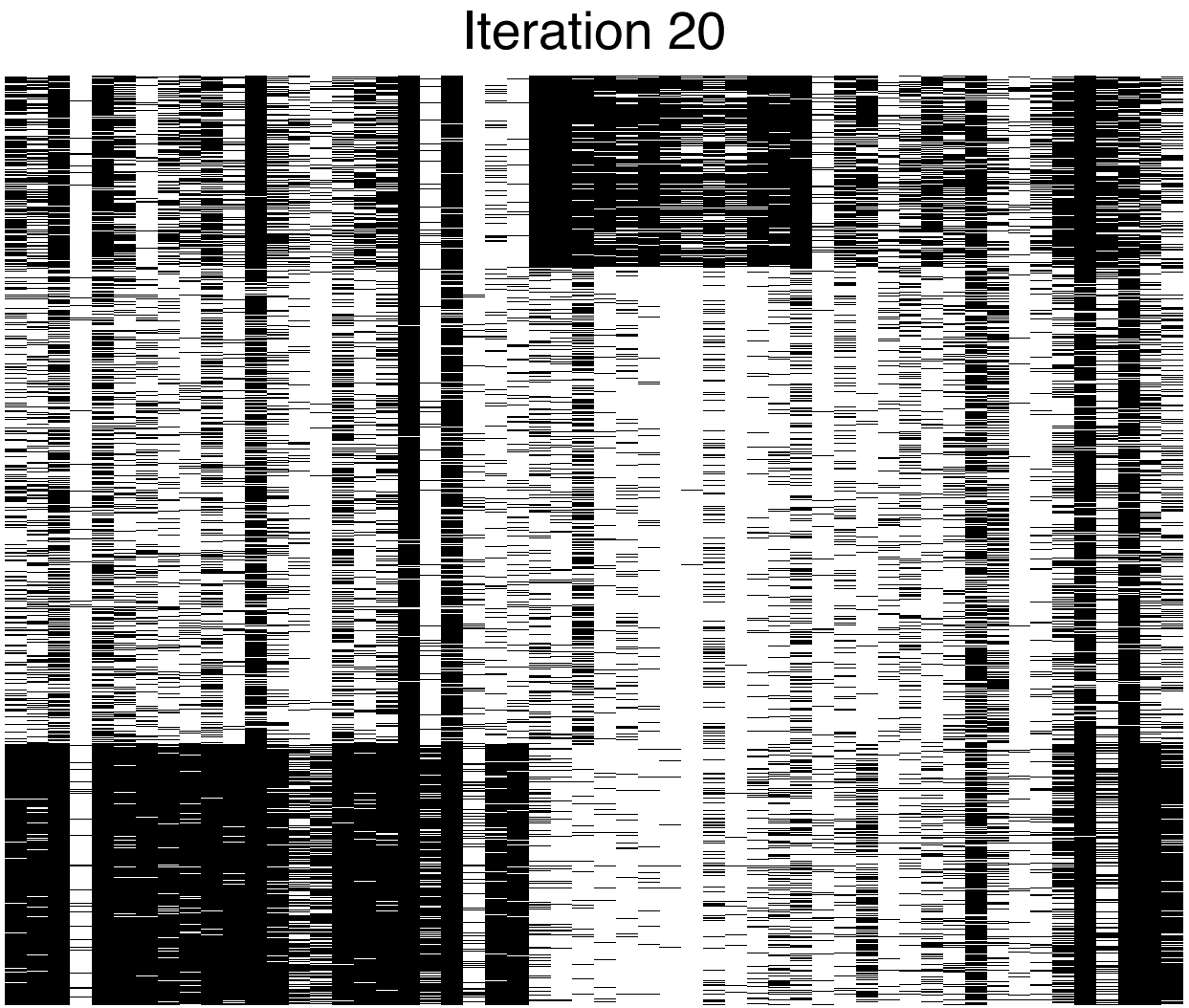
An EM run with a binary data set



An EM run with a binary data set



An EM run with a binary data set



Mixed data: classical approaches

Usually, unify data type by transformation :

- Quantify continuous variables: loose some information
- MCA of categorical variable: loose the meaning
- ...

Proposal

Model-based directly on raw data

Mixed data: conditional independence everywhere[9]

The aim is to combine continuous, categorical and integer data

$$\boldsymbol{x}_1 = (\boldsymbol{x}_1^{cont}, \boldsymbol{x}_1^{cat}, \boldsymbol{x}_1^{int})$$

The proposed solution is to mixed all types by inter-type conditional independence

$$\mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k) = \mathrm{p}(\mathbf{x}_1^{cont}; \boldsymbol{\alpha}_k^{cont}) \times \mathrm{p}(\mathbf{x}_1^{cat}; \boldsymbol{\alpha}_k^{cat}) \times \mathrm{p}(\mathbf{x}_1^{int}; \boldsymbol{\alpha}_k^{int})$$

In addition, for symmetry between types, intra-type conditional independence

> Only need to define the univariate pdf for each variable type!

- Continuous: Gaussian
- Categorical: multinomial
- Integer: Poisson

[9] MixtComp software on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

Missing data: current solutions

X_1	X_2	X_3	Cluster
1.23	?	3.42	?
?	?	4.10	?
4.53	1.50	5.35	?
?	5.67	?	?

Discarded solutions

- Suppress units and/or variables with missing data $\Rightarrow$ loss of information
- Imputation of the missing data by the mean or more evolved methods $\Rightarrow$ uncertainty of the prediction not taken into account

Retained solution

Use an integrated approach which allows to take into account all the available information to perform clustering

Missing data: MNAR assumption and estimation

Assumption on the missingness mecanism

Missing At Randon (MAR): the probability that a variable is missing does not depend on its own value given the observed variables.

Observed log-likelihood...

$$\ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \mathrm{p}(\boldsymbol{x}_i^{\mathrm{O}}; \boldsymbol{\alpha}_k) \right) = \ln \left[\sum_{k=1}^{K} \pi_k \underbrace{\int_{\mathbf{x}_i^M} \mathrm{p}(\mathbf{x}_i^O, \mathbf{x}_i^M; \boldsymbol{\alpha}_k) d\mathbf{x}_i^M}_{\text{MAR assumption}} \right]$$

Missing data: SEM algorithm[10]

A SEM algorithm to estimate $\boldsymbol{\theta}$ by maximizing the observed-data log-likelihood

- Initialisation: $\boldsymbol{\theta}^{(0)}$
- Iteration nb q:
  - E-step: compute conditional probabilities $\mathrm{p}(\boldsymbol{x}^M, \boldsymbol{z}|\boldsymbol{x}^0; \boldsymbol{\theta}^{(q)})$
  - S-step: draw $(\boldsymbol{x}^{M(q)}, \boldsymbol{z}^{(q)})$ from $\mathrm{p}(\boldsymbol{x}^M, \boldsymbol{z}|\boldsymbol{x}^0; \boldsymbol{\theta}^{(q)})$
  - M-step: maximize $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} \ln \mathrm{p}(\mathbf{x}^O, \boldsymbol{x}^{M(q)}, \boldsymbol{z}^{(q)}; \boldsymbol{\theta})$
- Stopping rule: iteration number

Properties: simpler than EM and interesting properties!

- Avoid possibly difficult E-step in an EM
- Classical M steps
- Avoids local maxima
- The mean of the sequence $(\boldsymbol{\theta}^{(q)})$ approximates $\hat{\boldsymbol{\theta}}$
- The variance of the sequence $(\boldsymbol{\theta}^{(q)})$ gives confidence intervals

