# MODEL BASED CO-CLUSTERING: HIGH DIMENSION AND ESTIMATION CHALLENGES

Christine Keribin, Christophe Biernacki, Julien Jacques

# MODEL BASED CO-CLUSTERING: HIGH DIMENSION AND ESTIMATION CHALLENGES

Christine KERIBIN [1]
Christophe BIERNACKI [2], Julien JACQUES [3]

[1] Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France. - Equipe CELESTE

[2] Université de Lille
INRIA - Lille - MODAL

[3] Université de Lyon, Lyon 2, ERIC UR 3083

Séminaire parisien de statistique, IHP, 11 mars 2024



# Outline

# Clustering

## Unsupervised ML framework

- data set $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ with $dim(\mathcal{X}) = d$
- to partition into $K$ clusters $G_1, \ldots, G_K$
- classification matrix $\mathbf{z} = (z_1, \ldots, z_n)$: $z_{ik} = \mathbb{1}_{i \in G_k}$

Model based clustering (MBC) reformulates cluster analysis in a well-posed estimation problem

- $x_i$ are iid observations of a mixture pdf

$$f(.;\theta) = \sum_{k=1}^{K} \pi_k \varphi(.;\alpha_k) \quad ; \sum_{i=1}^{K} \pi_K = 1, \pi_k \geq 0; \quad \mathbb{P}(z_{ik} = k) = \pi_k$$

- estimation with EM algorithm, partition is deduced from the conditional probability $p(z_i | x_i; \widehat{\theta})$

# HD settings: $d \sim n$, $d >> n$

More and more variables ($d$) and not so much observations ($n$)

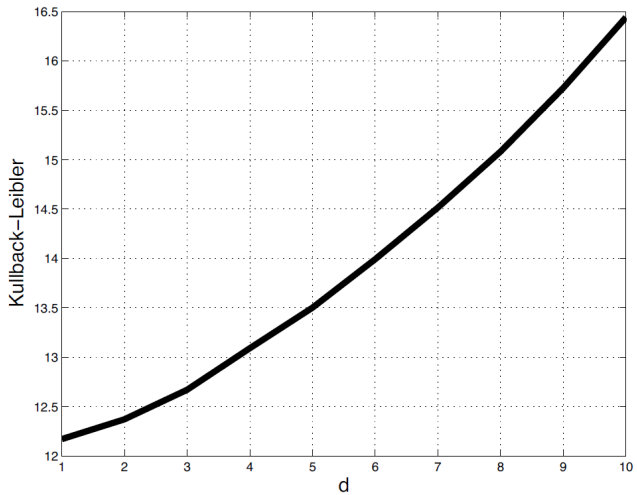
- marketing: $d \sim 10^2$ for $n = 10^4 - 10^6$
- microarray gene expression: $d \sim 10^2 - 10^4$ for $n = 20 - 50$ stress conditions
- fMRI images: $d \sim 10^4 - 10^5$ voxels for $n = 50 - 100$ patients
- textmining, curve, internet logs ...

Variables with mixed types

- different modes: qualitative, quantitative, functional
- different semantic

# Clustering in HD: curse (and blessing?)

- curse: density estimation quality decreases with $d$ for a same $n$



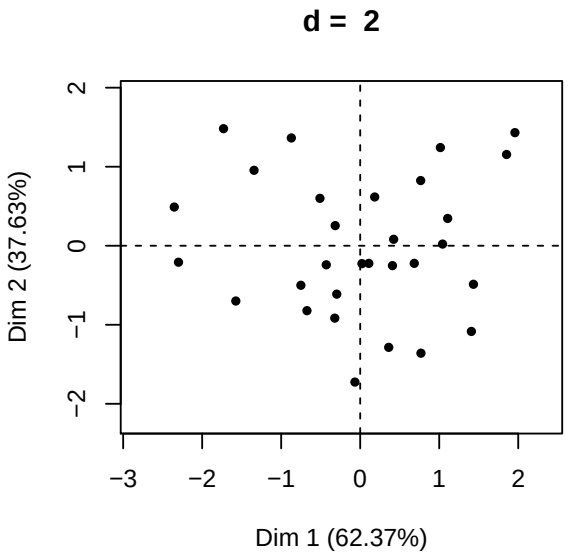

$Kullback(f, \widehat{f})$ for a two-component $d$- variate Gaussian mixture

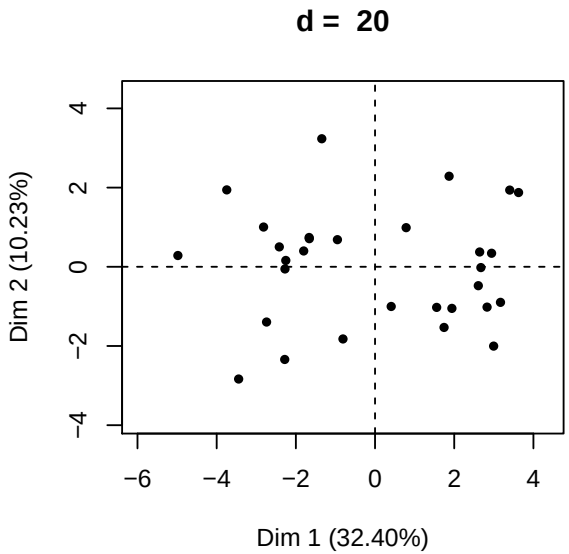
$$x_i|_{z_i} \sim \mathcal{N}_d(\mu_{z_i}, \Sigma); \pi_1 = \pi_2 = 0.5$$

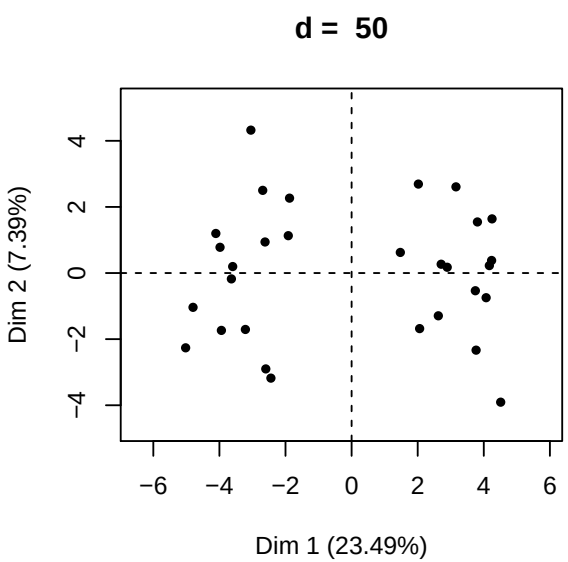
Basic case: $\mu_1 = 0_d; \mu_2 = 1_d; \Sigma = Id$

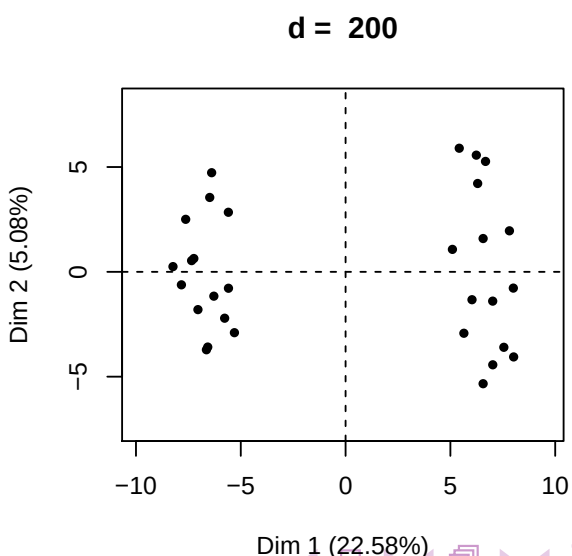
- blessing: but components could be more and more separated.
- Which impact on the misclassification error ?
  - what if the variables are correlated, redundant, non informative ?
  - what if $d >> n$ ?

