Classification non-supervisée de données de grande dimension et de graphes à l'aide de modèles à variables latentes discrètes

Soutenance de Thèse

Nicolas Jouvin

Directeurs : P. Latouche, C. Bouveyron & A. Livartowski

Vendredi 11 décembre 2020

institut Curie

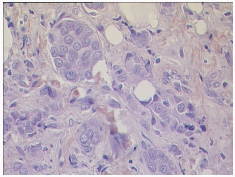
UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

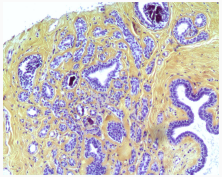
MATH INNOV

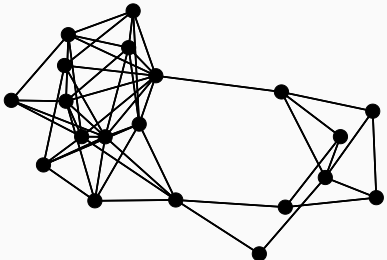
île de France

Clustering in a nutshell

It's a data's world...



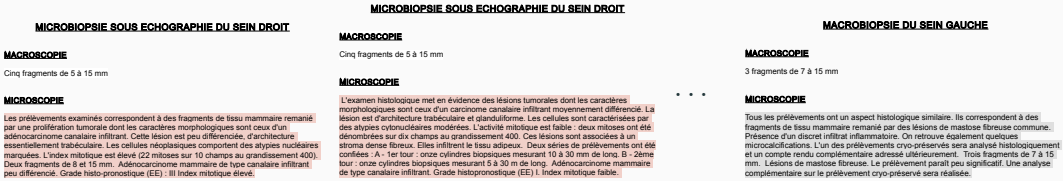




Clustering is the task of grouping objects together into classes or *clusters*, in an unsupervised fashion based on some criterion.

Example 1: document clustering

Grouping similar texts together based on their topics.



MICROBIOPSIE SOUS ECHOGRAPHIE DU SEIN DROIT

MACROSCOPIE

Cinq fragments de 5 à 15 mm

MICROSCOPIE

Les prélèvements examinés correspondent à des fragments de tissu mammaire remanié par une prolifération tumorale dont les caractères morphologiques sont ceux d'un adénocarcinome canalaire infiltrant. Cette lésion est peu différenciée, d'architecture essentiellement trabéculaire. Les cellules néoplasiques comportent des atypies nucléaires marquées. L'index mitotique est élevé (22 mitoses sur 10 champs au grandissement 400). Deux fragments de 8 et 15 mm. Adénocarcinome mammaire de type canalaire infiltrant peu différencié. Grade histo-pronostique (EE) : III Index mitotique élevé.

MICROBIOPSIE SOUS ECHOGRAPHIE DU SEIN DROIT

MACROSCOPIE

Cinq fragments de 5 à 15 mm

MICROSCOPIE

L'examen histologique met en évidence des lésions tumorales dont les caractères morphologiques sont ceux d'un carcinome canalaire infiltrant moyennement différencié. La lésion est d'architecture trabéculaire et glanduliforme. Les cellules sont caractérisées par des atypies cytonucléaires modérées. L'activité mitotique est faible : deux mitoses ont été dénombrées sur dix champs au grandissement 400. Ces lésions sont associées à un stroma dense fibreux. Elles infiltrent le tissu adipeux. Deux séries de prélèvements ont été confiées : A - 1er tour : onze cylindres biopsiques mesurant 10 à 30 mm de long. B - 2ème tour : onze cylindres biopsiques mesurant 5 à 30 m de long. Adénocarcinome mammaire de type canalaire infiltrant. Grade histopronostique (EE) I. Index mitotique faible.

. . .

MACROBIOPSIE DU SEIN GAUCHE

MACROSCOPIE

3 fragments de 7 à 15 mm

MICROSCOPIE

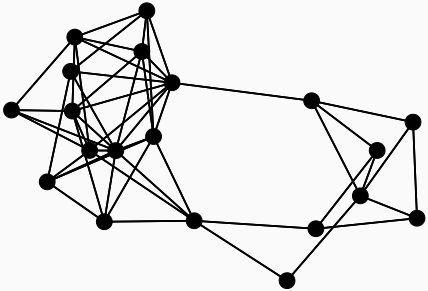
Tous les prélèvements ont un aspect histologique similaire. Ils correspondent à des fragments de tissu mammaire remanié par des lésions de mastose fibreuse commune. Présence d'un discret infiltrat inflammatoire. On retrouve également quelques microcalcifications. L'un des prélèvements cryo-préservés sera analysé histologiquement et un compte rendu complémentaire adressé ultérieurement. Trois fragments de 7 à 15 mm. Lésions de mastose fibreuse. Le prélèvement paraît peu significatif. Une analyse complémentaire sur le prélèvement cryo-préservé sera réalisée.

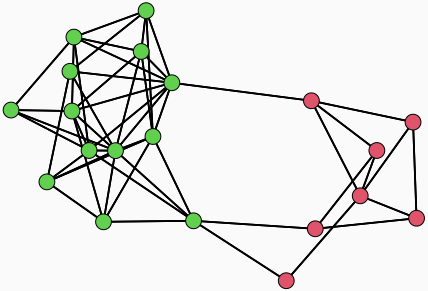
Doc 1	“Lésions cancéreuses (. . .) carcinome canalaire”
Doc 2	“Lésions cancéreuses (. . .) carcinome lobulaire”
. . .	. . .
Doc n	“Lésions bénignes (. . .) métaplasie”

Clustering is the task of grouping objects together into classes or *clusters*, in an unsupervised fashion based on some criterion.

