



HAL
open science

MODEL BASED CO-CLUSTERING: HIGH DIMENSION & ESTIMATION CHALLENGES

Christophe Biernacki, Julien Jacques, Christine Keribin

► **To cite this version:**

Christophe Biernacki, Julien Jacques, Christine Keribin. MODEL BASED CO-CLUSTERING: HIGH DIMENSION & ESTIMATION CHALLENGES. CFE-CMStatistics 2024 - The 18th International Joint Conference on Computational and Financial Econometrics (CFE) and Computational and Methodological Statistics (CMStatistics), Dec 2024, Londres, United Kingdom. hal-04867739

HAL Id: hal-04867739

<https://inria.hal.science/hal-04867739v1>

Submitted on 6 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

MODEL BASED CO-CLUSTERING: HIGH DIMENSION & ESTIMATION CHALLENGES

Christophe BIERNACKI¹, Julien JACQUES², Christine KERIBIN³

¹Inria, CNRS, Laboratoire de Mathématiques Painlevé, U. Lille, F59650 V. d'Ascq, France

²U. Lyon 2, ERIC, UR 3083, 5 Avenue Pierre Mendès-France, F69676 Bron Cedex, France

³U. Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, F91405 Orsay, France

18th International Joint Conference CFE-CMStatistics
14-16 December 2024, King's College London, UK



Outline

1 Motivation

2 Latent Block Model

3 Estimation challenges

4 Co-clustering for HD

Clustering

Unsupervised ML framework

- ▶ data set $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathcal{X}^n$ with $\dim(\mathcal{X}) = d$
- ▶ to partition into K clusters G_1, \dots, G_K
- ▶ classification matrix $\mathbf{z} = (z_1, \dots, z_n)$: $z_{ik} = \mathbb{1}_{i \in G_k}$

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

- ▶ \mathbf{x}_i are iid observations of a **mixture** pdf

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

- ▶ estimation with EM algorithm, partition is deduced from the conditional probability $p(z_i | \mathbf{x}_i; \hat{\theta})$, typically from the MAP principle, as soon as we have an estimate $\hat{\theta}$ of θ

High dimension (HD) settings: $d \sim n, d \gg n$

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

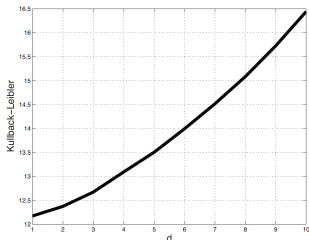
- ▶ marketing: $d \sim 10^2$ for $n = 10^4 - 10^6$
- ▶ microarray gene expression: $d \sim 10^2 - 10^4$ for $n = 20 - 50$ stress conditions
- ▶ fMRI images: $d \sim 10^4 - 10^5$ voxels for $n = 50 - 100$ patients
- ▶ text mining, curve, internet logs. . .

Moreover, variables can be of 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



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

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

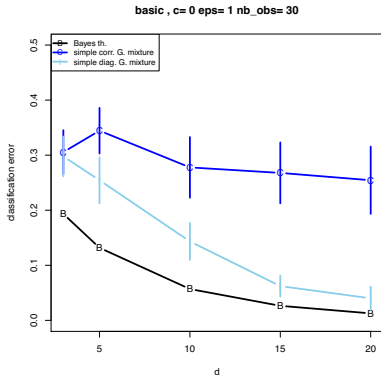
Basic case: $\mu_1 = \mathbf{0}_d; \mu_2 = \mathbf{1}_d; \Sigma = \mathbf{I}_d$

- **blessing**: but components are (in that case) more and more separated
- which impact on the misclassification error?
 - when the variables are correlated, redundant, non informative?
 - when $d \gg n$?

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

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

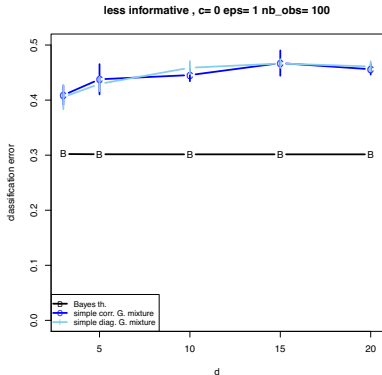
Basic case: $\Sigma = \mathbf{I}_d; \mu_1 = \mathbf{0}_d; \mu_2 = \mathbf{1}_d$