# *Basic case*: HD is a blessing for clustering



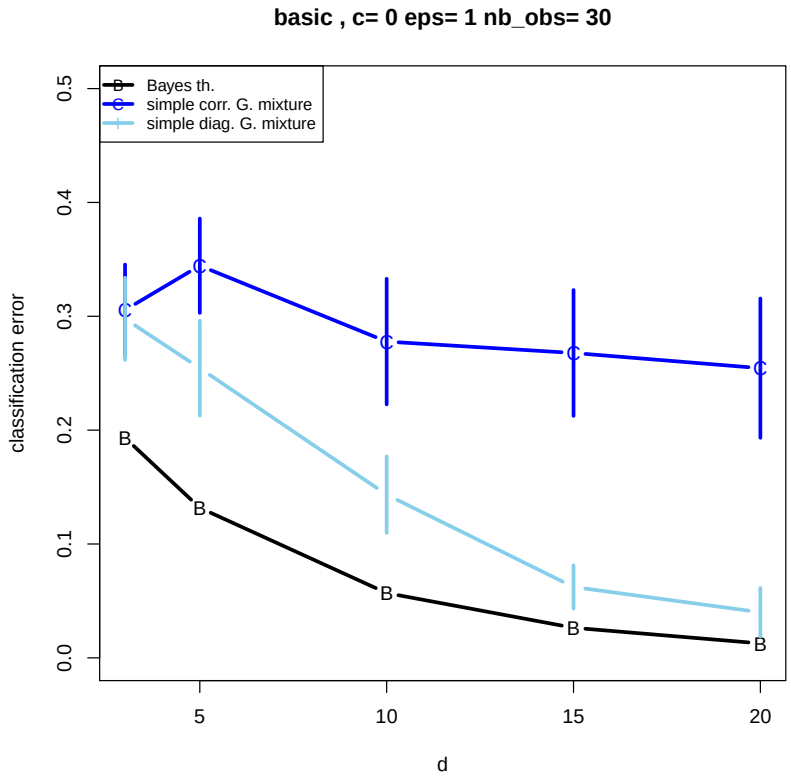



# *Basic case*: HD is a blessing for clustering

Each new variable is informative, sufficiently large $n \sim d$

- theoretical classification error decreases when $d$ grows: $err_{theo} = \Phi(-\sqrt{d}/2)$




A two-component $d$- variate Gaussian mixture

$$x_i|_{z_i} \sim \mathcal{N}_d(\mu_{z_i}, \Sigma);\ \pi_1 = \pi_2 = 0.5$$

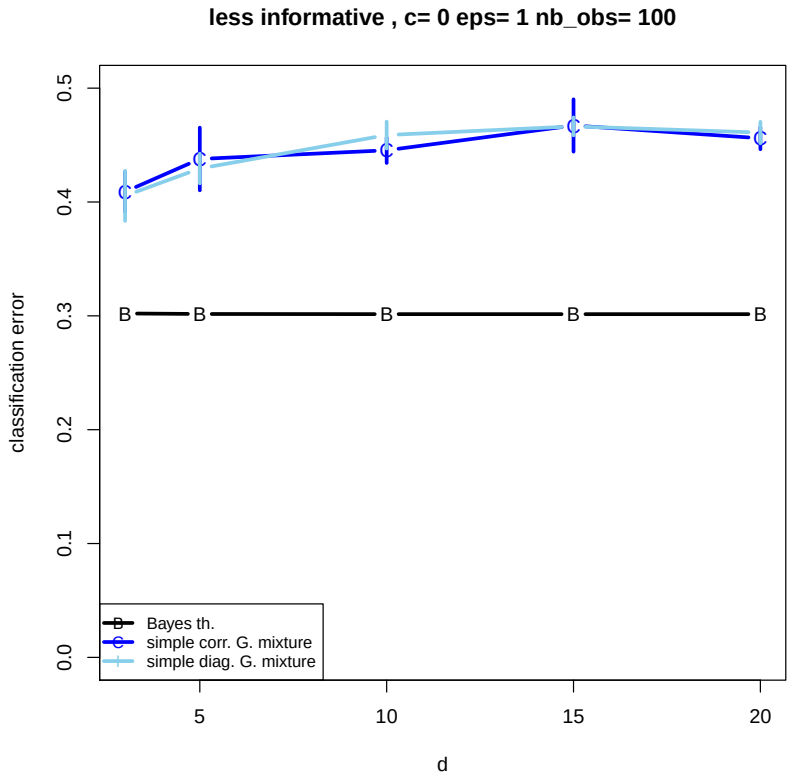
Basic case: $\Sigma = Id$; $\mu_1 = 0_d$; $\mu_2 = 1_d$

- and empirical error rate also decreases with $d$
- no model bias, classification ($\widehat{\mathbf{z}}$) variance decreases

# but HD can be less favorable for clustering

## (Many) less informative variables

- theoretical error slowly decreases when $d$ grows:
  $err_{theo} = \Phi(-||\mu_1 - \mu_2||/2)$




A two-component $d$- variate Gaussian mixture

$$x_i|_{z_i} \sim \mathcal{N}_d(\mu_{z_i}, \Sigma);\ \pi_1 = \pi_2 = 0.5$$

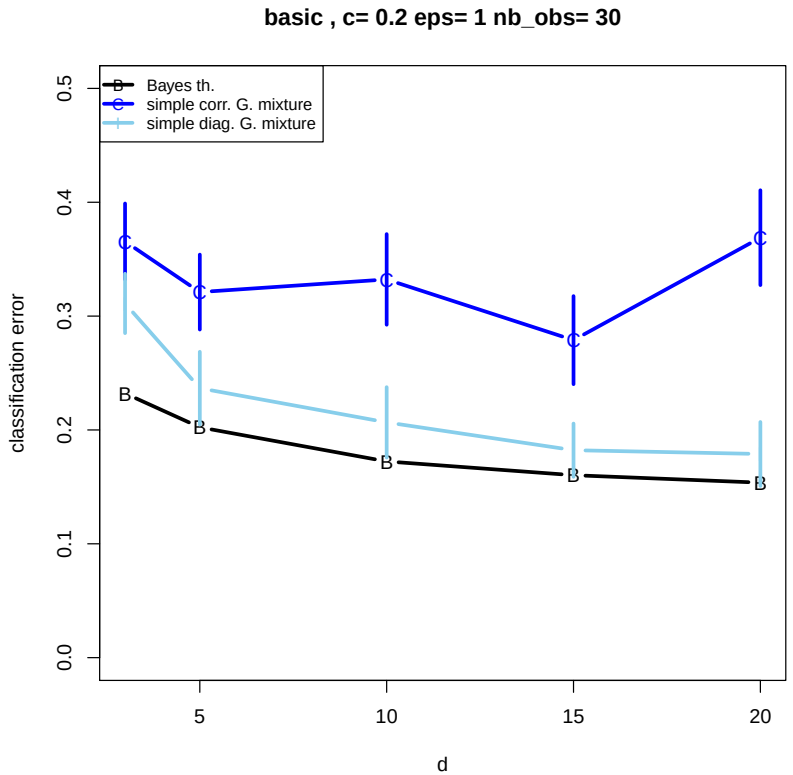
less informative: $\Sigma = Id;\ \mu_1 = 0_d$
$\mu_2 = (1, \frac{1}{2^2}, \ldots, \frac{1}{d^2})$

- but empirical error rate does not decrease any more...
- no model bias, classification ($\widehat{\mathbf{z}}$) variance increases

# Bias/variance trade-off for clustering

## Correlated variables

- theoretical error decreases $err_{theo} = \Phi(-||\mu_1 - \mu_2||_{\Sigma^{-1}}/2)$




A two-component $d$- variate Gaussian mixture

$$x_i|_{z_i} \sim \mathcal{N}_d(\mu_{z_i}, \Sigma);\ \pi_1 = \pi_2 = 0.5$$

Basic case+ correlated variables: $\Sigma = \Sigma(c);\ \mu_1 = 0_d;\ \mu_2 = 1_d$

- high error when clustered with the real (corr. G.) mixture model higher than with the biased (diag. G) mixture model, too much variance
- error decreases with the biased model which does not dammaged the separation

# Some alternatives for reducing variance

- Dimension reduction in non canonical space (PCA-like typically)
- Dimension reduction in the canonical space (variable selection)

  for Gaussian mixture, 3 variable-clusters (informative, non informative, linearly dependent) [*Maugis et al (2009)*] , for latent class analysis [*Fop et al (2018)*]
- Model parsimony in the initial HD space (constraints on model parameters)

  GMM [*Celeux and Govaert, '95*], mixture of factor analyzers [*[Ghahramani and Hinton, 97], [McLachlan et all], 03*], HD Gaussian models [*[Bouveyron et al, 07]*]

↪ trade-off between perfect modeling and what can be estimated in practice

# An alternative for reducing the variance

Main message:

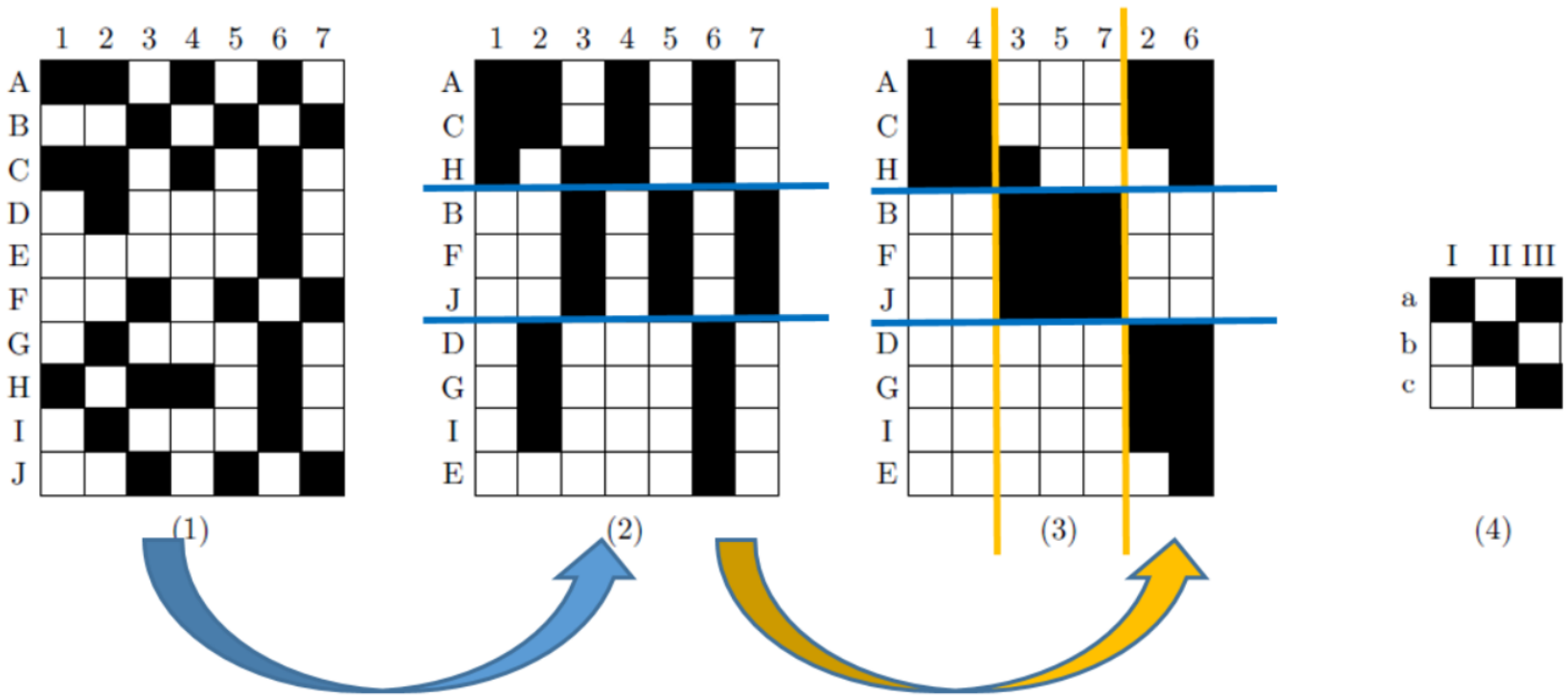
- clustering is a way for dealing with large $n$
  $\hookrightarrow$ why not also reusing this idea for large $d$?

## Co-clustering

- simultaneously clusters rows and columns
- performs parsimony of row clustering through variable clustering

$\hookrightarrow$ parsimonious, model-based, independent of data type, available

# From clustering to co-clustering




$(n = 10) \times (d = 7)$ $(K = 3) \times (d = 7)$ $(K = 3) \times (L = 3)$

[*Govaert (2008)*]

# Interpreting LBM as a MBC dimension reduction method[1]

- PCA for the $i$-th data individual $\mathbf{x}_i$ :

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

- specific CC case reduced to $K = 1$

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

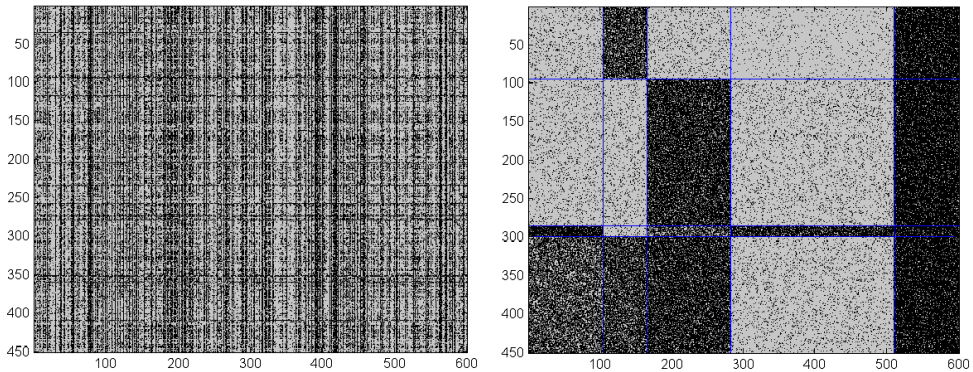
with $\tilde{w}_j = \ell \Leftrightarrow w_{j\ell} = 1$, $\varepsilon_i^j \sim \mathcal{N}(0, \sigma^2_{\tilde{w}_j})$, $\mathbf{v}^\ell = \sum_{\{j:\tilde{w}_j=\ell\}} \mathbf{e}^j$ and $\mathbf{r}_i = \sum_{j=1}^{d} \varepsilon_i^j \mathbf{e}^j$

- From sequential method *PCA then MBC* to simultaneous clustering with the dimension reduction process

[1] [*C. Biernacki, C. Keribin, J. Jacques, A survey on Model-Based Co-Clustering, JOC 2023*]

# Outline

# Unsupervised block clustering framework




- Co-clustering: to find a block clustering structure simultaneously on rows and columns

$H_1$: Blocks form a Cartesian product of a row-partition in $K$ row-groups by a column-partition in $L$ column-groups

$\mathbf{z} = (z_{ik})$ where $z_{ik} = 1$ if row $i$ belongs to row-group $k$

$\mathbf{w} = (w_{j\ell})$ where $w_{j\ell} = 1$ if column $j$ belongs to column-group $\ell$

- Application: huge data sets arising in recommendation systems (binary), genomic data analysis (Gaussian), text mining (Poisson),...
- In our case: dimension reduction for HD clustering

# Latent Block Model: a mixture model Govaert and Nadif (2008)

$H_2$: row and column labels are independently assigned :

$$z_i \overset{i.i.d}{\sim} \mathcal{M}(1, \boldsymbol{\pi}) \perp w_j \overset{i.i.d}{\sim} \mathcal{M}(1, \boldsymbol{\rho})$$

$H_3$: the $n \times d$ cells $x_{ij}$ are conditionally independent given $\mathbf{z}$ and $\mathbf{w}$ and follow the same conditional distribution $\varphi$ which parameter $\alpha$ only depends on the block:

$$x_{ij}|z_{ik} = 1, w_{j\ell} = 1 \sim \varphi(x_{ij}; \alpha_{k\ell})$$

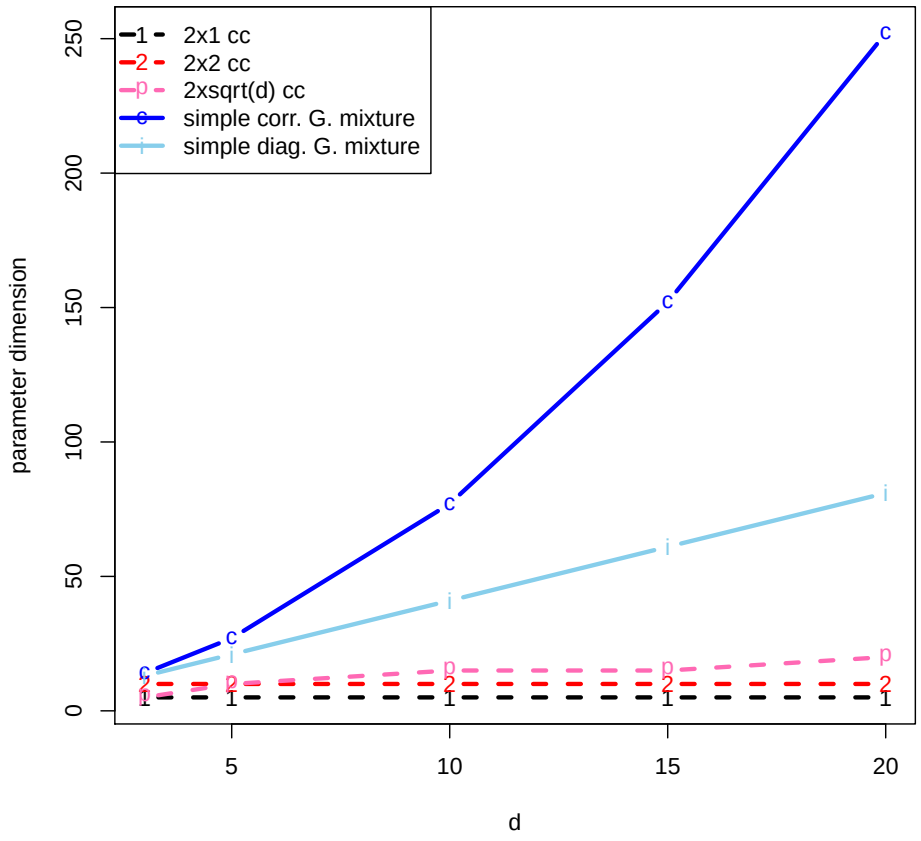
## Observed likelihood

$$p(x;\theta) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \underbrace{\prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}}}_{p(z,w;\theta)=p(z;\theta)p(w;\theta)} \underbrace{\prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}}_{p(x|z,w;\theta)}$$

with $\theta = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$

# Parametric model

- Very parsimonious model, good candidate to deal with HD settings.