Example 2: Network clustering

Group nodes of a network based on their connections with respect to others

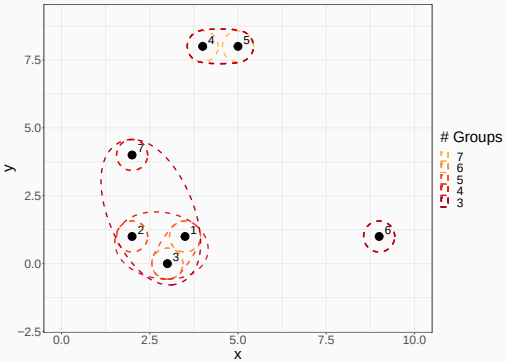


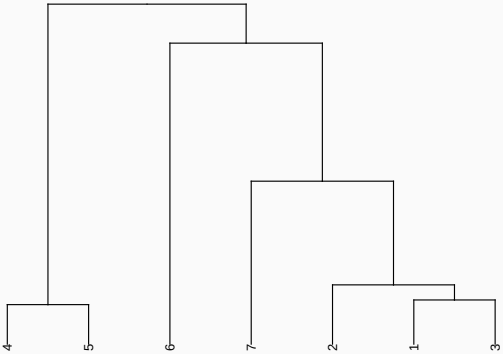


Clustering is the task of grouping objects together into classes or *clusters*, in an unsupervised fashion based on some criterion.

Example 3: Hierarchical clustering

Build a hierarchy of *nested* clusters





My contributions

Probabilistic approach for three types of data

- Count data, *e.g.* text documents
- Continuous data, *e.g.* images
- Graph data

Handle high-dimensionality: large number p of variables

Joint clustering & model selection: select the K number of clusters

Model-based clustering

The rationale of model-based clustering

Observe $\boldsymbol{Y}$ related to n objects

Search for $\boldsymbol{z}_i \in \{0,1\}^K$ the cluster assignment of object i

Assume $\boldsymbol{Z} = \{\boldsymbol{z}_i\}$ contains independent and identically distributed (*i.i.d.*) discrete latent variables

$$p(\boldsymbol{Z} \mid \boldsymbol{\pi}) = \prod_{i=1}^{n} \mathcal{M}_K(\boldsymbol{z}_i \mid 1, \boldsymbol{\pi})$$

Posit a statistical model on $\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}$

- perform inference, *e.g.* maximum-likelihood, to get $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})$
- use the posterior $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})$

A fundamental assumption: conditional independence

Discrete Latent Variable Models (DLVMs)

$$p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{\boldsymbol{y} \in \boldsymbol{Y}} p(\boldsymbol{y} \mid \boldsymbol{Z}, \boldsymbol{\theta}) \tag{1}$$

Example 1: Finite Mixture Models (FMM)

Observations $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ are *i.i.d.* inside a cluster

$$\forall i, \quad \boldsymbol{y}_i \mid \{z_{ik} = 1\} \sim p(\cdot \mid \boldsymbol{\theta}_k)$$

- Gaussian mixture model: $p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k) = \mathcal{N}_p(\boldsymbol{y}_i \mid \boldsymbol{m}_k, \boldsymbol{S}_k)$
- Mixture of multinomials: $p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k) = \mathcal{M}_p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k)$

$$p(\boldsymbol{Y} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^n p(\boldsymbol{y}_i \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k)$$

A fundamental assumption: conditional independence

Discrete Latent Variable Models (DLVMs)

$$p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{\boldsymbol{y} \in \boldsymbol{Y}} p(\boldsymbol{y} \mid \boldsymbol{Z}, \boldsymbol{\theta}) \tag{1}$$

Example 2: Stochastic Block Model (SBM)

Observe n^2 edges $\boldsymbol{Y} = \{y_{ij}\}_{ij}$, cluster n nodes

$$\forall (i,j), \quad y_{ij} \mid \{z_{ik} z_{jl} = 1\} \sim p(\cdot \mid \boldsymbol{\theta}_{kl})$$

Edges are *i.i.d.* inside a *block* of clusters, **not marginally**

- Binary SBM: $p(y_{ij} \mid \boldsymbol{\theta}_{kl}) = \mathcal{B}(y_{ij} \mid \theta_{kl})$
- Poisson SBM: $p(y_{ij} \mid \boldsymbol{\theta}_{kl}) = \mathcal{P}(y_{ij} \mid \theta_{kl})$

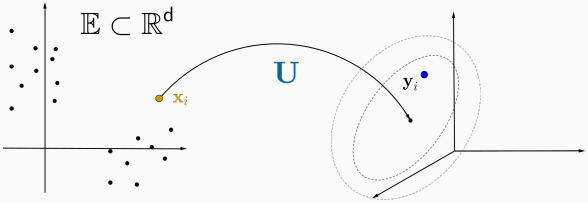
Main challenges tackled in this thesis

High-dimensional clustering: mixture estimation is cumbersome

- Gaussian mixtures: $\mathcal{O}(p^2)$ free parameters
- Small-sample scenario $n << p$

Probabilistic dimension reduction

$$\boldsymbol{x}_i \in \mathbb{E} \subset \mathbb{R}^d \longrightarrow \boldsymbol{y}_i \approx \boldsymbol{U}\boldsymbol{x}_i$$



Greedy clustering: discrete optimization w.r.t. $\boldsymbol{Z}$

- Joint inference and clustering
- Joint model selection and clustering

Outline

High-dimensional count data clustering

The Bayesian Fisher-EM algorithm

Hierarchical model-based clustering in DLVMs

Conclusion

High-dimensional count data clustering

MICROBIOPSIE SOUS ECHOGRAPHIE DU SEIN DROIT

MACROSCOPIE

Cinq fragments de 5 à 15 mm

MICROSCOPIE

Les prélèvements examinés correspondent à des fragments de tissu mammaire remanié par une prolifération tumorale dont les caractères morphologiques sont ceux d'un adénocarcinome canalaire infiltrant. Cette lésion est peu différenciée, d'architecture essentiellement trabéculaire. Les cellules néoplasiques comportent des atypies nucléaires marquées. L'index mitotique est élevé (22 mitoses sur 10 champs au grandissement 400). Deux fragments de 8 et 15 mm. Adénocarcinome mammaire de type canalaire infiltrant peu différencié. Grade histo-pronostique (EE) : III Index mitotique élevé.