[10] MixtComp software on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

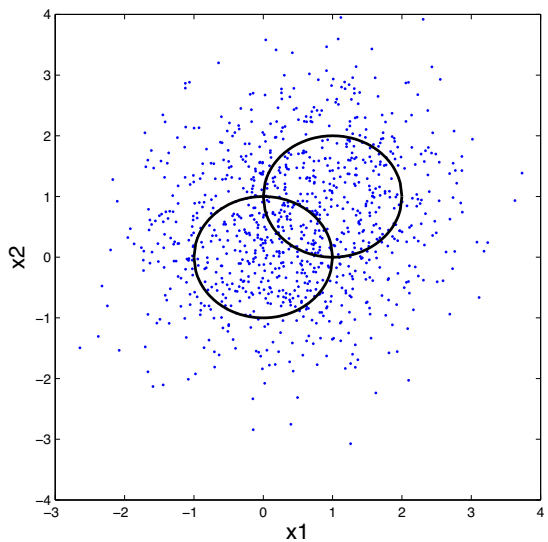
Outline

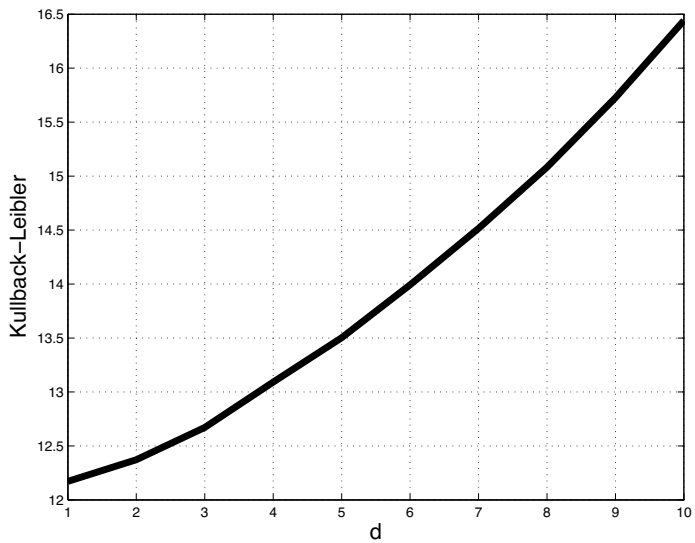
Curse: HD density estimation

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: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{I}} = \sqrt{d}\ldots$





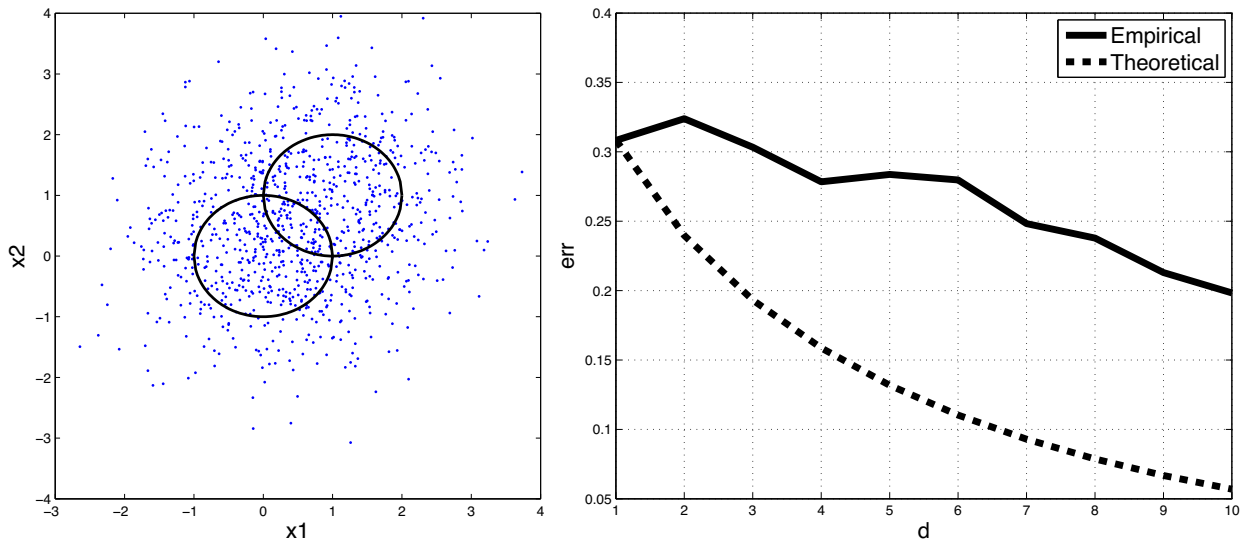
. . . but density estimation quality decreases with d

Blessing: HD clustering (1/2)

Each variable provides equal and own separation information

(Same parameter setting as before)

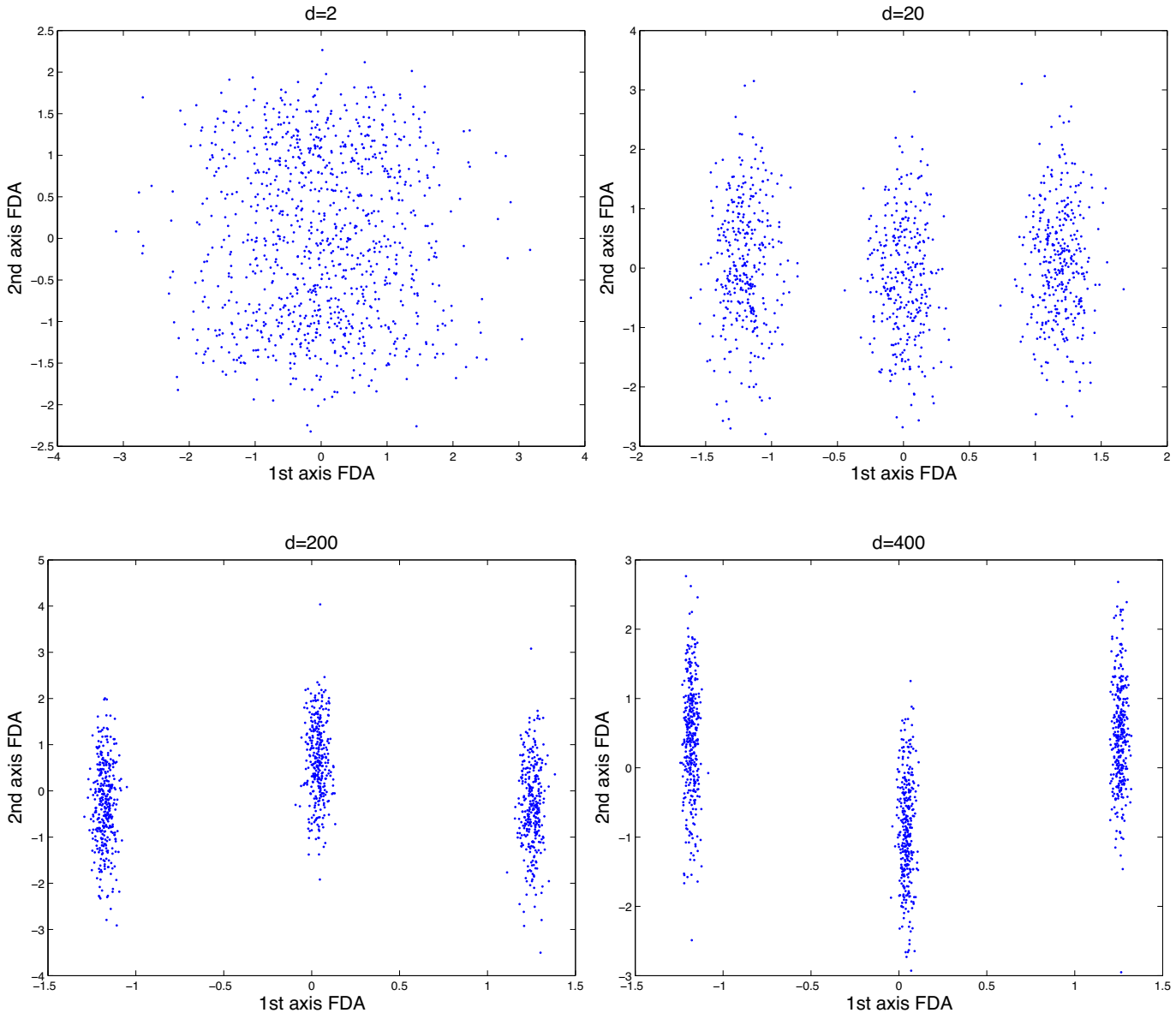
Theoretical error decreases when d grows: $\text{err}_{theo} = \Phi(-\sqrt{d}/2)\ldots$



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

Blessing: HD clustering (2/2)

FDA



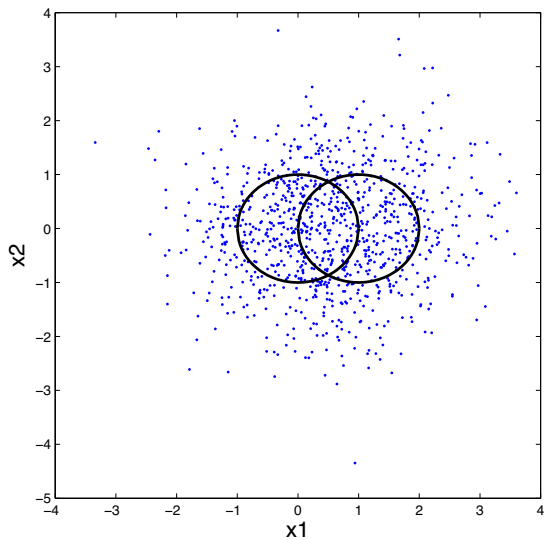
Curse: HD clustering (1/2)