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

$$x_j | z_j \sim \mathcal{N}_d(\mu_{z_j}, \Sigma); \pi_1 = \pi_2 = 0.5$$

Less informative: $\Sigma = \mathbf{I}_d; \mu_1 = \mathbf{0}_d,$

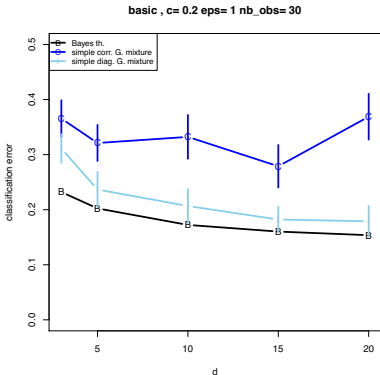
$$\mu_2 = \left(1, \frac{1}{2^2}, \dots, \frac{1}{d^2}\right)$$

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

$$x_j | z_j \sim \mathcal{N}_d(\mu_{z_j}, \Sigma); \pi_1 = \pi_2 = 0.5$$

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

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

Some alternatives for reducing variance

- ▶ Dimension reduction in non canonical space (PCA-like typically)
 - ▶ Dimension reduction in the canonical space (variable selection)
 - Gaussian mixture with 3 variable-clusters (informative, non informative, linearly dependent) [Maugis et al (2009)]
 - latent class analysis [Fop et al (2018)]
 - ▶ Model parsimony in the initial HD space (constraints on model parameters)
 - parsimonious Gaussian mixture [Celeux and Govaert, '95]
 - mixture of factor analyzers [[Ghahramani and Hinton, 97], [McLachlan et al], 03]
 - HD Gaussian models [[Bouveyron et al, 07]]
- ↪ trade-off between “meaningful” modeling and what can be estimated in practice

An alternative for reducing the variance

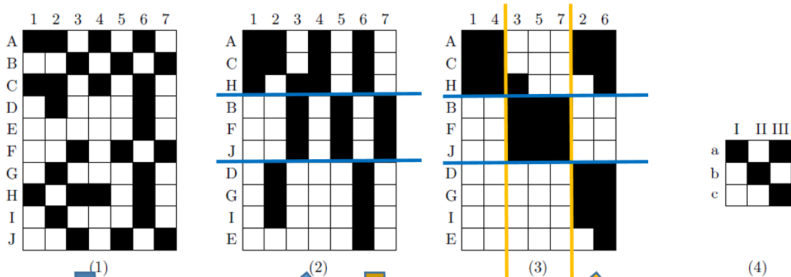
Main message:

- ▶ clustering is a way for dealing with large n
- ↪ why not also reusing this idea for large d ?

Co-clustering (CC)

- ▶ simultaneously clusters rows and columns
- ▶ performs parsimony of row clustering through variable clustering
- ↪ 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 an MBC dimension reduction method¹

(See the Latent Block Model (LBM) definition in the next section...)

- ▶ **PCA** for the i -th data individual \mathbf{x}_i : acts as a dimension reduction method

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

- ▶ specific **CC** case reduced to $K = 1$: again, acts as a dim. reduction method

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

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

- ↪ return now to the clustering task:
- **sequential clustering** method: (1) PCA and then (2) MBC
 - **simultaneous clustering** method: only CC, since dim. reduction is included

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

Outline

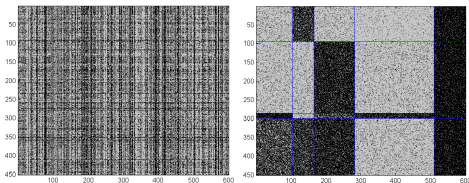
1 Motivation

2 Latent Block Model

3 Estimation challenges

4 Co-clustering for HD

Unsupervised block clustering framework



- ▶ **co-clustering (CC):** 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

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

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

- ▶ **traditional applications:** huge data sets arising in recommendation systems (binary), genomic data analysis (Gaussian), text mining (Poisson),...
- ▶ **our proposed application:** dimension reduction for HD clustering

Latent Block Model (LBM): a mixture model²

H_2 : row and column labels are independently assigned :

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

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

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

Observed likelihood

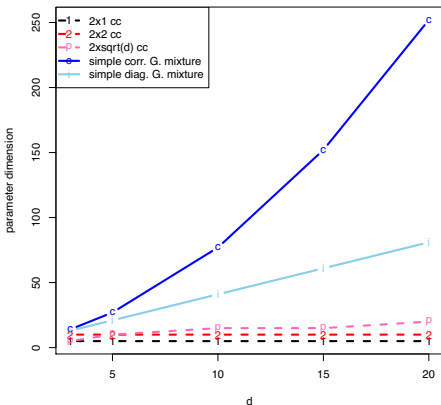
$$p(x; \theta) = \sum_{(z, w) \in \mathcal{Z} \times \mathcal{W}} \underbrace{\prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}}}_{p(z, w; \theta) = p(z; \theta) p(w; \theta)} \underbrace{\prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik} w_{j\ell}}}_{p(x|z, w; \theta)}$$

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

²[Govaert and Nadif (2008)]

Parametric model

- ▶ very **parsimonious** model, good candidate to deal with HD settings...