. . .

MACROBIOPSIE DU SEIN GAUCHE

MACROSCOPIE

3 fragments de 7 à 15 mm

MICROSCOPIE

Tous les prélèvements ont un aspect histologique similaire. Ils correspondent à des fragments de tissu mammaire remanié par des lésions de mastose fibreuse commune. Présence d'un discret infiltrat inflammatoire. On retrouve également quelques microcalcifications. L'un des prélèvements cryo-préservés sera analysé histologiquement et un compte rendu complémentaire adressé ultérieurement. Trois fragments de 7 à 15 mm. Lésions de mastose fibreuse. Le prélèvement paraît peu significatif. Une analyse complémentaire sur le prélèvement cryo-préservé sera réalisée.

Document-term matrix

Documents \ Terms	lésions	canalaire	...	lobulaire	métaplasie
"Lésions (...) carcinome canalaire"	2	1	...	0	0
"Lésions bénignes (...) métaplasie"	3	0	...	0	1

Probabilistic dimension reduction for count data

Count data: $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ with $\boldsymbol{y}_i \in \mathbb{N}^p$

- *e.g.* word count in a document, *read* count in a gene
- Total count $c_i := \sum_j y_{ij}$
- Zero inflated data, high-dimensional problems $n << p$

Multinomial PCA (MPCA, Buntine 2002)

$$\boldsymbol{x}_i \sim \mathcal{D}_d(\boldsymbol{\alpha}) \qquad \text{(latent space: } \Delta_d\text{)}$$

$$\boldsymbol{y}_i \mid \boldsymbol{x}_i \sim \mathcal{M}_p(c_i, \boldsymbol{U}\boldsymbol{x}_i) \qquad \text{(observation space)}$$

- $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d] \in (\Delta_p)^d$ is called the *topic* matrix
- also known as Latent Dirichlet Allocation (Blei et al. 2003)

Mixture of multinomial PCA (Yu et al. 2005)

One latent variable per cluster:

$$\boldsymbol{x} = (\boldsymbol{x}_k)_k, \quad \boldsymbol{x}_k \overset{i.i.d.}{\sim} \mathcal{D}_d(\boldsymbol{\alpha})$$

$$\forall i, \quad \boldsymbol{y}_i \mid \boldsymbol{x} \sim \sum_{k=1}^{K} \pi_k \, \mathcal{M}_p(c_i, \boldsymbol{U}\boldsymbol{x}_k)$$

Constrained multinomial model: $\boldsymbol{\theta}_k = \boldsymbol{U}\boldsymbol{x}_k$ (Carel et al. 2017)

Property

Suppose $\boldsymbol{Z}$ known and fixed, construct K meta-observations

$$\tilde{\boldsymbol{Y}}_k(\boldsymbol{Z}) = \sum_{i=1}^{n} z_{ik} \boldsymbol{y}_i$$

Then, $\boldsymbol{Y} \mid \boldsymbol{Z}$ follows a MPCA model on $\tilde{\boldsymbol{Y}}(\boldsymbol{Z})$

Classification likelihood approach

$$\underset{\boldsymbol{Z},\boldsymbol{U},\boldsymbol{\pi}}{\arg\max}\left\{\log p(\boldsymbol{Y},\boldsymbol{Z}\mid\boldsymbol{\pi},\boldsymbol{U}) = \log p(\boldsymbol{Z}\mid\boldsymbol{\pi}) + \underbrace{\log p(\boldsymbol{Y}\mid\boldsymbol{Z},\boldsymbol{U})}_{(*)\text{ MPCA on }\tilde{\boldsymbol{Y}}(\boldsymbol{Z})}\right\}$$

Problems:

1. Combinatorics: number of partitions exponential with n
2. $(*)$ is intractable because of marginal over $\boldsymbol{x}$

Solutions:

1. Variational inference layer on $\boldsymbol{x}$

$$\boldsymbol{x}\sim q, \qquad \log p(\boldsymbol{Y},\boldsymbol{Z}\mid\boldsymbol{\pi},\boldsymbol{U}) \geq \mathcal{J}(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{U},q)$$

2. Greedy algorithm for joint inference and clustering

Branch & bound C-VEM algorithm

Algorithm: Explore partition space using $\mathcal{J}$ as a surrogate objective

Input: K, d, $\boldsymbol{Z}^{(0)}, \boldsymbol{\pi}^{(0)}$, $\boldsymbol{U}$

while $\boldsymbol{Z}$ *has not converged* **do**

For all $i = 1, \ldots, n$, try individual swaps: $z_{ik}^{(t)} = 1 \rightarrow z_{il}^{(tmp)} = 1$

`// Difference with standard greedy approaches`

Use variational inference to update q

$$(\mathcal{J}_l, q_l) = \arg\max_q \mathcal{J}(\boldsymbol{Z}^{(tmp)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{U}, q)$$

Select $l^\star = \arg\max_l \mathcal{J}_l$

$$z_{il^\star}^{(t+1)} = 1, \qquad q^{(t+1)} = q_{l^\star} \qquad \boldsymbol{\pi}^{(t+1)} = \textstyle\sum_i \boldsymbol{z}_i^{(t+1)}/n$$

end

Model selection

How to choose the pair (K, d) ?