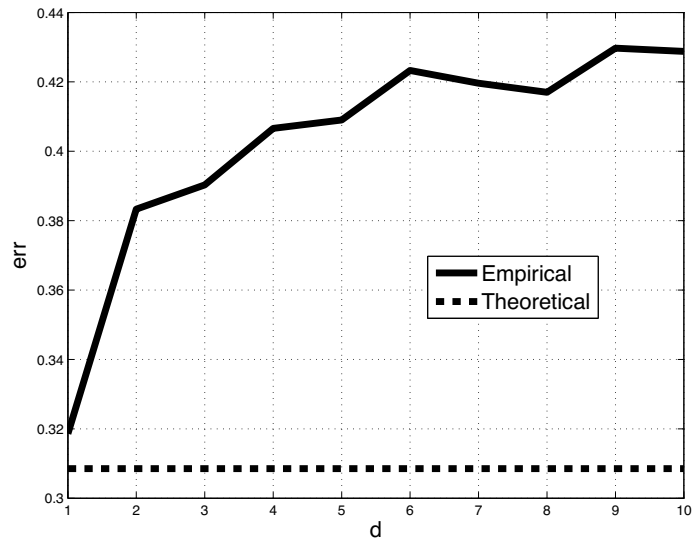
Many variables provide no separation information

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: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{I}} = 1 \ldots$





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

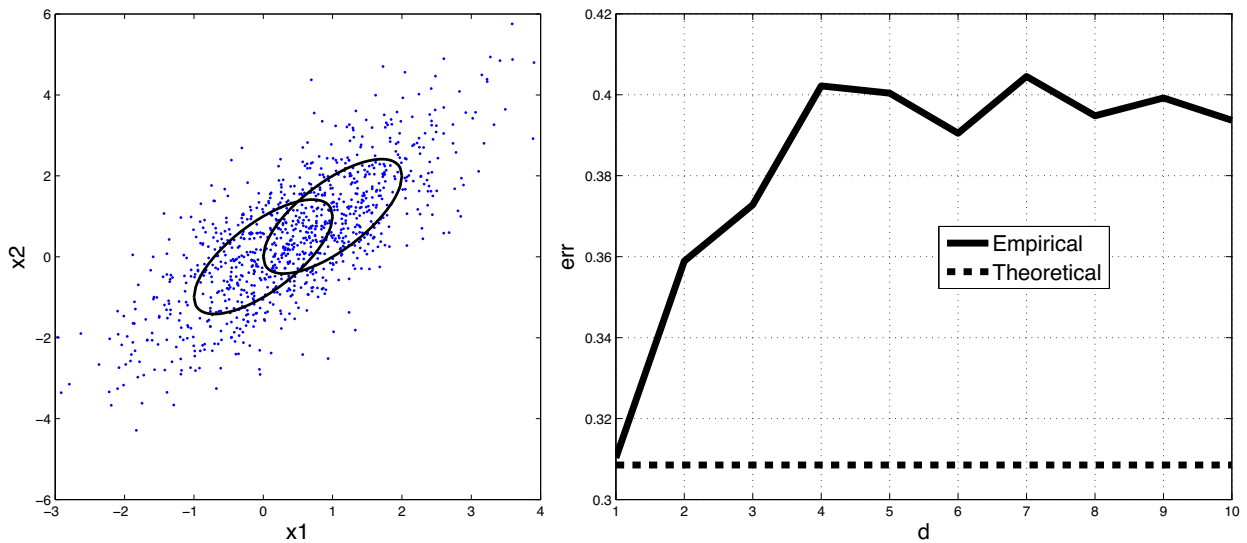
Curse: HD clustering (2/2)

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

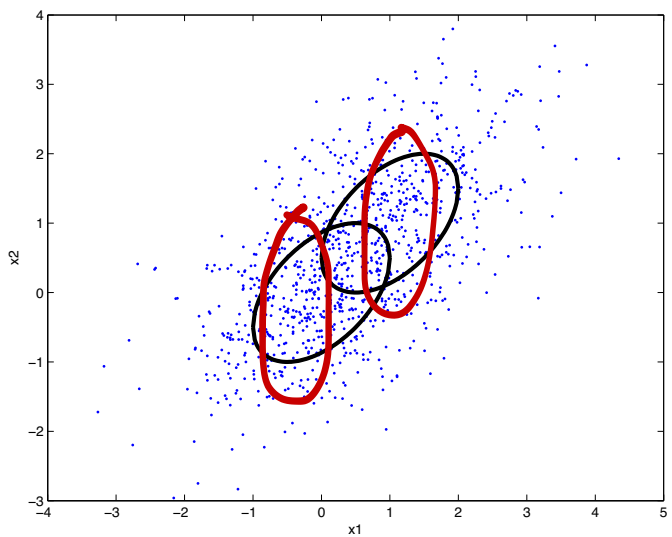
$$\begin{aligned}
\mathrm{err}(\mathbf{z}, \hat{\mathbf{z}}) &= \Big\{\mathrm{err}(\mathbf{z}, \mathbf{z}^*) - \mathrm{err}(\mathbf{z}, \mathbf{z})\Big\} + \Big\{\mathrm{err}(\mathbf{z}, \hat{\mathbf{z}}) - \mathrm{err}(\mathbf{z}, \mathbf{z}^*)\Big\} \\
&= \Big\{\text{bias}\Big\} + \Big\{\text{variance}\Big\}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

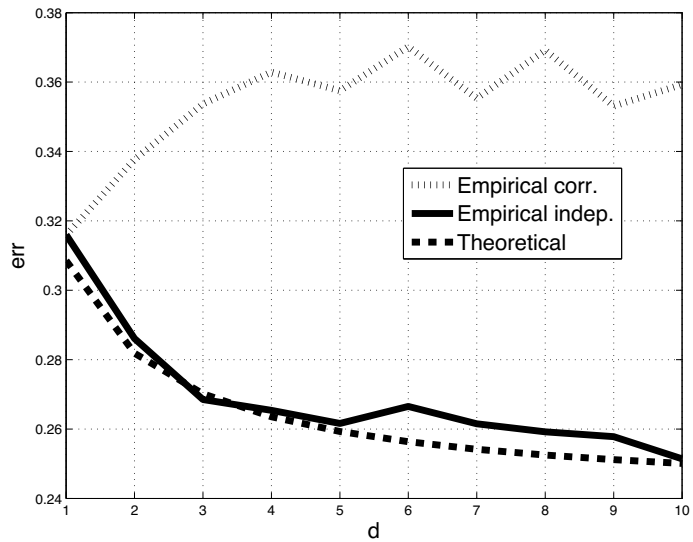
Bias/variance in HD: reduce variance, accept bias

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!





Intermediate conclusion

Blessing consequences

- Perform clustering in the whole data space
- Do not use "filter" methods where variable selection is performed before the clustering task [Jouve and Nicoloyannis, 05]
- Thus, prefer "wrapper" methods (see many examples later)

Curse consequences

- Impose parsimony on models designed in this whole data space (see [Bouveyron and Brunet, 14] for a review)
- Two kinds of wrapper methods: parsimony in the canonical variable space, or not
- Do not hesitate to introduce bias (it justifies somewhat conditional independence)

Outline

Gaussian mixture of factor analysers[11]

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

Complexity (some more parsimonious versions exist)

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

[11] pgmm package: `http://cran.r-project.org/web/packages/pgmm/index.html`

HD Gaussian models (1/2)[12]

Definition

[Bouveyron *et al.*, 07]

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

where

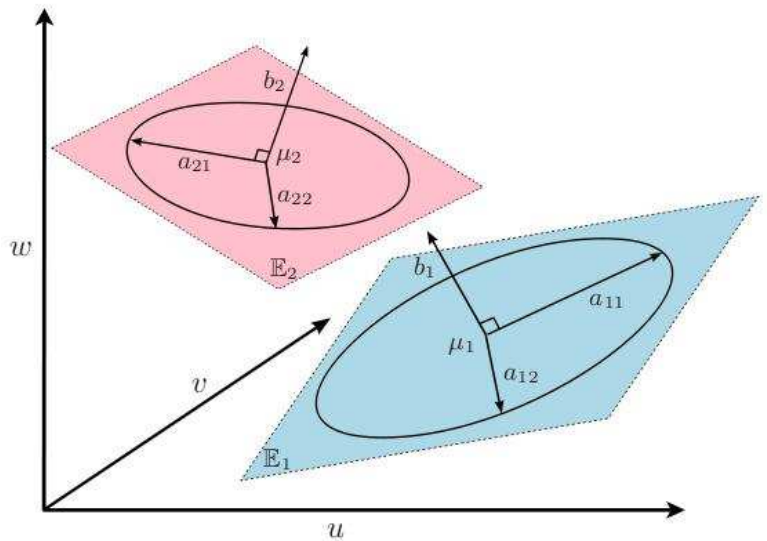
- $\mathbf{D}_k$ is the orthogonal matrix of the eigenvectors of $\boldsymbol{\Sigma}_k$
- $\boldsymbol{\Delta}_k$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}_k$