# Many appealing extensions

- variable type diversity: ordinal data [*Biernacki and Jacques, '18*], functional data [*Bouveyron et al, 18*], mixed-type data [*Selosse et al, '20*], textual interaction data [*Bergé et al, '19*]
- dynamic, e.g. [*Marchello, '23*] (zero inflated dLBM)

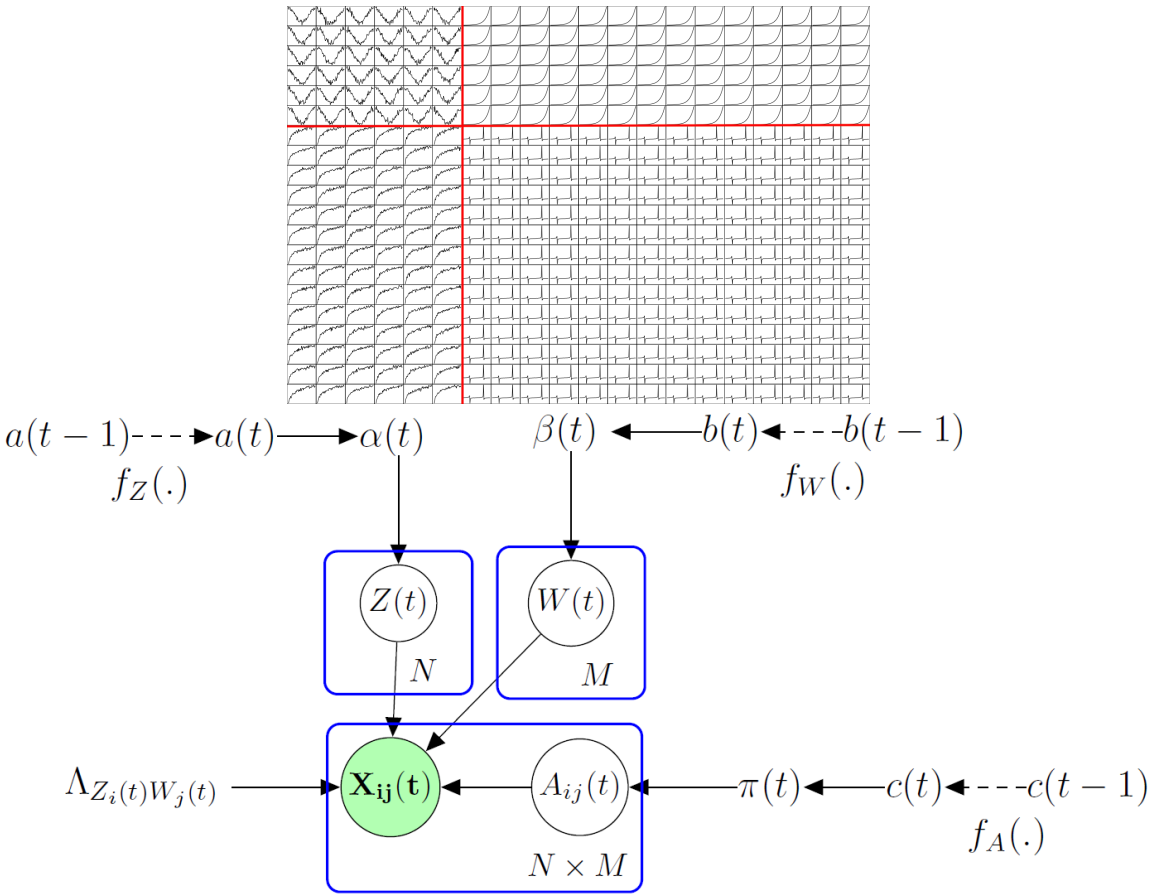
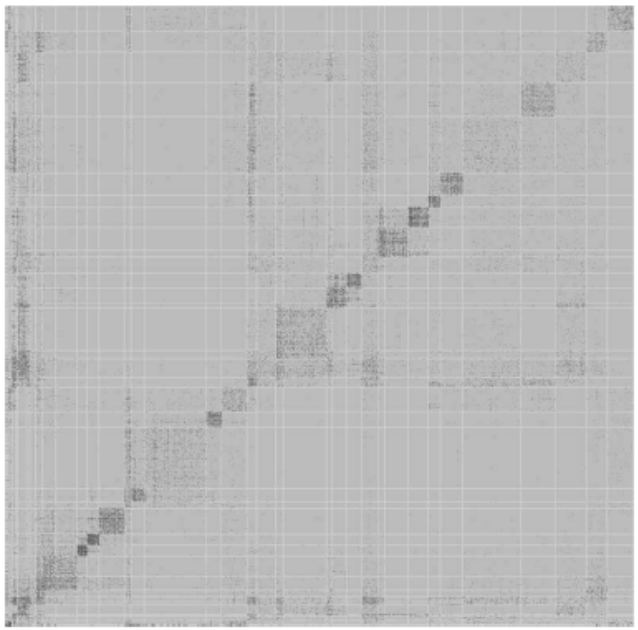
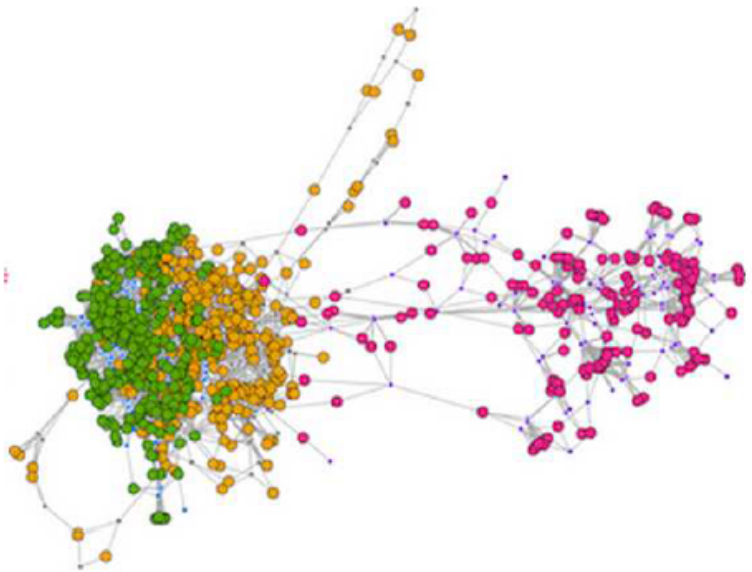



Figure 4.1: Graphical representation of the Zero-Inflated dLBM model.

# Particular case: graph clustering

- Stochastic Block Model (SBM): $x$ is the adjacency matrix of a graph
  $\hookrightarrow n = d$, $K = L$, $\mathbf{z} = \mathbf{w}$





# Estimation

$$p(x;\theta) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}$$

- generic identifiability [*K. et al, '15*]

- likelihood for $2 \times 2$ blocks and $20 \times 20$ matrix: $\approx 2^{20} \times 2^{20} \approx 10^{12}$ terms: intractable
- E-step: intractable
  ↪ Approximation with Variational EM [*Govaert and Nadif '08*]
  ↪ Estimation with SEM [*K. et al '10*]
  ↪ Stochastic gradient descent [*Baey et al. '23*]
  ↪ Bayesian versions: Gibbs and V-Bayes algorithms [*K., Brault, Celeux, Govaert, '15*]

- Consistency and asymptotic normality of ML- and V- estimators (when $\log(n)/d \to +\infty$ and $\log(d)/n \to +\infty$)[*Brault, K., Mariadassou, '20*]
- model selection (ICL)

# EM for LBM

$$\log p(\mathbf{x};\theta) = \underbrace{\mathbb{E}[\log p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}^{(c)}]}_{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(c)})} - \mathbb{E}[\log p(\mathbf{z},\mathbf{w}|\mathbf{x};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}^{(c)}]$$

## Expectation Maximization (iteration $c$)

- **E-step:** computing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(c)})$

$$Q(\theta|\theta^{(c)}) = \sum_{i,k} \underbrace{\mathbb{P}(z_{ik}=1|\theta^{(c)},\mathbf{x})}_{s_{ik}^{(c)}} \log \pi_k + \sum_{j,\ell} \underbrace{\mathbb{P}(w_{j\ell}|\theta^{(c)},\mathbf{x})}_{t_{j\ell}^{(c)}} \log \rho_\ell$$
$$+ \sum_{i,j,k,\ell} \underbrace{\mathbb{P}(z_{ik}w_{j\ell}=1|\theta^{(c)},\mathbf{x})}_{e_{i,j,k,\ell}^{(c)}} \log \varphi(x_{ij};\alpha_{k\ell})$$

  $\hookrightarrow$ $s_{ik}^{(c)}$, $t_{j\ell}^{(c)}$, $e_{i,j,k,\ell}^{(c)}$ are intractable

- **M-step:** maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(c)})$ in $\boldsymbol{\theta}$, classical

  $\hookrightarrow$ $\theta^{(c+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(c)})$

# VEM for LBM

$$\log p(\mathbf{x};\theta) = \underbrace{\mathbb{E}_{q_z q_w}[\log \frac{p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})}{q_z q_w}|\mathbf{x}]}_{\mathcal{F}(\boldsymbol{\theta};q_z,q_w)} + KL(q_z q_w, \log p(\mathbf{z},\mathbf{w}|\mathbf{x};\boldsymbol{\theta}))]$$