Integrated Classification Likelihood (ICL, Biernacki et al. 2000)

$$\log p(\boldsymbol{Y}, \boldsymbol{Z}) = \int_{\boldsymbol{\pi}} \int_{\boldsymbol{U}} \log p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{U}) \, \mathrm{d}\boldsymbol{U} \, \mathrm{d}\boldsymbol{\pi}$$

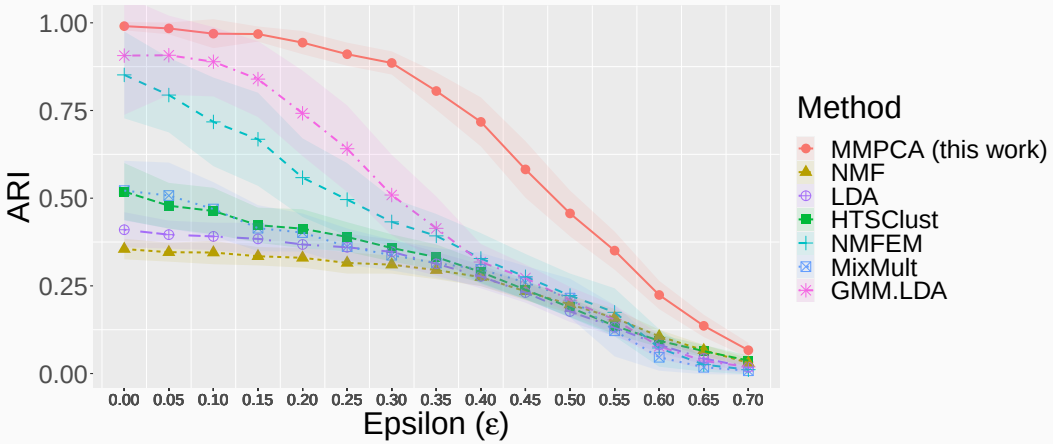
ICL criterion for MMPCA

Laplace and Stirling approximations combined with a variational approximation on $p(\boldsymbol{Y} \mid \boldsymbol{Z})$ lead to

$$\mathrm{ICL}_{MMPCA}(K, d) = \mathcal{J}(\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{U}}, \hat{q}) - \frac{d(p-1)}{2} \log(K) - \frac{K-1}{2} \log(n)$$

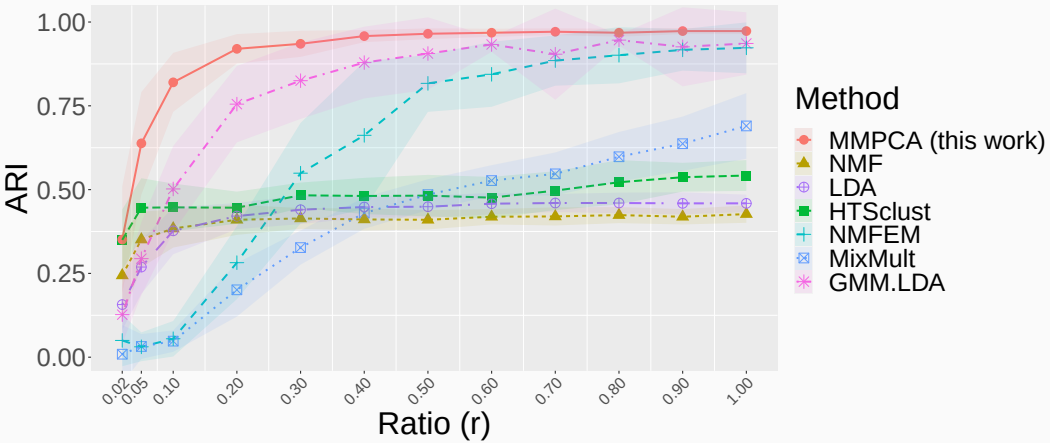
Scenario 1: noisy setting

$n = 400,$ $p = 1000,$ Noise level: $\epsilon \in [0, 1]$



Scenario 2: small-sample sizes

$p = 1000,$ $\epsilon = 0.2,$ $n = r \times p,\ r \in [0, 1]$



Application: clustering of anatomopathological reports

Context: textual reports describing histopathological slides

- Benign
- Lobular carcinoma
- Non Special Type (NST) carcinoma, *e.g.* ductal

Unsupervised analysis: select $K = 7$ and $d = 5$

	Benign	NST carcinoma	Lobular carcinoma
1	0	0	43
2	1	31	1
3	0	106	0
4	231	3	0
5	0	211	0
6	0	126	0
7	0	113	0

Application: clustering of anatomopathological reports

Context: textual reports describing histopathological slides

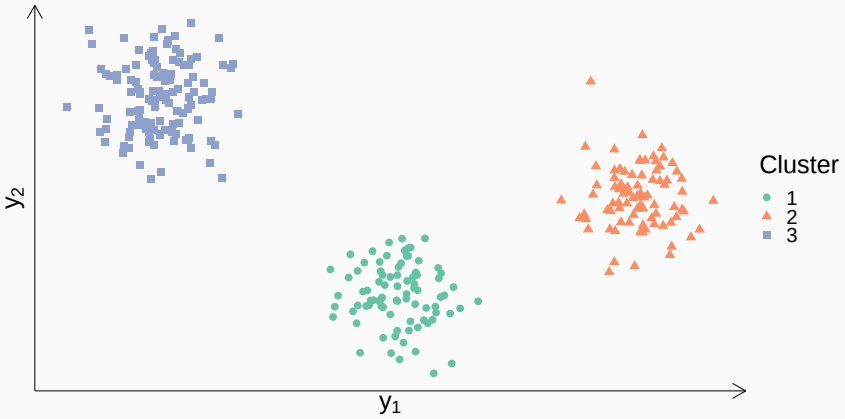
- Benign
- Lobular carcinoma
- Non Special Type (NST) carcinoma, *e.g.* ductal

Unsupervised analysis: select $K = 7$ and $d = 5$

	Benign	NST carcinoma	Lobular carcinoma
1	0	0	43
2	1	31	1
3	0	106	0
4	231	3	0
5	0	211	0
6	0	126	0
7	0	113	0

The Bayesian Fisher-EM algorithm

Low-dimensional mixture



What if p is large ?