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

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

[12] Mixmod software on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

HD Gaussian models (2/2)



Complexity (some more parsimonious versions exist)

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

Functional data (1/4)[13]

[Jacques and Preda, 13]

Data: n data of ordered m_i time-points $\{X(t_{is}), 0 \leq s \leq m_i, t_{is\in[0,T]}\}$ $(i = 1, \ldots, n)$

Model:

- n curves $\mathbf{Y}_i = \{\mathbf{Y}_i(t), t \in [0, T]\}$ discretized each in m_i time-points $\{Y(t_{is}), 0 \leq s \leq m_i, t_{is\in[0,T]}\}$
- a basis of d (B-splines) functions $\{\phi_j\}_{j=1,\ldots,d}$

$$Y_i(t) = \sum_{j=1}^{d} \gamma_{ij}\phi_j(t)$$

- error on observation

$$X_i(t_{is}) = Y_i(t_{is}) + \varepsilon_{is}$$

Estimation: regression

$$\hat{\gamma}_i = (\mathbf{\Phi}_i'\mathbf{\Phi}_i)^{-1}\mathbf{\Phi}_i'\mathbf{X}_i$$

where $\mathbf{\Phi}_i = (\phi_j(t_{is}))_{j,s}$ and $\mathbf{X}_i = (X_i(t_{i0}), \ldots, X_i(t_{im_i}))'$

[13] FunClustering package: `http://cran.r-project.org/web/packages/Funclustering/index.html`

Functional data (2/4)

Functional PCA:

- Matrix of coefficients $\mathbf{\Gamma} = (\gamma_{ij})$ $n \times d$
- Matrix of weights for centering curves $\mathbf{T} = \frac{1}{n}\mathbf{I}$
- Matrix of centered coefficients $\tilde{\mathbf{\Gamma}}$ of $\boldsymbol{\gamma}$ $n \times d$
- Matrix of the inner products $\mathbf{W} = (w_{jj'}) = \int_0^T \phi_j(t)\phi_{j'}(t)dt$ $(1 \leq j, j' \leq d)$
- Principal components (centered): the jth principal component score $\mathbf{C}_j$ is the jth eigenvector associated to the jth eigenvalue

$$\tilde{\mathbf{\Gamma}}\mathbf{W}\tilde{\mathbf{\Gamma}}'\mathbf{T}\mathbf{C}_j = \alpha_j \mathbf{C}_j$$

Trick: it is a kind of variable ordering

Gaussian process: if data $\{X(t_{is})\}$ arise from a Gaussian process $\{\boldsymbol{X}(t), t \in [0, T]\}$

$$\mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}) \approx \underbrace{\prod_{j=1}^{q}}_{\text{indep.}} \mathrm{p}(C_i^j; \underbrace{0}_{\text{centered}}, \alpha_j)$$

with $\mathrm{p}(\cdot; 0, \alpha_j)$ the univariate Gaussian of center 0 and variance α_j

Functional data (3/4)

Gaussian mixture model: for K groups, it is assumed the mixture

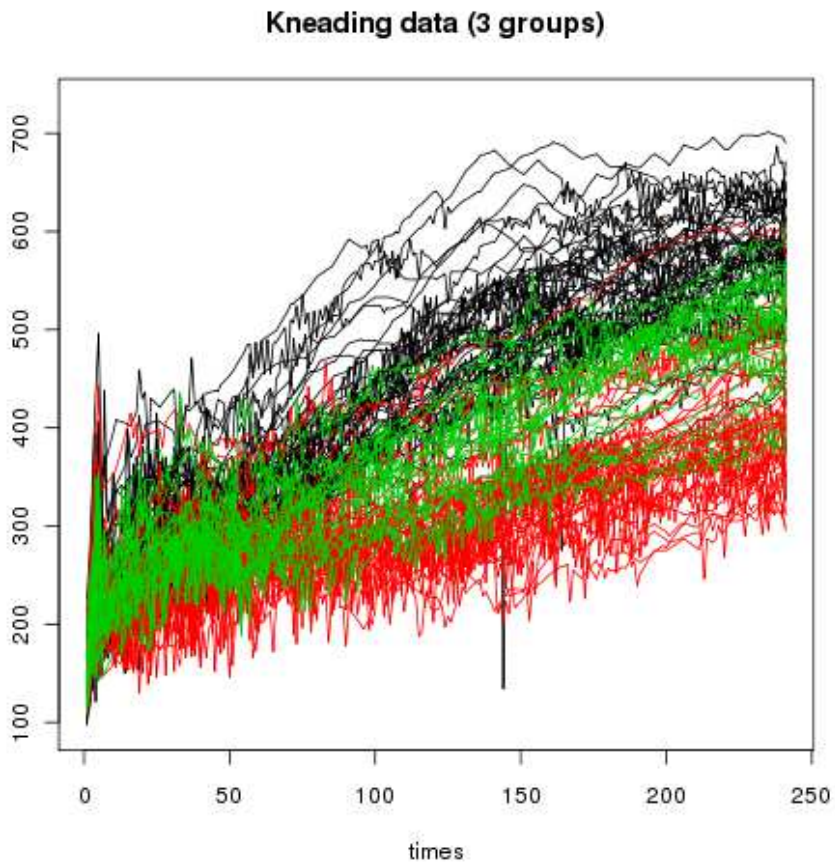
$$\mathrm{p}(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} \mathrm{p}(C_{ik}^j; 0, \alpha_{jk})$$

where C_{ik}^j is a group conditional score

Parameter estimation: EM-like algorithm for maximizing the pseudo log-likelihood

- E-step:
$$t_{ik} \propto \pi_k \mathrm{p}(C_{ik}^j; 0, \alpha_{jk})$$
- M-step:
  - Principal score upadate: weights T_k depends now on t_{ik}, also $\mathbf{\Gamma}_k$
  - q_k selection: a kind of elbow in the eigenvalues...
  - Parameters: π_k as usual, α_k from previous conditional PCA

Functional data (4/4)



Outline

Spherical and diagonal Gaussians[14]

Definition

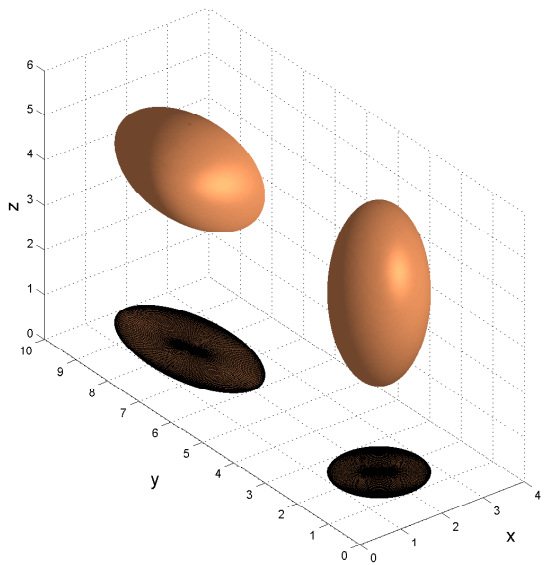
[Celeux and Govaert, 95]

$$\text{spherical: } \mathbf{\Sigma}_k = \lambda_k \mathbf{I} \quad \text{or} \quad \text{diagonal: } \mathbf{\Sigma}_k = \lambda_k \mathbf{B}_k$$

where $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$ and $\mathbf{B_k}$ diagonal with $|\mathbf{B_k}| = \mathbf{1}$

Complexity (more parsimonious versions exist)

$$\text{Spherical} : \nu = (K-1) + Kd + K, \quad \text{Diagonal} : \nu = (K-1) + Kd + Kd$$



[14] Mixmod software on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

Latent class model[15]

[Goodman, 74]

Categorical variables: d variables with m_j modalities each, $\mathbf{x}_i^j \in \{0,1\}^{m_j}$ and

$$\mathbf{x}_i^{jh} = 1 \quad \Leftrightarrow \quad \text{variable } j \text{ of } \mathbf{x}_i \text{ takes modality } h$$

Conditional independence:

$$p(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}$$

and

$$\alpha_k^{jh} = p(X_i^{jh} = 1 | Z_{ik} = 1)$$

with $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$

Complexity (more parsimonious versions exist)

$$\nu = (K-1) + d \prod_{j=1}^{d} (m_j - 1)$$

[15] Mixmod and MixtComp software on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

Mixed data model[16]

High dimensional can be mixed: categorical and continuous variables together

Model: combine (diagonal)parsimonious Gaussians and latent class model by conditional independence

$$\mathrm{p}_k(\mathbf{x}^{cont}, \mathbf{x}^{cat}) = \mathrm{p}_k(\mathbf{x}^{cont}) \times \mathrm{p}_k(\mathbf{x}^{cat})$$

Complexity
Still depend on d, thus not so parsimonious. . .