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	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + 2KL$

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

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** [Keribin et al., '15]
- ▶ **likelihood intractable**: 2×2 blocks and 20×20 matrix: $\approx 2^{20} \times 2^{20} \approx 10^{12}$ terms
- ▶ **E-step intractable**
 - ↪ approximation with Variational EM (VEM) [Govaert and Nadif '08]
 - ↪ estimation with Stochastic EM (SEM) [Keribin et al. '10]
 - ↪ stochastic gradient descent [Baey et al. '23]
 - ↪ Bayesian versions: Gibbs and V-Bayes algorithms [Keribin, Brault, Celeux, Govaert, '15]
- ▶ **consistency and asymptotic normality** of ML- and V- estimators (when $\log(n)/d \rightarrow 0$ and $\log(d)/n \rightarrow 0$) [Brault, Keribin, Mariadassou, '20]
- ▶ **model selection** (ICL)

VEM for LBM

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

VEM (iteration r)

► approximated E-step:

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

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

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

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

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

Outline

- 1 Motivation
- 2 Latent Block Model
- 3 Estimation challenges**
- 4 Co-clustering for HD

Estimation challenges

Challenges in the (classical) MBC context:

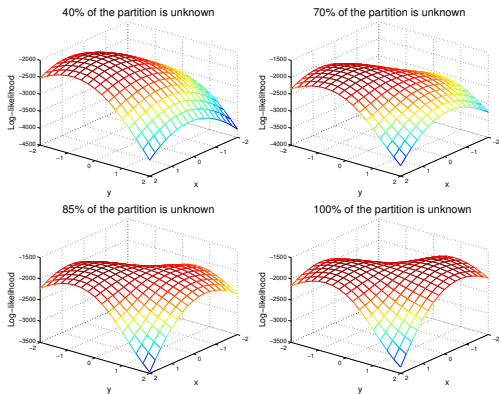
- ▶ local maxima solutions
- ▶ empty cluster solutions
- ▶ degenerate solutions

Question to be addressed here

What about LBM?

Local maxima

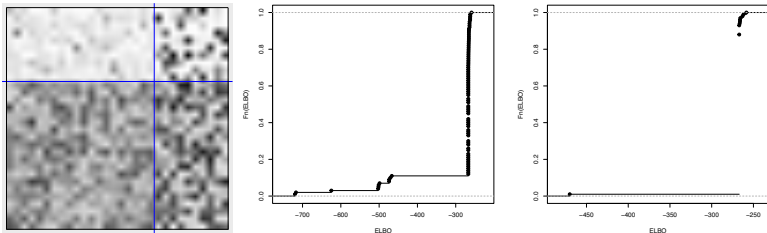
- ▶ not just linked to the number of parameters!
- ▶ linked to the number of latent variables (here a toy two-param. Gaussian mixture)



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

Local maxima in Gaussian LBM: illustration³

Figure: Gaussian 2×2 LBM (*left*) ; e.c.d.f. of ELBO values obtained from $B = 100$ initializations, at both 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

Empty clusters in LBM

Really frequent (and reported)!

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

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

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

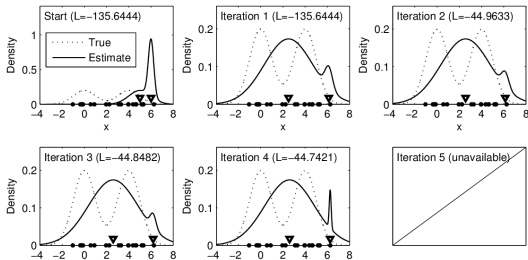
- ▶ **exponential speed** even with very small sample sizes

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

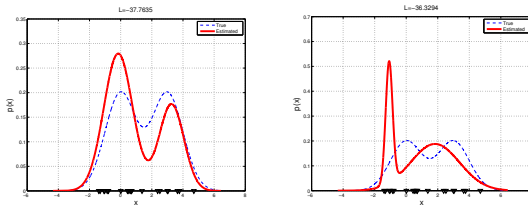
↔ initialization often achieved by performing **independent clustering** of the rows and columns but the best initialization strategy stays **an open question**

Degenerate and spurious solutions in MBC

- ▶ **degenerate** solutions lay on the border of the 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?