## VEM (iteration $c$)

- approximated **E-step:**

$$q_z^{(c)}, q_w^{(c)} = \arg\max_{q_z,q_w} \mathcal{F}(\boldsymbol{\theta}; q_z, q_w)$$

$$\hookrightarrow s_{ik}^{(c)} \approx q_{zik}^{(c)},\ t_{j\ell}^{(c)} \approx q_{wj\ell}^{(c)},\ e_{ijk\ell}^{(c)} \approx q_{zik}^{(c)} q_{wj\ell}^{(c)}$$

$$Q(\theta|\theta^{(c)}) \approx \sum_{i,k} q_{zik}^{(c)} \log \pi_k + \sum_{j,\ell} q_{wj\ell}^{(c)} \log \rho_\ell + \sum_{i,j,k,\ell} q_{zik}^{(c)} q_{wj\ell}^{(c)} \log \varphi(x_{ij};\alpha_{k\ell})$$

- **M-step:** $\theta^{(c+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\theta}; q_z^{(c)}, q_w^{(c)})$ in $\boldsymbol{\theta}$, classical

$$\max_{\boldsymbol{\theta}} \log p(\mathbf{x};\theta) \geq \max_{\boldsymbol{\theta},q_z,q_w} \mathcal{F}(\boldsymbol{\theta}; q_z, q_w)$$

# SEM for LBM

$$\log p(\mathbf{x};\theta) = \underbrace{\mathbb{E}[\log p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}^{(c)}]}_{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(c)})} - \mathbb{E}[\log p(\mathbf{z},\mathbf{w}|\mathbf{x};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}^{(c)}]$$

## SEM-Gibbs (iteration $c$)

- **step SE:** impute missing values $(\mathbf{z},\mathbf{w}) \sim p(\mathbf{z},\mathbf{w}|\mathbf{x};\boldsymbol{\theta}^{(c)})$ *[Celeux and Diebolt (1986)]* with Gibbs-sampling
  - $\mathbf{z}^{(t+1)} \sim p(\mathbf{z}|\mathbf{w}^{(t)},\mathbf{x};\boldsymbol{\theta}^{(c)})$
  - $\mathbf{w}^{(t+1)} \sim p(\mathbf{w}|\mathbf{z}^{(t+1)},\mathbf{x};\boldsymbol{\theta}^{(c)})$

  $\hookrightarrow (\mathbf{z}^{(c)},\mathbf{w}^{(c)})$
- **step M:** maximize $\log p(\mathbf{x},\mathbf{z}^{(c)},\mathbf{w}^{(c)};\boldsymbol{\theta})$ in $\theta$

$\hookrightarrow$ stochastic, result fluctuates around the optimal value

# Stochastic gradient descent for latent variables [2]

- Fisher Identity [*Cappé et al. '05*]: $\nabla_\theta \log p(x;\theta) = \mathbb{E}(\nabla_\theta \log p(\mathbf{x},\mathbf{z}|\mathbf{x};\theta))$

## SGD-Gibbs for LBM (iteration $c$)

- **step S:** Gibbs-sampling: $(\mathbf{z}^{(c)},\mathbf{w}^{(c)}) \sim p(.,.|\mathbf{x};\theta^{(c)})$
- gradient for observations $i = 1,\ldots,n; j = 1,\ldots,d$

$$J_{ij}^{(c),obs} = \nabla_\theta \log p(x_{ij}|z_i^{(c)}, w_j^{(c)};\theta^{(c)}); \Delta_{ij}^{(c),obs} = (1-\gamma^{(c)})\Delta_{ij}^{(c),obs} + \gamma^{(c)} J_{ij}^{(c),obs}$$

- gradient for latent variables $i = 1,\ldots,n; j = 1,\ldots,d$

$$J_i^{(c),z} = \nabla_\theta \log p(z_i^{(c)};\theta^{(c)});\ \Delta_i^{(c),z} = (1-\gamma^{(c)})\Delta_i^{(c-1),z} + \gamma^{(c)} J_i^{(c),z}$$

$$J_j^{(c),w} = \nabla_\theta \log p(w_j^{(c)};\theta^{(c)});\ \Delta_j^{(c),w} = (1-\gamma^{(c)})\Delta_j^{(c-1),z} + \gamma^{(c)} J_j^{(c),w}$$

- $\theta^{(c+1)} = \theta^{(c)} + \gamma^{(k)} P_k v_k$

[2] [*Baey, Delattre, Kuhn, Leger, Lemler '23*]

# Outline

# Estimation challenges

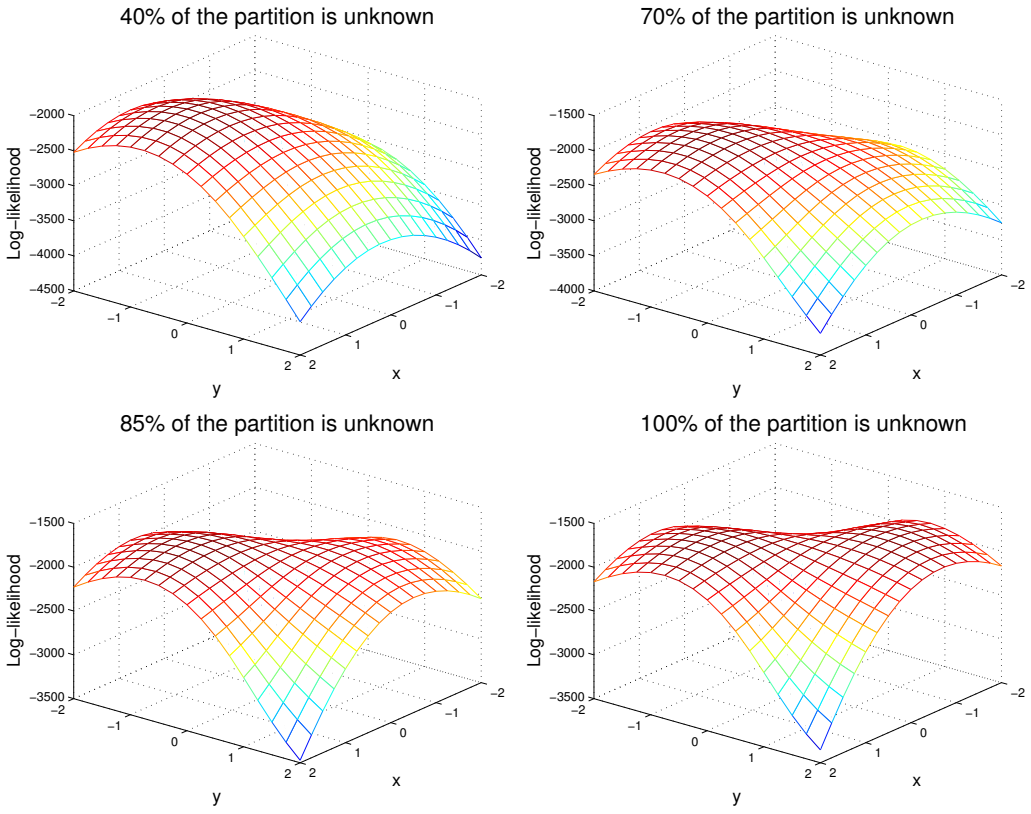
in model based clustering context:

- "classical" local maxima
- empty cluster solutions
- degenerate solutions

↪ worse for LBM ?

# Classical local maxima

- linked to the number of latent variables




- additional order of magnitude: from $K^n$ to $K^n L^d$ $\hookrightarrow$ pb for LBM? but no way to draw a likelihood map

# Empty clusters in LBM

Really frequent (and reported) !

- a component such that $\pi_k^{(c)} \simeq 0$
- number of empty clusters in MBC: $\#\mathcal{Z}^0$
- number of empty clusters in LBM : $\#(\mathcal{Z} \times \mathcal{W})^0$

$$\frac{\#(\mathcal{Z} \times \mathcal{W})^0}{\#\mathcal{Z}^0} = \frac{K^n L^d - S(n,K)S(d,L)K!L!}{K^n - S(n,K)K!} \to \infty \text{ when } n \to \infty \text{ or } d \to \infty$$

$S(n, K)$ is the Stirling number of second kind

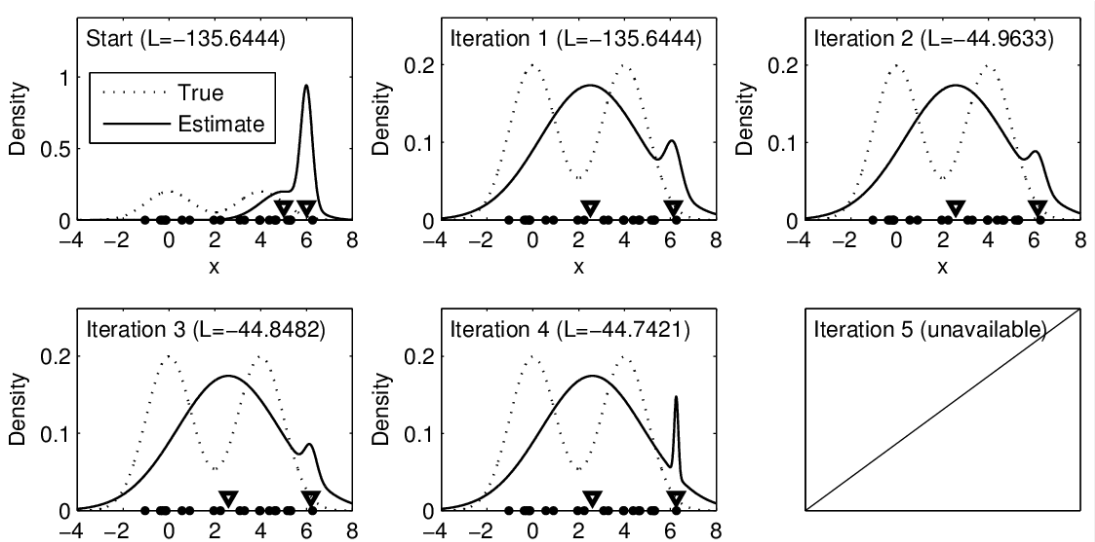
- exponential speed even with very small sample sizes

$$\frac{\#(\mathcal{Z} \times \mathcal{W})^0}{\#\mathcal{Z}^0} = \begin{cases} 62 & \text{when } n = d = 5 \text{ and } K = L = 2, \\ 710\,768 & \text{when } n = d = 9 \text{ and } K = L = 4. \end{cases}$$