[16] Mixmod and MixtComp software on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

Gaussian "variable selection"[17][18]

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

[17] selvarclust package: `http://www.math.univ-toulouse.fr/~maugis/SelvarClustHomepage.html`

[18] selvarmix package: `http://cran.r-project.org/web/packages/SelvarMix/index.html`

Gaussian "variable selection": cruder version

Definition

[Pan and Shen, 07], [Zhou *et al.*, 09], [Meynet, 10]

$$\mathrm{p}(\mathbf{x}_1|\boldsymbol{\theta}) = \underbrace{\left\{\sum_{k=1}^{K} \mathrm{p}(\mathbf{x}_1^{J_r}; \boldsymbol{\mu}_k, \sigma^2\mathbf{I})\right\}}_{\text{relevant variables}} \times \underbrace{\left\{\mathrm{p}(\mathbf{x}_1^{J_a}; \boldsymbol{\mu}, \sigma^2\mathbf{I})\right\}}_{\text{active variables}} \times \underbrace{\left\{\mathrm{p}(\mathbf{x}_1^{J_i}; \mathbf{0}, \sigma^2\mathbf{I})\right\}}_{\text{irrelevant variables}}$$

where

- all parts are Gaussians
- $\{J_r, J_a, J_i\}$ is a partition of $\{1, \ldots, d\}$
- $\mathrm{p}(\mathbf{x}_1^{J_i}; \mathbf{0}, \sigma^2\mathbf{I})$: "variance killer" (crude assumption)

Outline

Some alternatives for reducing variance

Limitation of previous models

- They are often not parsimonious enough for (very) HD
- For instance, difficult as soon as $n < d$
- The most parsimonious versions are restricted to the Gaussian case

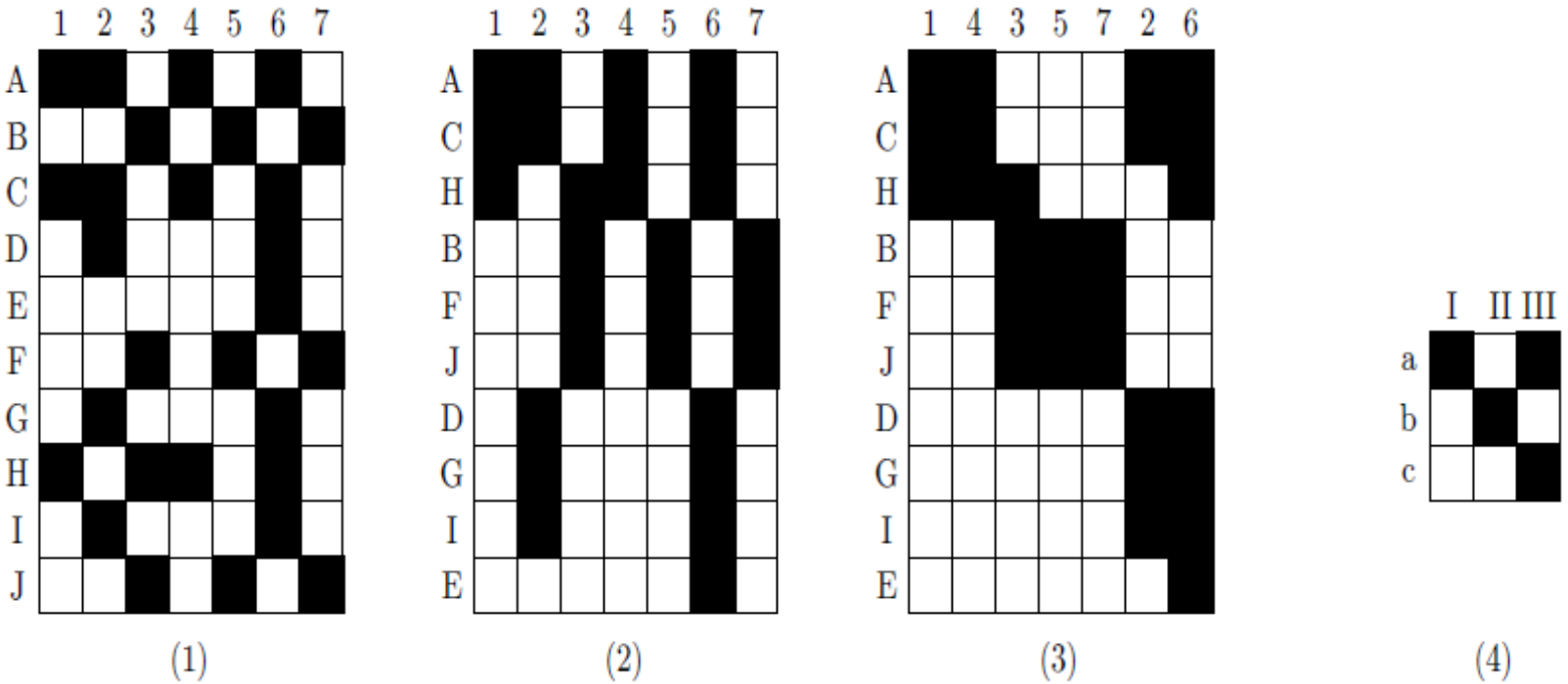
How to overcome these limitations?

- 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

From clustering to co-clustering



[Govaert, 2011]

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 clusters 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)

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[20]

[Govaert and Nadif, 2014]

The pdf $p(\cdot; \boldsymbol{\alpha}_{z_i w_j})$ depends on the kind of data x_{ij}:

- Binary data: $x_{ij} \in \{0, 1\}$, $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{B}(\alpha_{kl})$
- Categorical data with m levels:
  $\mathbf{x}_{ij} = \{x_{ijh}\} \in \{0, 1\}^m$ with $\sum_{h=1}^{m} x_{ijh} = 1$ and $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{M}(\boldsymbol{\alpha}_{kl})$ with $\boldsymbol{\alpha}_{kl} = \{\alpha_{kjh}\}$
- Count data: $x_i^j \in \mathbb{N}$, $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{P}(\mu_k \nu_l \gamma_{kl})$[19]
- Continuous data: $x_i^j \in \mathbb{R}$, $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{N}(\mu_{kl}, \sigma_{kl}^2)$

[19] 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.

[20] BlockCluster package on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