Gaussian subspace clustering

Continuous data: $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, $\boldsymbol{y}_i \in \mathbb{R}^p$

$$\forall i, \quad \boldsymbol{y}_i \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}_p(\boldsymbol{m}_k, \boldsymbol{S}_k) \tag{GMM}$$

Problem when p is large: over-parameterized $\boldsymbol{S}_k$

Gaussian subspace clustering

Continuous data: $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, $\boldsymbol{y}_i \in \mathbb{R}^p$

$$\forall i, \quad \boldsymbol{y}_i \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}_p(\boldsymbol{m}_k, \boldsymbol{S}_k) \tag{GMM}$$

Problem when p is large: over-parameterized $\boldsymbol{S}_k$

Constrained GMM: **low-rank covariance**

$$\boldsymbol{m}_k = \boldsymbol{U}\boldsymbol{\mu}_k \qquad \boldsymbol{S}_k = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{U} + \boldsymbol{\Psi}_k, \quad \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d$$

Factor analysis formulation: low-dimensional embedding

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \text{(Latent space: } \mathbb{R}^d\text{)}$$

$$\boldsymbol{y}_i = \boldsymbol{U}\boldsymbol{x}_i + \epsilon_{ik}, \quad \epsilon_{ik} \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Psi}_k) \qquad \text{(Observation space)}$$

The tension between clustering and density estimation

Maximum-likelihood estimation (MLE):

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{U}}) \in \underset{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{U}}{\arg\max} \ \log p(\boldsymbol{Y} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{U})$$

PCA-like objective: preserves variance of the signal

Discriminative subspace

Supervised: Fisher Discriminant Analysis (FDA, Fisher 1936)

1. $\boldsymbol{Z}$ is known: construct scatter matrices

$$\boldsymbol{S}_B = \sum_k n_k (\boldsymbol{m}_k - \bar{\boldsymbol{y}})(\boldsymbol{m}_k - \bar{\boldsymbol{y}})^\top \qquad \text{(between-class)}$$

$$\boldsymbol{S}_W = \sum_k \sum_i z_{ik} (\boldsymbol{y}_i - \boldsymbol{m}_k)(\boldsymbol{y}_i - \boldsymbol{m}_k)^\top \qquad \text{(within-class)}$$

2. maximize their ratio in the latent space, $d \leq K - 1$

$$\hat{\boldsymbol{U}} = \arg\max_{U} \left\{ F(\boldsymbol{U}) := \mathrm{Tr}\left[(\boldsymbol{U}^\top \boldsymbol{S}_W \boldsymbol{U})^{-1} \boldsymbol{U}^\top \boldsymbol{S}_B \boldsymbol{U} \right] \right\}$$

Discriminative subspace

Supervised: Fisher Discriminant Analysis (FDA, Fisher 1936)

1. $\boldsymbol{Z}$ is known: construct scatter matrices

$$\boldsymbol{S}_B = \sum_k n_k(\boldsymbol{m}_k - \bar{\boldsymbol{y}})(\boldsymbol{m}_k - \bar{\boldsymbol{y}})^\top \qquad \text{(between-class)}$$
$$\boldsymbol{S}_W = \sum_k \sum_i z_{ik}(\boldsymbol{y}_i - \boldsymbol{m}_k)(\boldsymbol{y}_i - \boldsymbol{m}_k)^\top \qquad \text{(within-class)}$$

2. maximize their ratio in the latent space, $d \leq K - 1$

$$\hat{\boldsymbol{U}} = \arg\max_{\boldsymbol{U}} \left\{ F(\boldsymbol{U}) := \operatorname{Tr}\left[(\boldsymbol{U}^\top \boldsymbol{S}_W \boldsymbol{U})^{-1} \boldsymbol{U}^\top \boldsymbol{S}_B \boldsymbol{U} \right] \right\}$$

Unsupervised: use posterior (Bouveyron et al. 2012)

$$\tau_{ik} = p(z_{ik} = 1 \mid \boldsymbol{y}_i, \boldsymbol{\pi}, \boldsymbol{\Sigma}_k, \beta_k, \boldsymbol{U})$$

Bayesian discriminative latent mixture model (BDLM)

Problem: algorithmic instability, bad conditioning of the scatter matrices

New Gaussian subspace clustering model:

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_k), \quad \boldsymbol{\mu}_k \overset{i.i.d.}{\sim} \mathcal{N}_d(\boldsymbol{\nu}, \lambda \boldsymbol{I}_d)$$

$$\boldsymbol{y}_i \mid \boldsymbol{\mu} \overset{i.i.d.}{\sim} \textstyle\sum_k \pi_k \mathcal{N}_p(\boldsymbol{U}\boldsymbol{\mu}_k, \underbrace{\boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{U} + \boldsymbol{\Psi}_k}_{\boldsymbol{S}_k})$$

- λ controls the separation in the latent space
- $\boldsymbol{U}$ is supposed to be discriminative
- Block diagonal hypothesis on $\boldsymbol{S}_k = \boldsymbol{D}\boldsymbol{\Delta}_k\boldsymbol{D}^\top$:

$$\boldsymbol{\Delta}_k = \begin{pmatrix} \boldsymbol{\Sigma}_k & 0 \\ 0 & \beta_k \boldsymbol{I}_{p-d} \end{pmatrix}, \boldsymbol{D} = [\boldsymbol{U}, \boldsymbol{U}^\perp]$$

- Possible constraints on $\boldsymbol{\Sigma}_k$, 12 submodels

Joint inference and clustering

Objective: joint MLE and FDA

- Maximize likelihood with respect to $\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta}$
- Update τ_{ik} and maximize $F(\boldsymbol{U})$ w.r.t $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_d$

Problems:

1. orthonormal FDA $\rightarrow$ no closed-form solution
2. intractable likelihood because of marginalization on $(\boldsymbol{Z}, \boldsymbol{\mu})$

Solutions:

1. Use variational inference layer on $(\boldsymbol{Z}, \boldsymbol{\mu})$

$$(\boldsymbol{Z}, \boldsymbol{\mu}) \sim q \qquad \log p(\boldsymbol{Y} \mid \boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{U}) \geq \mathcal{J}(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{U}, q)$$

2. Use iterative procedure solving 1-D FDA problems

Bayesian Fisher-EM algorithm (BFEM)

Fix $\boldsymbol{U}^{(0)}$ and $(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})^{(0)}$ and iterate over

- VE-step: Find

$$q^{(t+1)} = \arg\max_{q} \mathcal{J}(\boldsymbol{\pi}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{U}^{(t)}, q)$$

- M-step: Find

$$(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})^{(t+1)} = \arg\max_{\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta}} \mathcal{J}(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{U}^{(t)}, q^{(t+1)})$$

- F-step: use $q^{(t+1)}(\boldsymbol{Z})$ to construct $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$, then

$$\boldsymbol{U}^{(t+1)} := \left[\boldsymbol{u}_1 \mid \ldots \mid \boldsymbol{u}_d\right], \text{ with } \forall h = 1, \ldots, d,$$

$$\boldsymbol{u}_h = \arg\max_{\boldsymbol{u} \in \mathbb{R}^p} F(\boldsymbol{u}) \text{ s.t. } \boldsymbol{u} \in \left\{\forall r < h, \boldsymbol{u}^\top \boldsymbol{u}_r = 0\right\}$$

Hyper-parameter and model selection

Questions:

1. How to set $(\boldsymbol{\nu}, \lambda)$? $\lambda \to +\infty \implies$ frequentist setting
2. How to choose K and a submodel $\mathcal{M}$?

Empirical Bayes:

$$(\hat{\boldsymbol{\nu}}, \hat{\lambda}) = \arg\max_{\boldsymbol{\nu}, \lambda} \mathcal{J}(\boldsymbol{\nu}, \lambda)$$

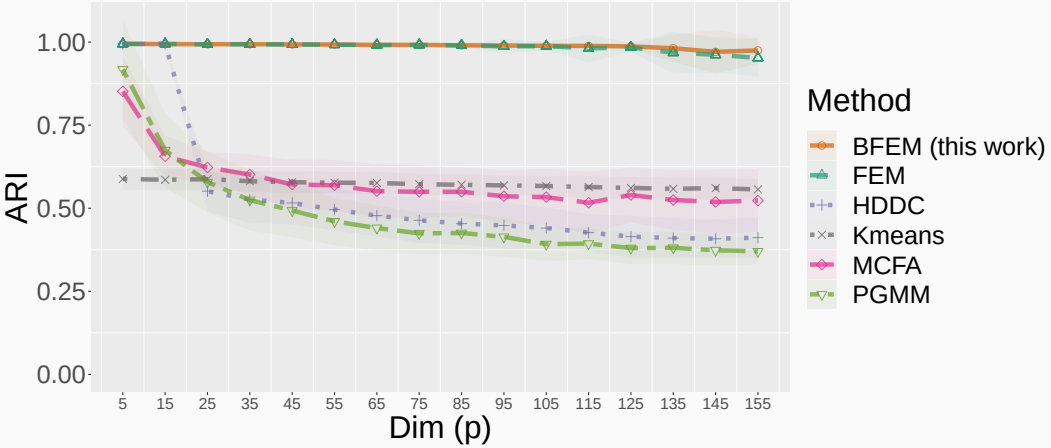
ICL criterion for BFEM

Denote $\gamma_{\mathcal{M},K}$ the number of free parameters in model $\mathcal{M}$ with K clusters

$$\mathrm{ICL}_{BIC}(\mathcal{M}, K) = \log p(\boldsymbol{Y}, \hat{\boldsymbol{Z}} \mid \hat{\boldsymbol{\vartheta}}, \mathcal{M}, K) - \frac{\gamma_{\mathcal{M},K}}{2} \log(n),$$