- ▶ freq. of obtaining a **cluster** with a single element in the whole partition latent space

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

- ▶ freq. of obtaining a **block** with a single element in the whole partition latent space

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

↔ 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)

Outline

- 1 Motivation
- 2 Latent Block Model
- 3 Estimation challenges
- 4 Co-clustering for HD**

Properties of LBM in view of HD clustering

- ▶ parsimonious model
- ▶ the true partition is recovered when there is a consistent estimator [Mariadassou and Matias, '15]

$$\hat{\theta} \xrightarrow{n, d \rightarrow \infty} \theta^* \Rightarrow p(\hat{\mathbf{z}} = \mathbf{z}^*, \hat{\mathbf{w}} = \mathbf{w}^* | \mathbf{x}; \hat{\theta}) \xrightarrow{n, d \rightarrow \infty} 1,$$

- ▶ non asymptotic properties (binary case) [Brault, '14]

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

↪ LBM is a natural regularized candidate for HD clustering

Numerical illustrations: experimental design

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

	\mathbf{I}_d	$\Sigma(c)$
$\mu_1 = \mathbf{0}_d, \mu_2 = \mathbf{1}_d$	(M_1)	(M_2)
$\mu_1 = \mathbf{0}_d, \mu_2 = (1, 2^{-2}, \dots, d^{-2})$	(M_3)	(M_6)
(M_1) of size $n \times d/2$ duplicated twice	(M_4)	
$\mu_1 \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d), \mu_2 \sim \mathcal{N}_d(\underbrace{(1, \dots, 1)}_{\sqrt{d}}, 0, \dots, 0), \mathbf{I}_d)$	(M_5)	

→ only (M_1) is a nominal LBM (2×1)

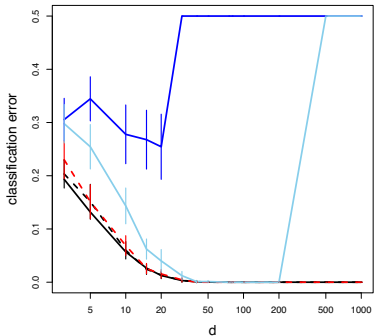
Row clustering is performed using four different methods

- ▶ 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 (2×1) LBM
- ▶ co-clustering with a Gaussian (2×2) LBM

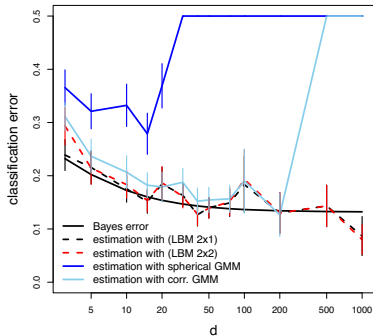
Classification error averaged over 30 samples of size $n = 30$

Numerical illustrations: results

Model (M1)



Model (M2)

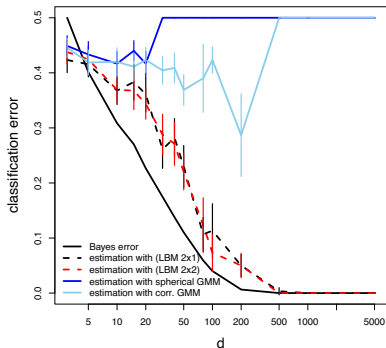


LBM resists to HD

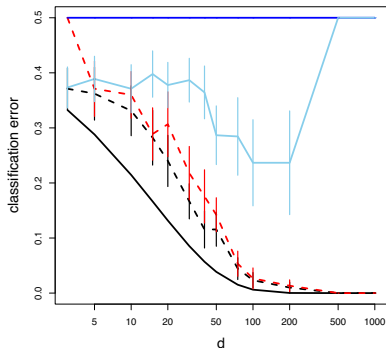
LBM performs better than simple mixture, although the underlying model overrides the co-clustering conditional independence assumptions

Numerical illustrations: results (cont'd)

Model (M5)



Model (M4)



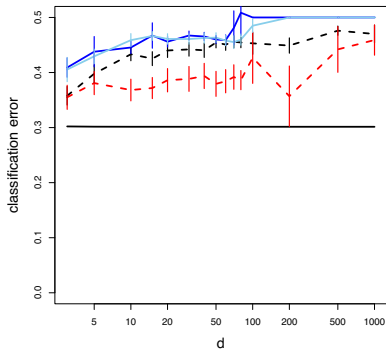
non informative variables

redundancy: duplicated variables

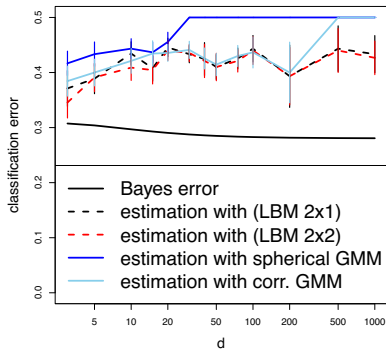
Same conclusion for LBM, which resists to such HD scenarios

Numerical illustrations: results (cont'd)

Model (M3)



Model (M6)



too many noninformative variables

- ▶ 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
- ▶ CC 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. Biernacki¹ · J. Jacques² · C. Keribin³

Accepted: 9 May 2023

© The Author(s) under exclusive licence to The Classification Society 2023