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

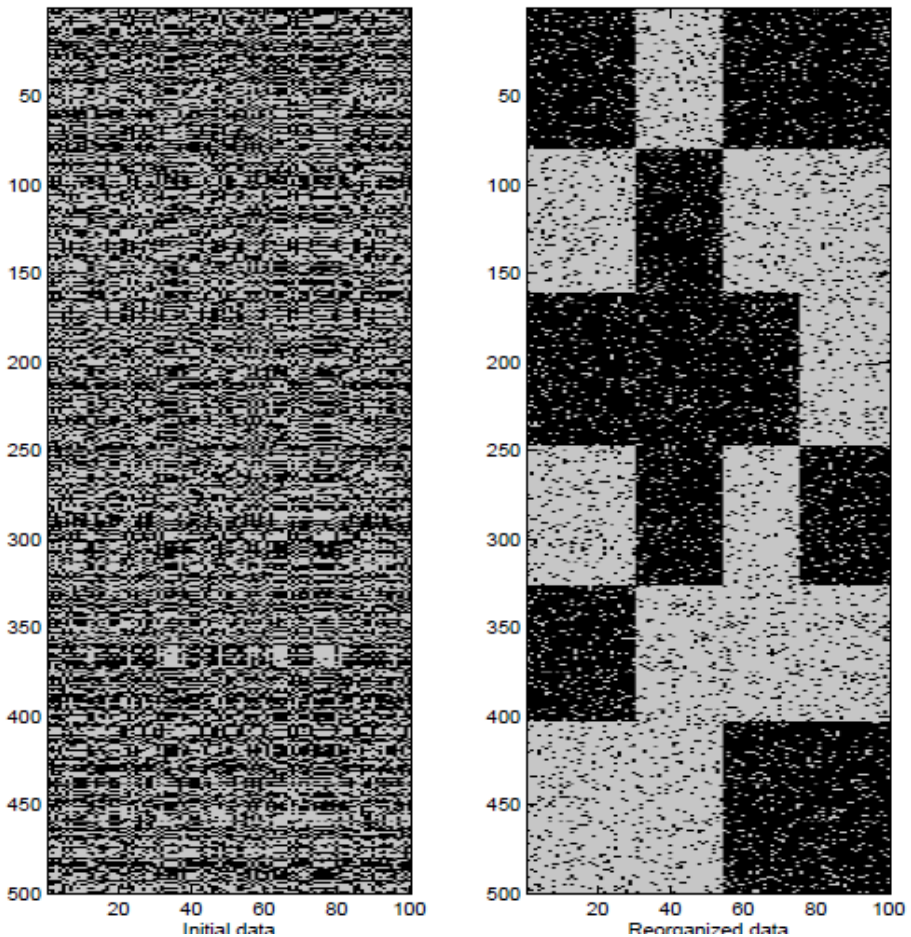
Homogeneity

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 (with missing values)

[Jacques and Biernacki, 2018]

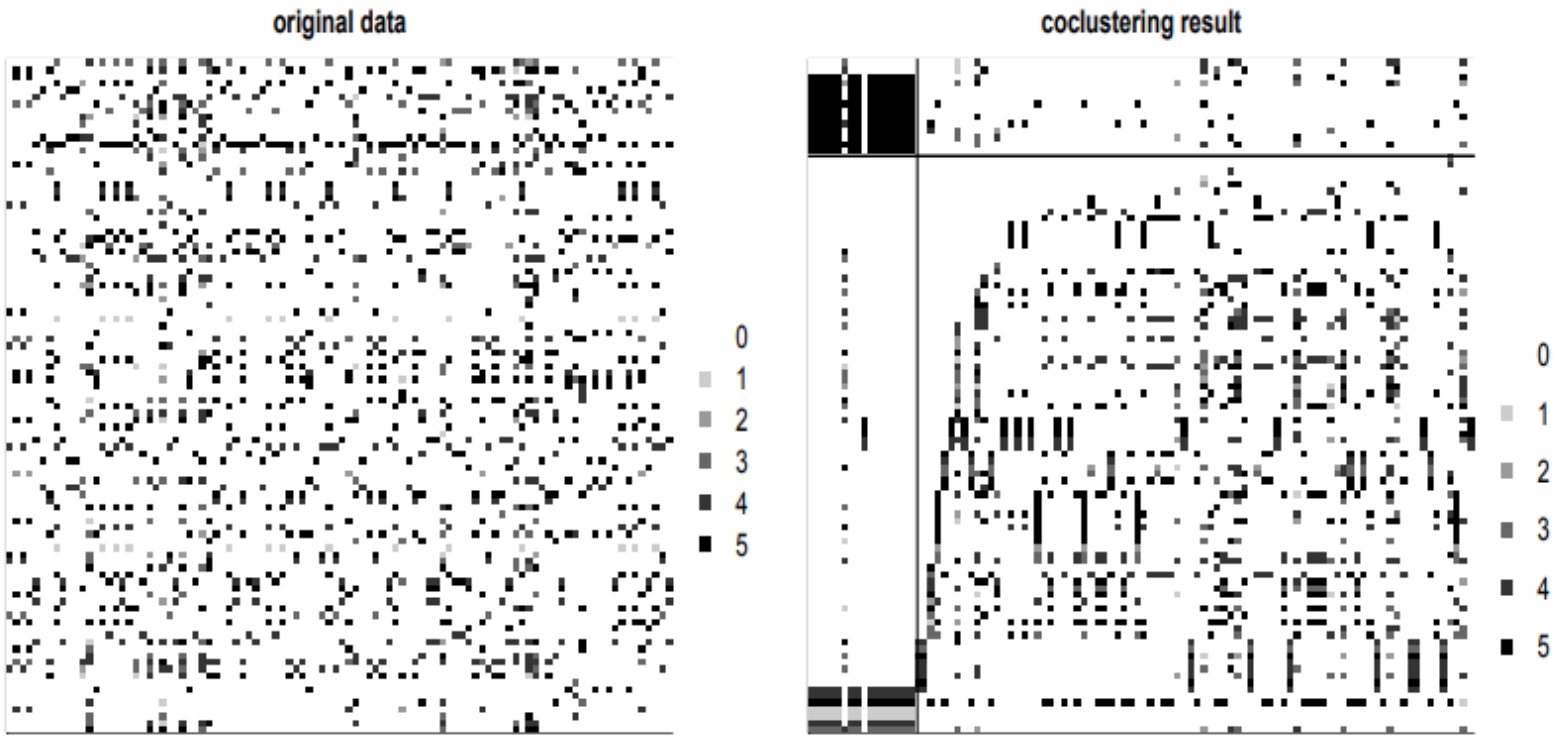
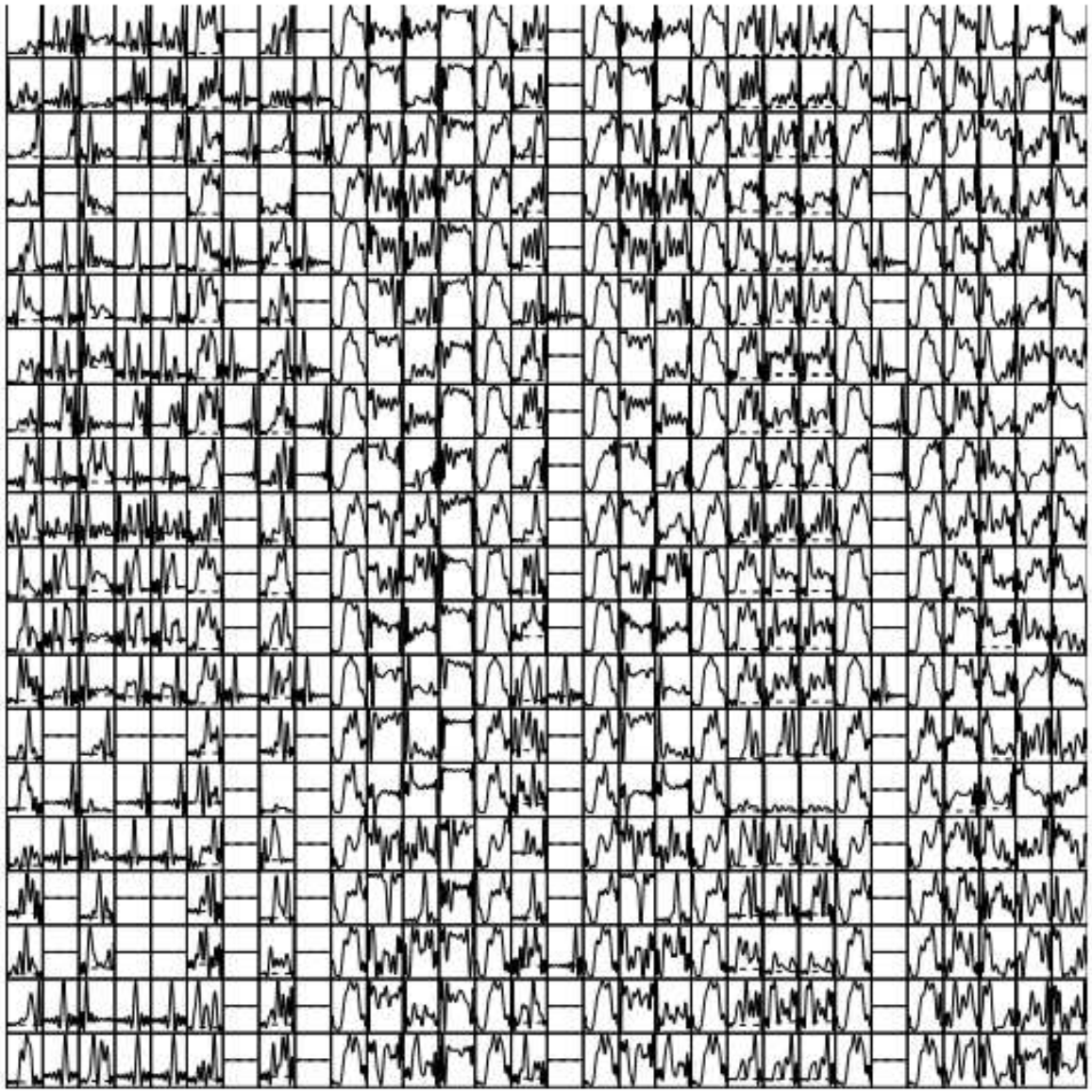