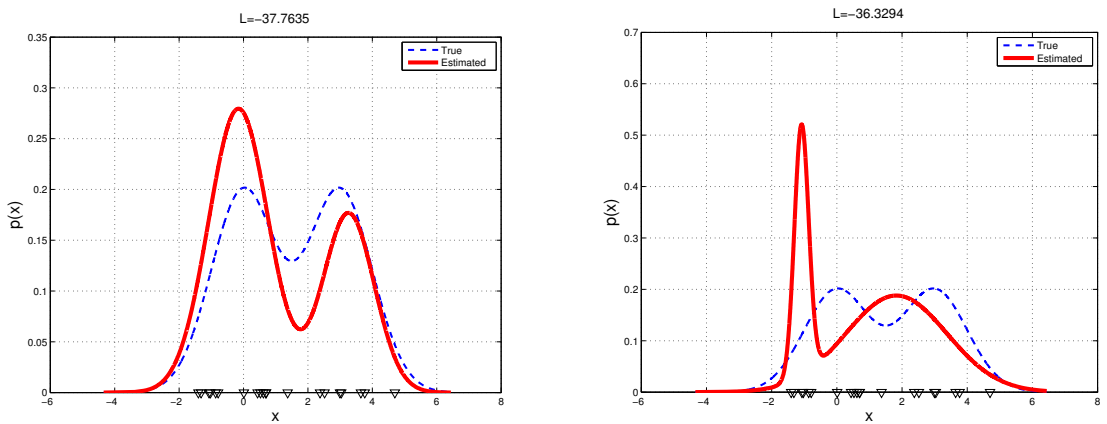
$\hookrightarrow$ initialization often achieved by performing independent clustering of the rows and columns

# Degenerate and spurious solutions in MBC

- Degenerate solutions lay on the border parameter space




- spurious local maximizers due to one or several relatively small (but not null) generalized variance



# Degenerate and spurious solutions in LBM

no reported such problems ?

- frequency of obtaining a cluster with a single element in the whole partition latent space

$$\frac{\#\mathcal{Z}_1^1}{\#\mathcal{Z}} = \left(\frac{K-1}{K}\right)^{n-1}.$$

- frequency of obtaining a block with a single element in the whole partition latent space

$$\frac{\#(\mathcal{Z}\times\mathcal{W})_{1,1}^1}{\#(\mathcal{Z}\times\mathcal{W})} = \frac{\#\mathcal{Z}_1^1}{\#\mathcal{Z}}\left(\frac{L-1}{L}\right)^{d-1}.$$

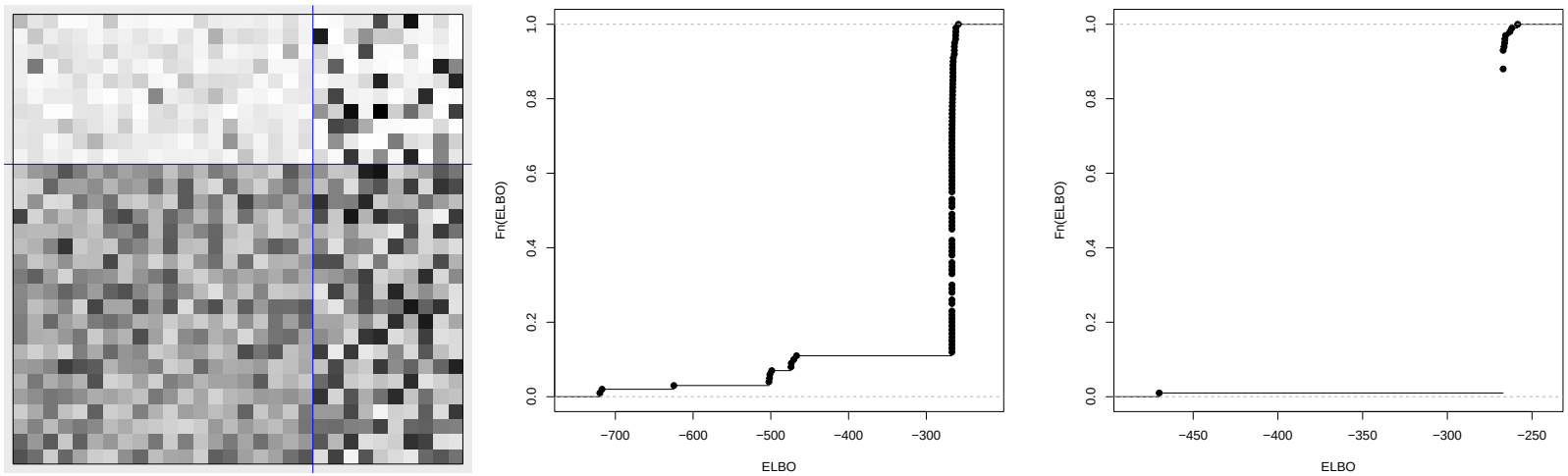
  $\hookrightarrow$ degeneracy situations are expected to be significantly less present in co-clustering than in clustering

  Ex with $L = 4$ and $d = 50$: $(1 - 1/L)^{d-1} = 7.5510^{-7}$

- Spurious case: same combinatorial arguments (but more tedious computations)

# Local maxima in Gaussian LBM: illustration[3]

Figure: *Left*: Gaussian $2 \times 2$ LBM ; e.c.d.f. of ELBO values obtained from $B = 100$ initializations, standard precision (*center*) and high precision (*right*)

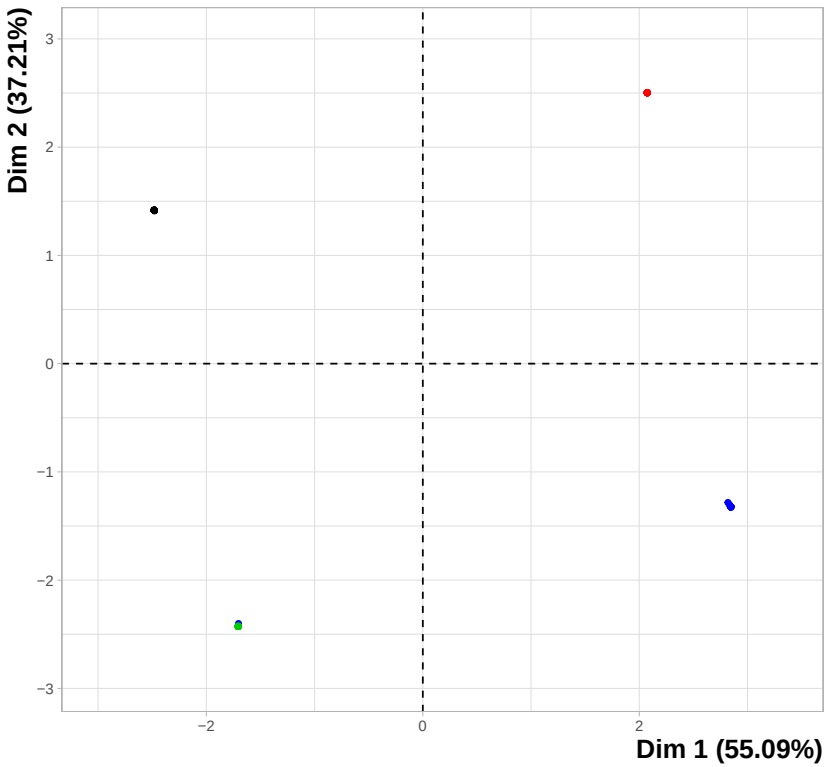


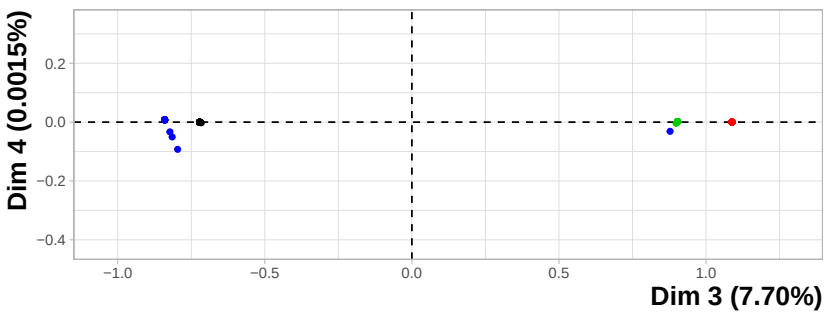

- some potential slow convergence effect
- different parameters but the same partition
  ↪ clustering is easier than estimation, no need to estimate with too much precision