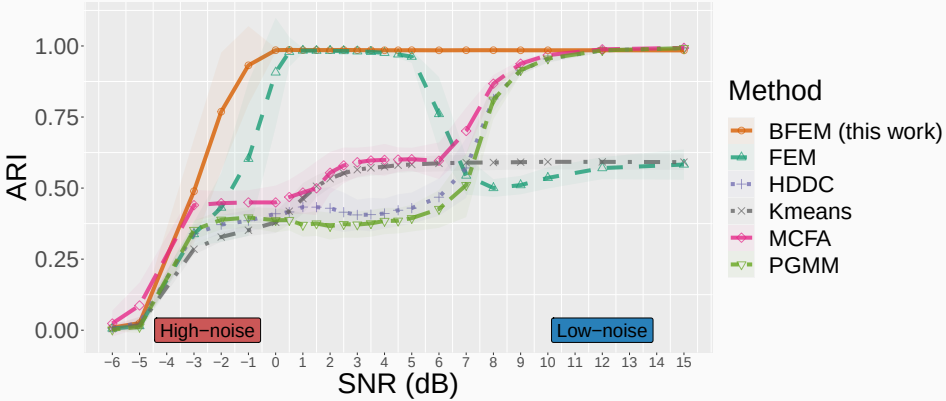
Scenario 1: increasing dimension

$n = 900, \qquad d = 2, \qquad \beta_k = 1, \qquad p \in \{5, 15, \ldots, 155\}$



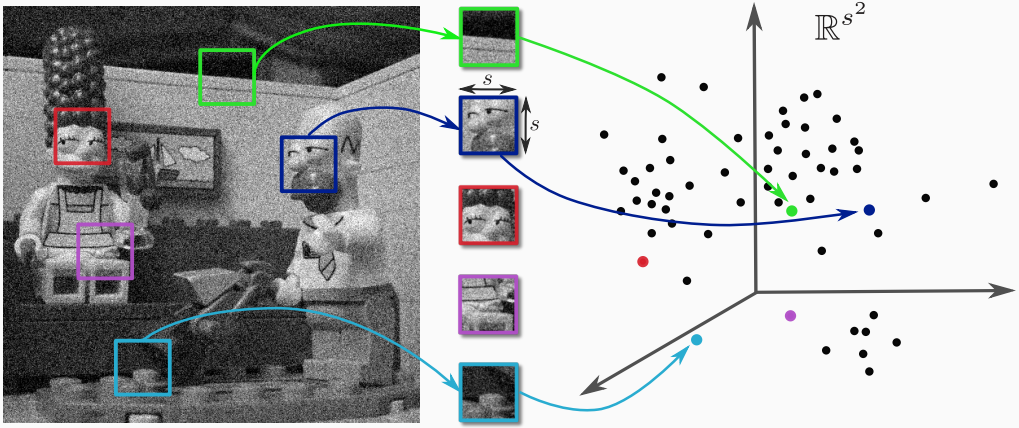
Scenario 2: signal-to-noise ratio

$$n = 900, \qquad d = 2, \qquad p = 150, \qquad \beta \in \{8, \ldots, 0.8, \ldots, 0.08\}$$



Application: patch-based image denoising (Houdard et al. 2018)

$$I = I_0 + N, \qquad N \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$



Alley image

Original

Noisy, $\sigma = 30$

Alley image

S-PLE, $PSNR = 28.22dB$

BFEM, $PSNR = 28.95dB$

Hierarchical model-based clustering in DLVMs

Overview of contributions

- Generic approach: applies in the framework of DLVMs
- Model selection criterion as a clustering objective

$$\mathrm{ICL}_{ex}(\boldsymbol{Z}) = \log \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\theta}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\pi}$$

Two contributions:

1. Genetic algorithm: greedy maximization w.r.t $\boldsymbol{Z}$
   - Based on selection mechanisms: *mutation* and *cross-over* operators
   - Perform clustering and model selection, return $\boldsymbol{Z}^{(K^\star)}$
   - Bypass inference
2. Hierarchical algorithm: start from $\boldsymbol{Z}^{(K^\star)}$ and merge clusters

$$\boldsymbol{Z}^{(K^\star)} \leq \ldots \leq \boldsymbol{Z}^{(1)}$$

Overview of contributions

- Generic approach: applies in the framework of DLVMs
- Model selection criterion as a clustering objective

$$\mathrm{ICL}_{ex}(\boldsymbol{Z}) = \log \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\theta}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\pi}$$

Two contributions:

1. Genetic algorithm: greedy maximization w.r.t $\boldsymbol{Z}$
   - Based on selection mechanisms: *mutation* and *cross-over* operators
   - Perform clustering and model selection, return $\boldsymbol{Z}^{(K^\star)}$
   - Bypass inference
2. Hierarchical algorithm: start from $\boldsymbol{Z}^{(K^\star)}$ and merge clusters

$$\boldsymbol{Z}^{(K^\star)} \leq \ldots \leq \boldsymbol{Z}^{(1)}$$

Exact integrated classification likelihood

Proposition (Fubini)

With a factorized prior: $p(\boldsymbol{\theta}, \boldsymbol{\pi}) = p(\boldsymbol{\theta} \mid \boldsymbol{\beta})\, p(\boldsymbol{\pi} \mid \boldsymbol{\alpha})$

$$\mathrm{ICL}_{\mathrm{ex}}(\boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \underbrace{\log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\beta})}_{(1)} + \underbrace{\log p(\boldsymbol{Z} \mid \boldsymbol{\alpha})}_{(2)}$$

Conjugate prior for exact (1) available in standard DLVMs

- MoM (Biernacki et al. 2010)
- Binary SBM (Côme et al. 2015)
- GMM (Bertoletti et al. 2015), modulo *informative* prior

Suppose $\boldsymbol{\beta}$ fixed and denote

$$D(\boldsymbol{Z}) := \log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\beta})$$

Exact integrated classification likelihood

Proposition (Fubini)

With a factorized prior: $p(\boldsymbol{\theta}, \boldsymbol{\pi}) = p(\boldsymbol{\theta} \mid \boldsymbol{\beta})\, p(\boldsymbol{\pi} \mid \boldsymbol{\alpha})$

$$\mathrm{ICL}_{ex}(\boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \underbrace{\log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\beta})}_{(1)} + \underbrace{\log p(\boldsymbol{Z} \mid \boldsymbol{\alpha})}_{(2)}$$

Exact expression of (2) with universal prior

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \mathcal{D}_K \left(\boldsymbol{\pi} \mid \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)\right)$$

Set $\forall k, \alpha_k = \alpha$,

- $\alpha = 1$: uniform prior on the simplex
- $\alpha = 1/2$: Jeffreys prior

A sketch of the hierarchical algorithm

Standard agglomerative method

- Starts from $\boldsymbol{Z}^{(K)}$ with $K \leq n$ cluster
- At each stage, find the best fusion w.r.t ICL_{ex}

Problem: a fusion is not always possible

Solution:

- Use α as a regularization parameter
- Extract a set of dominating *nested* partitions

A novel approximation for the $\mathrm{ICL}_{\mathrm{ex}}$

$$\log p(\boldsymbol{Z} \mid \alpha) = \log \frac{\Gamma(K\alpha) \prod_k \Gamma(\alpha + n_k)}{\Gamma(\alpha)^K \Gamma(n + \alpha K)}, \qquad n_k = \sum_i z_{ik}$$