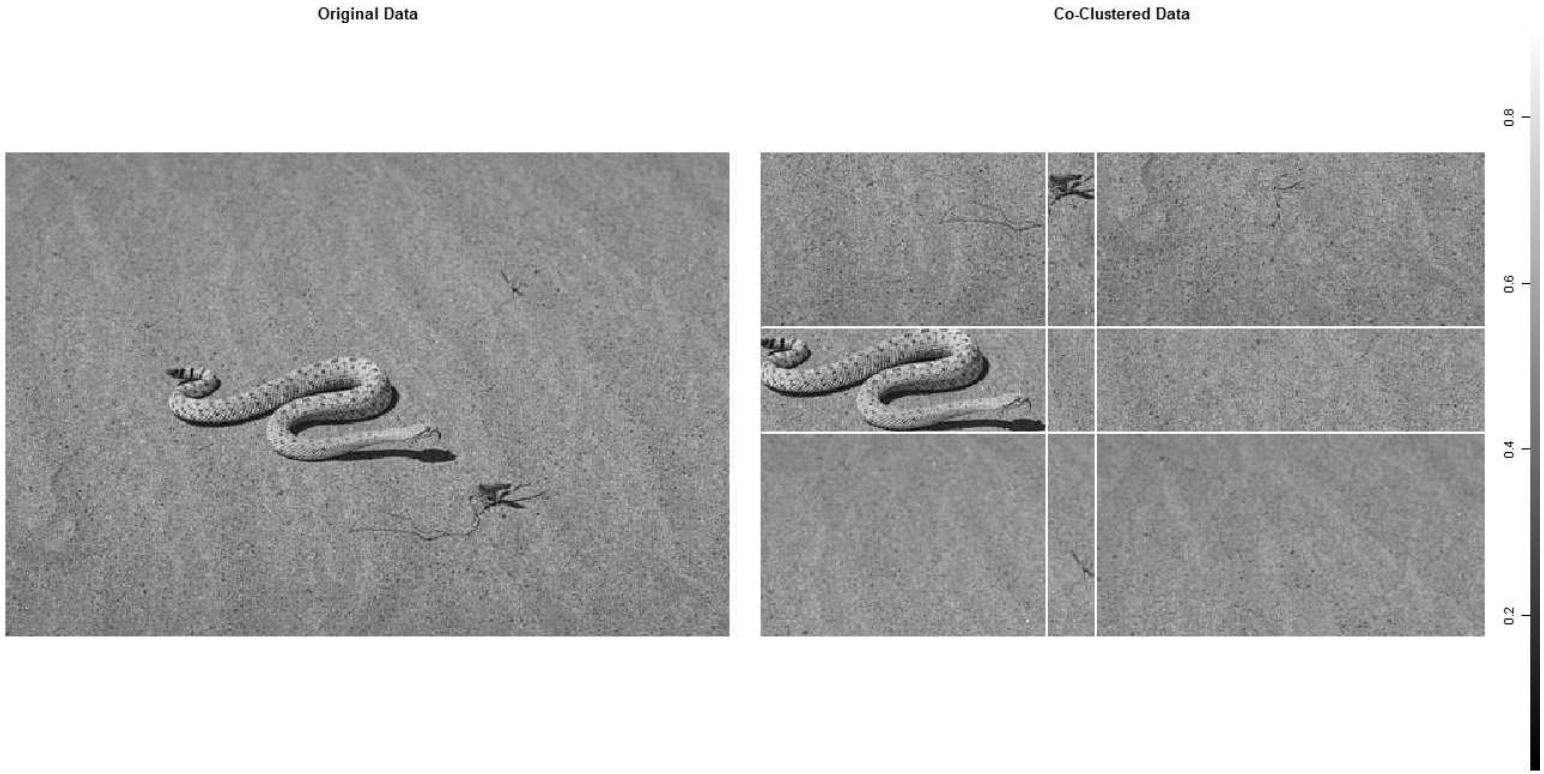
Figure 11: Top 100 Amazone Fine Food Review data (left) and co-clustering result (right).

Other kind of data: functional

[Jacques, 2016]

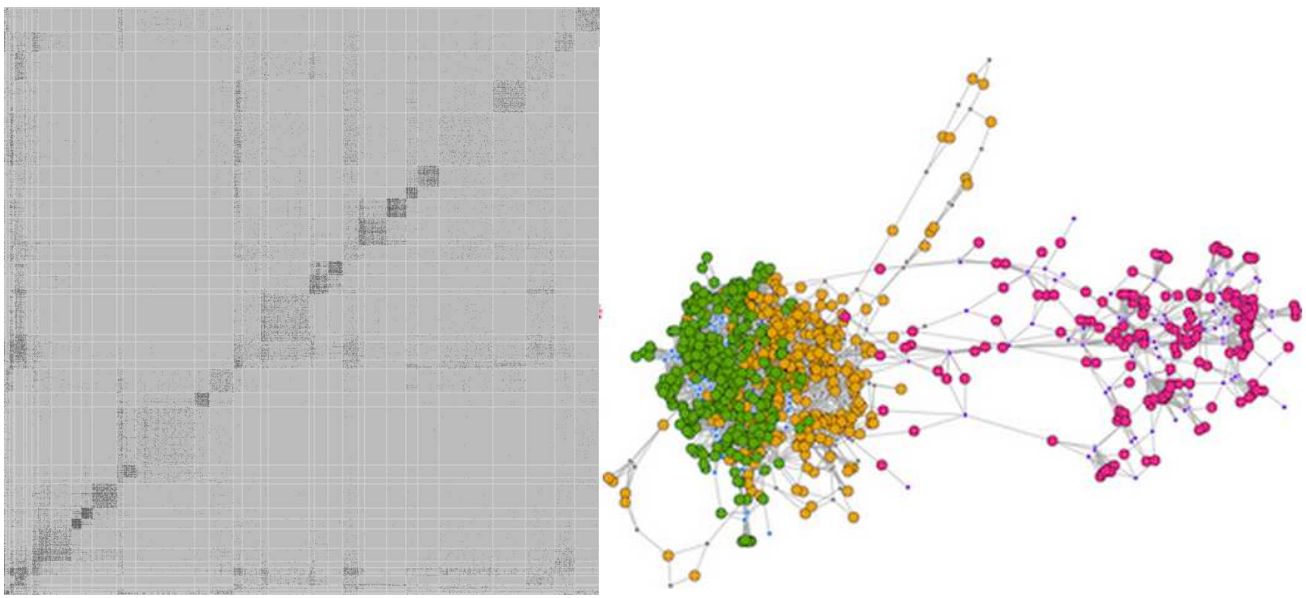


Other kind of data: image



Particular case: graph clustering

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



MLE estimation: log-likelihood(s)

- Similar to clustering: 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 earlier (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

Block estimation: estimate

- Once we have a parameter estimate $\hat{\boldsymbol{\theta}}$, we need to have a 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: consistency