[3] R package `blockcluster`

# Local maxima in Gaussian LBM: illustration (cont'd)

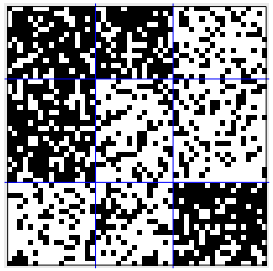
Projection of the solutions on the PCA planes



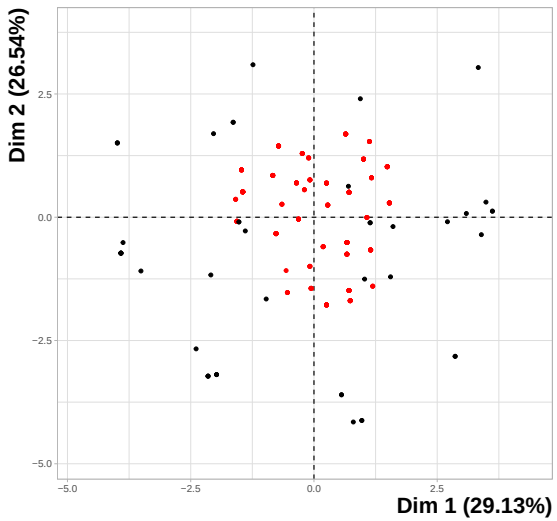




- `blockcluster` initialization performs several initial small VEM steps, that can explain these results

# Local maxima in binary LBM: illustration[4]



| | | | |
|---|---|---|---|
| ELBO | $-1615.48$ | $-1615.43$ | $-1596.67$ |
| count | 9 | 38 | 254 |
| nb steps | $> 10^4$ | $> 10^4$ | $< 100$ |




- high ELBO values (red) : 36 relabelled configurations
- low ELBO values (black) : 30 co-clustering configurations, all with an empty row or cluster
  $\hookrightarrow$ very lazy convergence

[4] R package `bikm1`

# Initilization strategies: crucial !

- two independent clustering (row / column)
- using several initializations with other algorithms (CEM, SEM, EM-VBayes) on few iterations
- resampling in order to avoid getting empty blocks during the first iterations
- `blockmodels`[*Leger, '16*] smart and robust initialization: combining spectral clustering adapted for LBM and a forward/backward reinitialization strategy:
  - forward exploration of the space of models (K, L) by splitting already existing clusters [*Robert '17*]
  - backward exploration by merging groups

# Outline

# Properties of LBM in view of HD clustering

- parcimonious model
- the true partition is recovered when there is a consistent estimator

  [*Mariadassou and Matias, '15*]

$$\hat{\boldsymbol{\theta}} \overset{n,d\to\infty}{\longrightarrow} \boldsymbol{\theta}^* \quad \Rightarrow \quad p(\hat{\mathbf{z}} = \mathbf{z}^*, \hat{\mathbf{w}} = \mathbf{w}^* | \mathbf{x}; \hat{\boldsymbol{\theta}}) \overset{n,d\to\infty}{\longrightarrow} 1,$$

↪ naturally regularized candidate for HD clustering

# Numerical illustrations

Generating models, with two balanced ($\pi_1 = \pi_2$) row clusters in $\mathbb{R}^d$.

| | $\mathbf{I}_d$ | $\mathbf{\Sigma}_d(c)$ |
|---|---|---|
| $\boldsymbol{\mu}_1 = \mathbf{0}_d$, $\boldsymbol{\mu}_2 = \mathbf{1}_d$ | $(M_1)$ | $(M_2)$ |
| $\boldsymbol{\mu}_1 = \mathbf{0}_d$, $\boldsymbol{\mu}_2 = (1, 2^{-2}, \ldots, d^{-2})$ | $(M_3)$ | $(M_6)$ |
| $(M_1)$ of size $n \times d/2$ duplicated twice | $(M_4)$ | |
| $\boldsymbol{\mu}_1 \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$, $\boldsymbol{\mu}_2 \sim \mathcal{N}_d((\underbrace{1, \ldots, 1}_{\sqrt{d}}, 0, \ldots, 0), \mathbf{I}_d)$ | $(M_5)$ | |

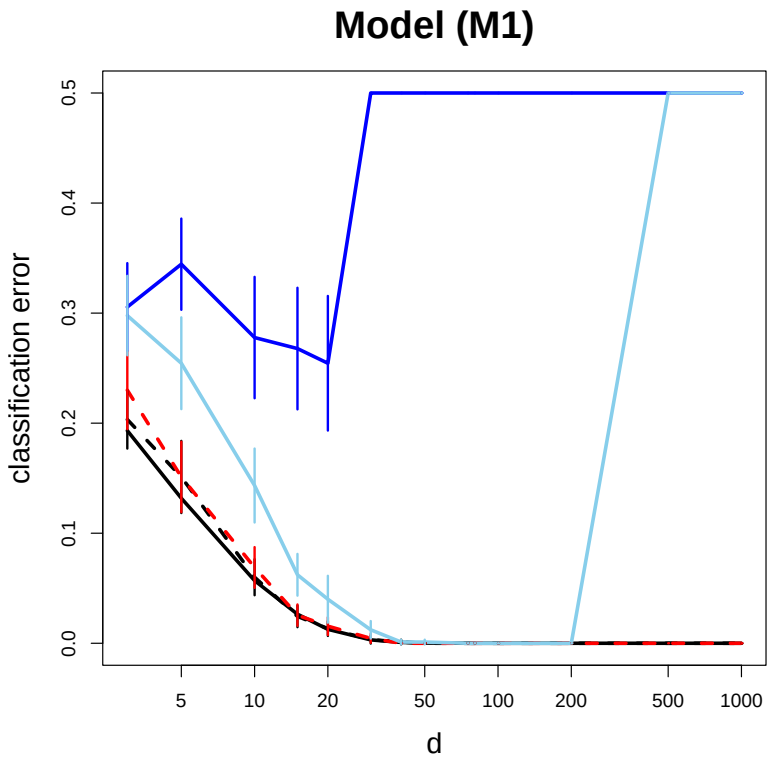
$\hookrightarrow$ only $(M_1)$ is a nomimal LBM (2x1)

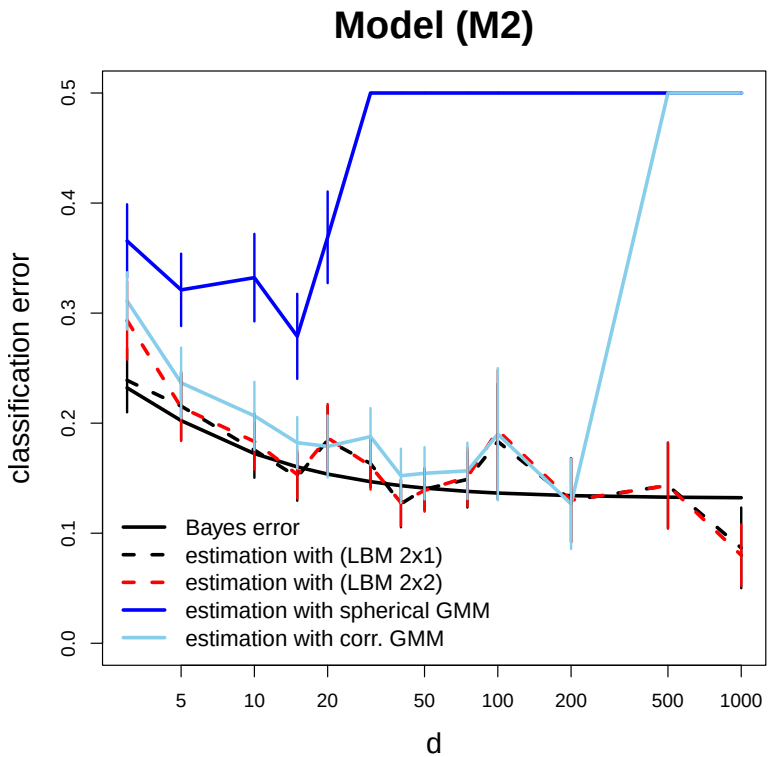
Row clustering is performed using four different methods

- clustering with a mixture of two spherical $d$-dim. Gaussian distributions,
- clustering with a mixture of two full-covariance $d$-dim. Gaussian distr.,
- co-clustering with a Gaussian $(2 \times 1)$ LBM
- co-clustering with a Gaussian $(2 \times 2)$ LBM

Classification error averaged over 30 samples of size $n = 30$

# Numerical illustration

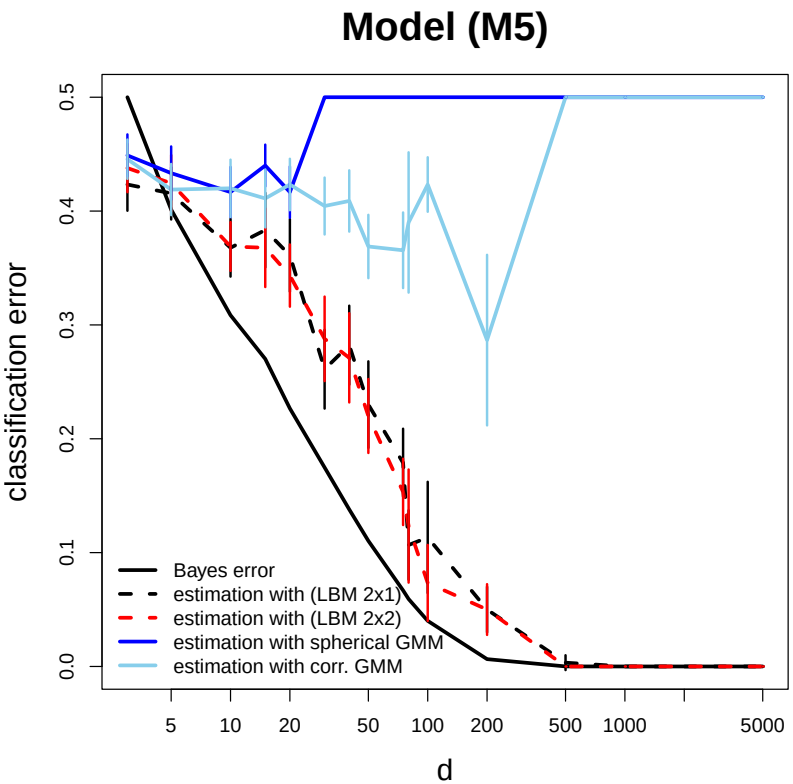


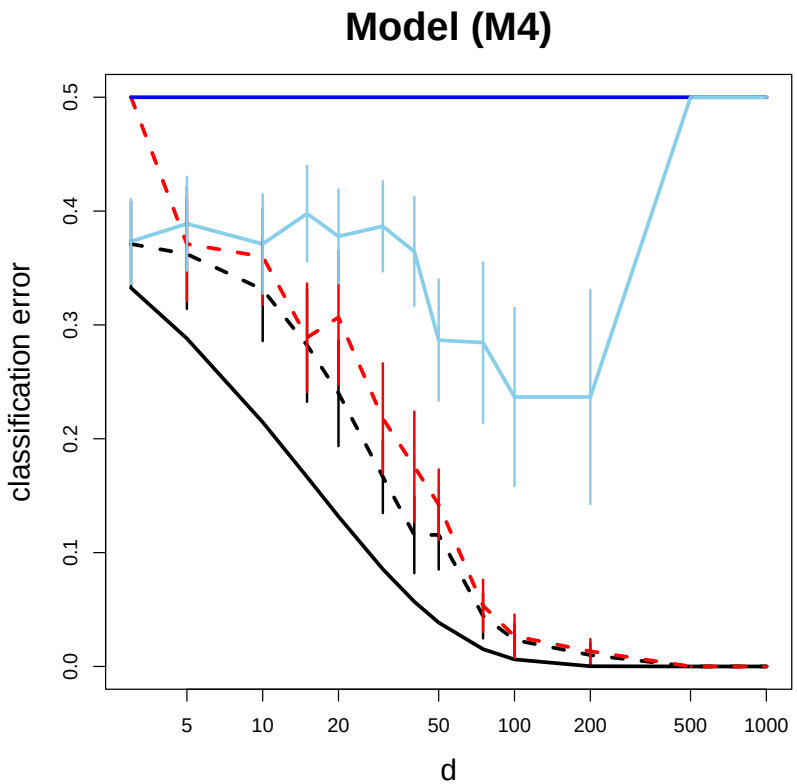




Resists to HD

Performs better than simple mixture, although the underlying model overrides the co-clustering conditional independence assumptions

# Numerical illustration




non info. variables

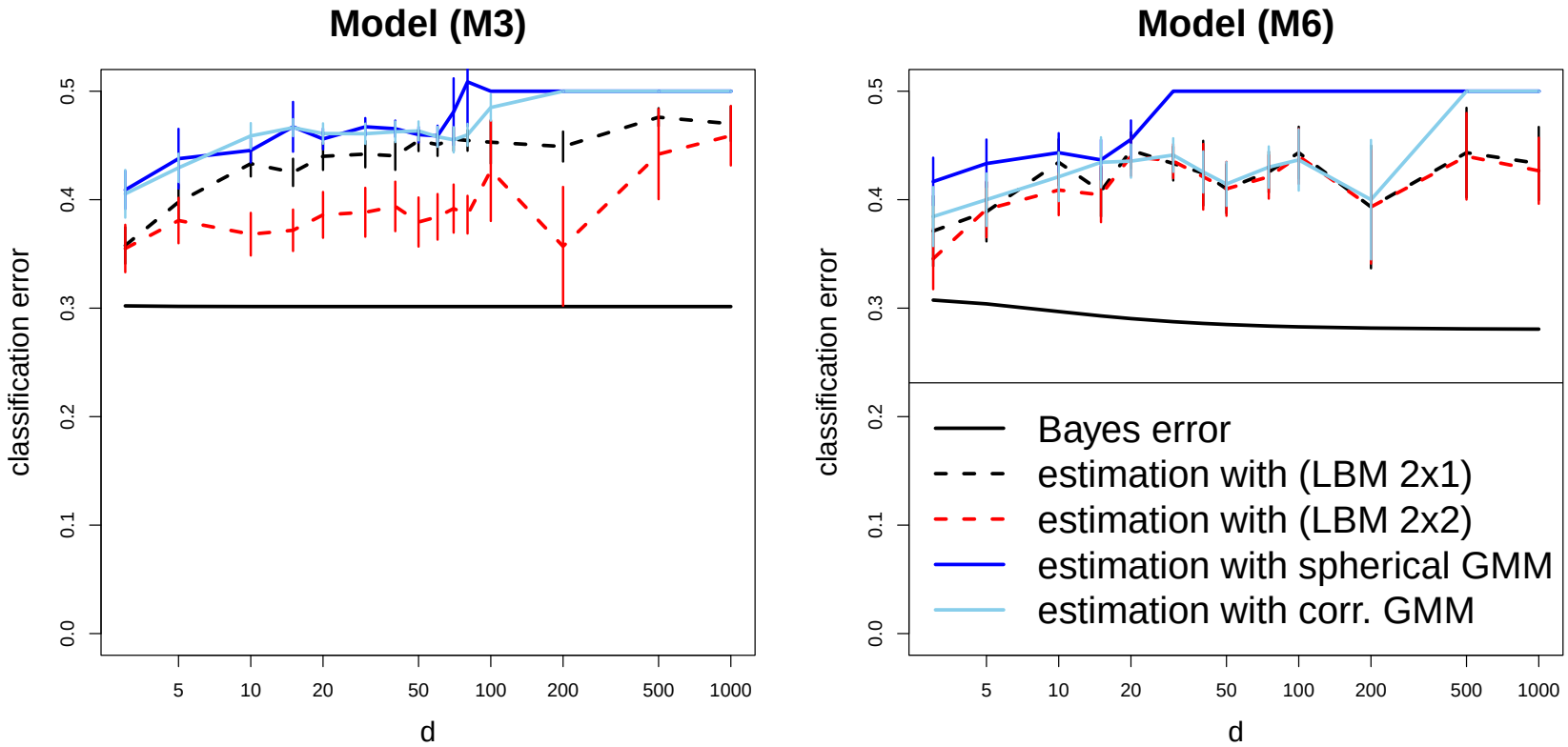
$\mu_1 \sim \mathcal{N}_d(0, Id)$

$\mu_2 \sim \mathcal{N}_d((\underbrace{1, \ldots, 1}_{\sqrt{d}}, 0, \ldots, 0), Id)$

redundancy: duplicated variables

# Numerical illustration: too many noninformative variables

$\mu_1 = \mathbf{0}_d$, $\mu_2 = (1, 2^{-2}, \ldots, d^{-2})$




- too much bias, classification error with CC no longer lines up with Bayes risk
- but CC methods still perform better than simple mixture

# Take home message

- in HD clustering, accept model bias to reduce variance on the classification
- co-clustering acts as a regularized tool to perform clustering
  ↪ very parsimonious model, with model bias but often offers better classification error
- LBM also offers a level of flexibility despite its parsimony
  ↪ using LBM for HD clustering should be more emphasized
  ↪ estimation is challenging, but there are solutions
  ↪ less cases of spurious / degenerate solutions than in simple mixture models

For further research...

- Interpret the groups of variables in co-clustering
- Extend the empirical results by theoretical guidelines
- Robust estimation
- Model selection: to be fixed by theoretical results

# Merci / Thank you!

Journal of Classification
https://doi.org/10.1007/s00357-023-09441-3

INVITED ARTICLE


## A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges

C. Biernacki[1] · J. Jacques[2] · C. Keribin[3]

Accepted: 9 May 2023