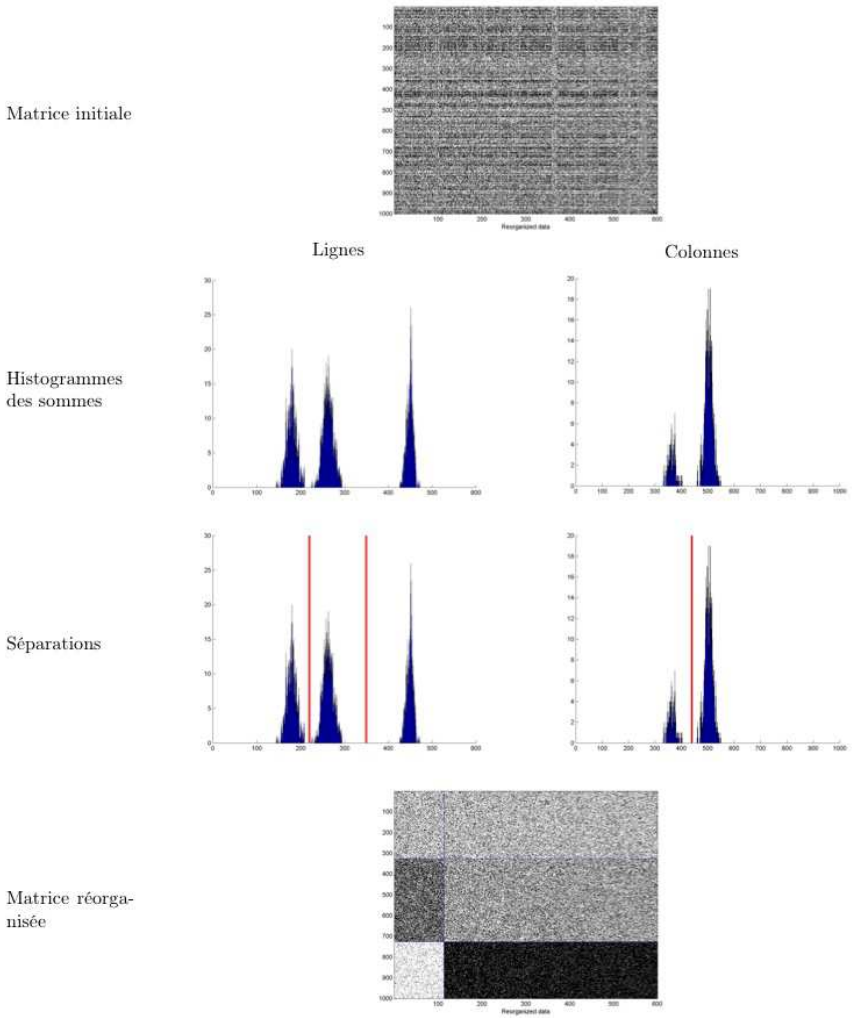
[Mariadassou and Matias, 12]

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

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 partition 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

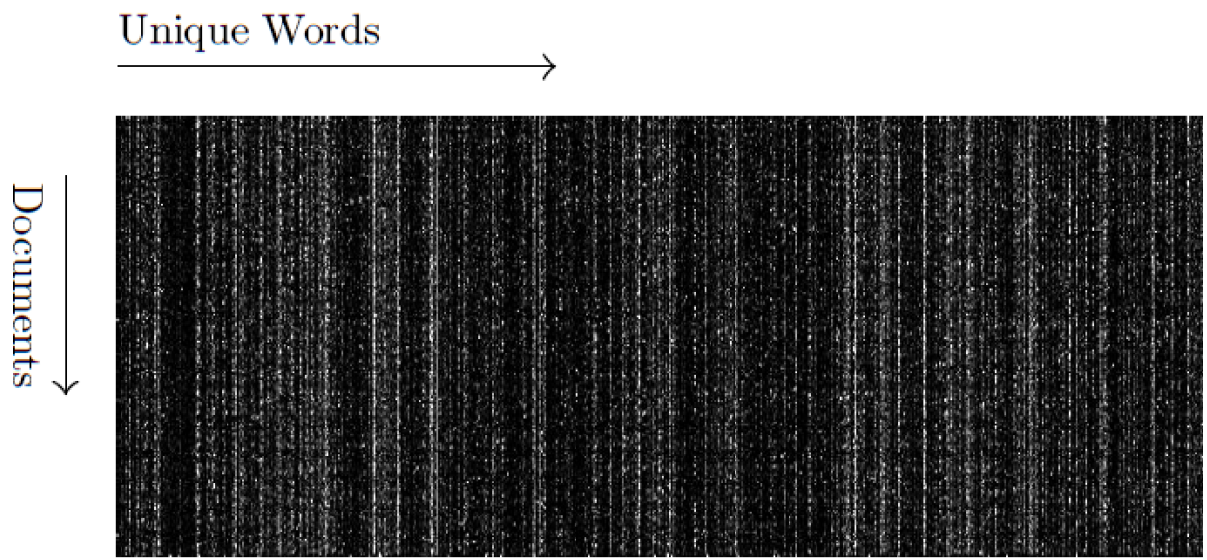
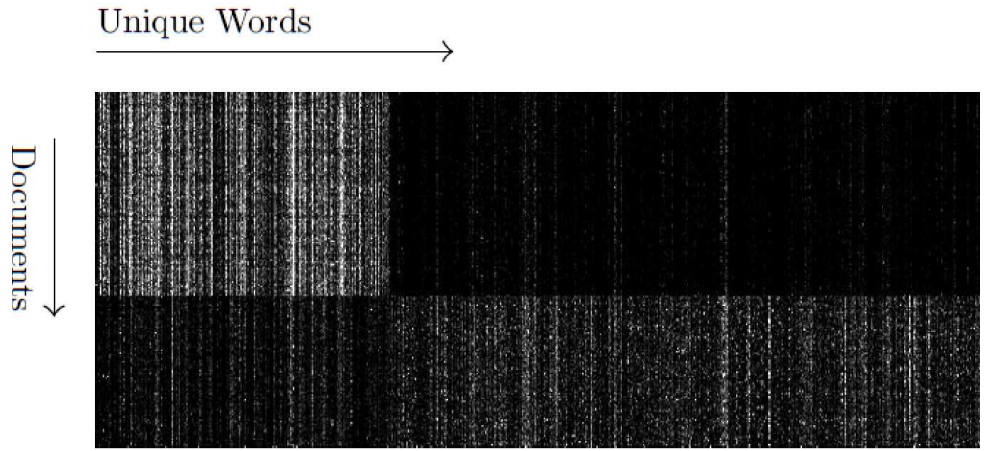


Illustration: document clustering (2/2)



Results with 2×2 blocks

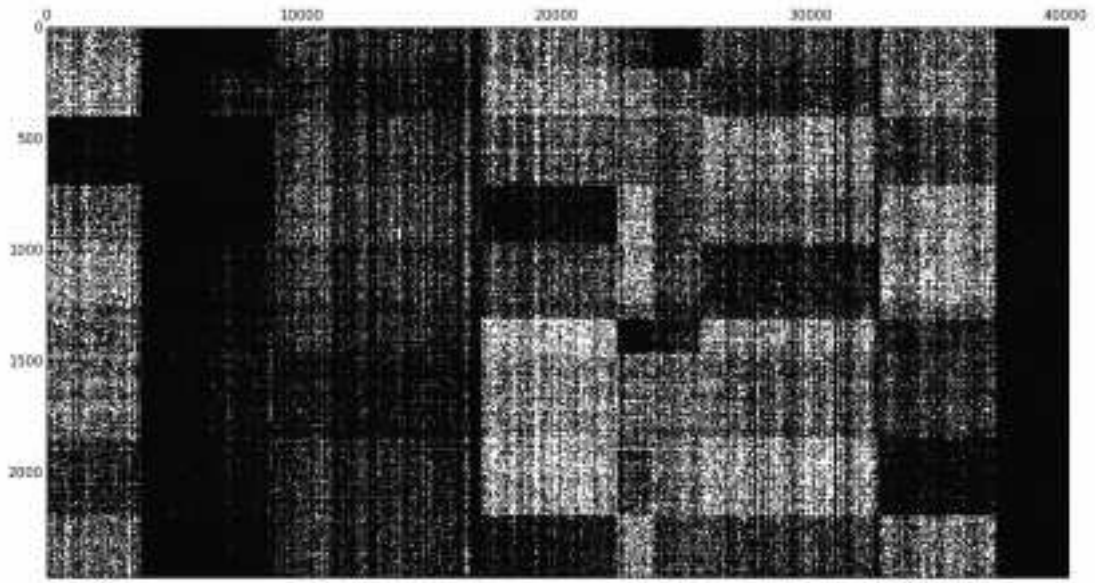
	Medline	Cranfield
Medline	1033	0
Cranfield	0	1398

Experiment illustrates previous theory: HD clustering is blessing

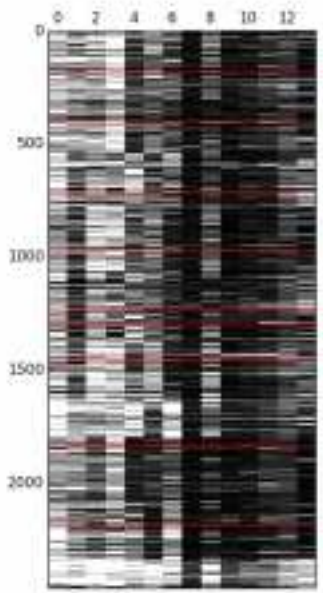
Outline

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