Our proposition: asymptotic of $\log \Gamma$ near 0

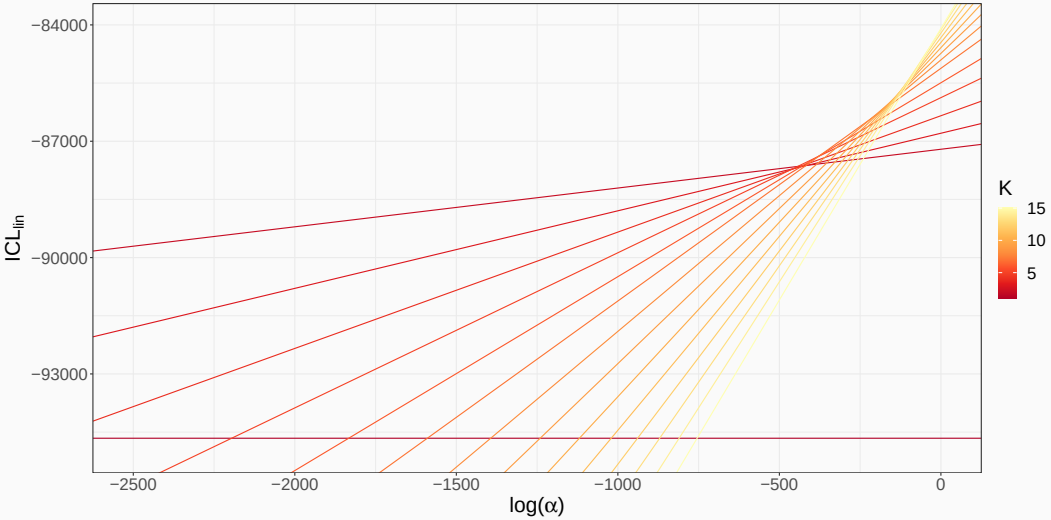
$$\log \Gamma(\alpha) \underset{\alpha \to 0}{\sim} -\log(\alpha)$$

Log-linear ICL

$$\mathrm{ICL}_{lin}(\boldsymbol{Z}, \alpha) := (K-1)\log(\alpha) + I(\boldsymbol{Z})$$

$$I(\boldsymbol{Z}) = D(\boldsymbol{Z}) + \sum_k \log \Gamma(n_k) - \log \Gamma(n) - \log(K)$$

$\mathrm{ICL}_{\mathbf{lin}}$ increasing slope with K



Fusion opportunity at stage (k)

Fixed partition $\boldsymbol{Z}^{(k)}$ with k clusters

Two clusters (g, h): ICL_{lin} change for $g \cup h$?

$$\Delta_{g\cup h}(\alpha) = \mathrm{ICL}_{lin}\left(\boldsymbol{Z}_{g\cup h}, \alpha\right) - \mathrm{ICL}_{lin}\left(\boldsymbol{Z}^{(k)}, \alpha\right)$$

Proposition

$$\forall g \neq h,\ \Delta_{g\cup h}(\alpha) > 0 \iff \log(\alpha) < I(\boldsymbol{Z}_{g\cup h}) - I(\boldsymbol{Z}^{(k)})$$

Regularization parameter: α unlocks fusions

Question: $k(k-1)/2$ fusions, which one is the best ?

$$(g^\star, h^\star) = \arg\max_{g,h} I(\boldsymbol{Z}_{g\cup h})$$

Hierarchy construction and dendrogram representation

Repeat procedure at each stage $\boldsymbol{Z}^{(k)}$

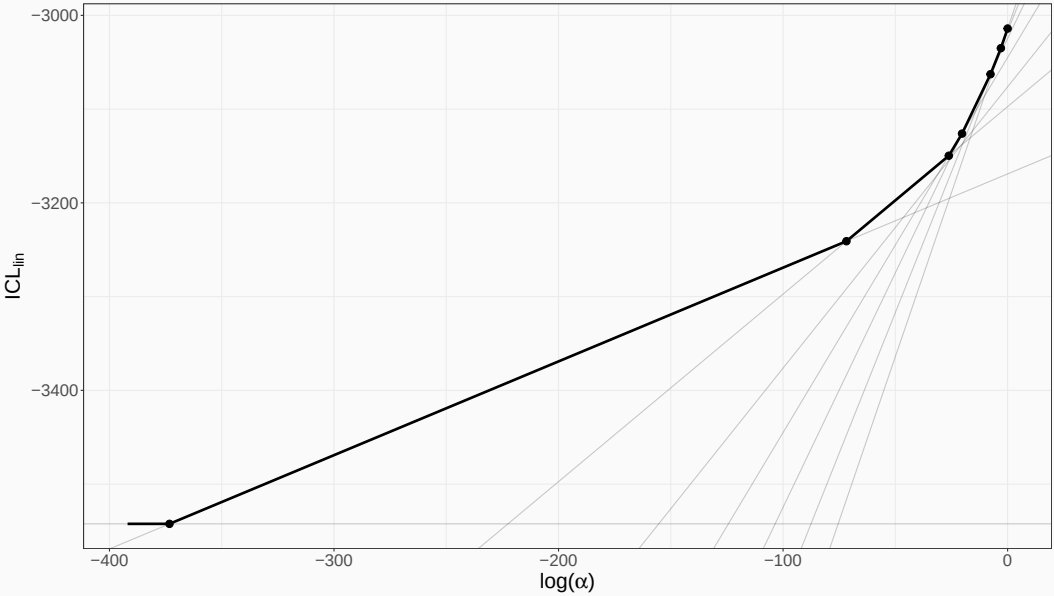
$$\log \alpha^{(k)} := I(\boldsymbol{Z}_{g^\star \cup h^\star}) - I(\boldsymbol{Z}^{(k)})$$

Outputs a hierarchy of partitions

Dendrogram representation:

- $\alpha^{(k)}$ is the amount of regularization needed for the fusion
- Extract a front of dominating partitions on range $[\alpha^{(k-1)}, \alpha^{(k)}]$

A discrete Pareto frontier



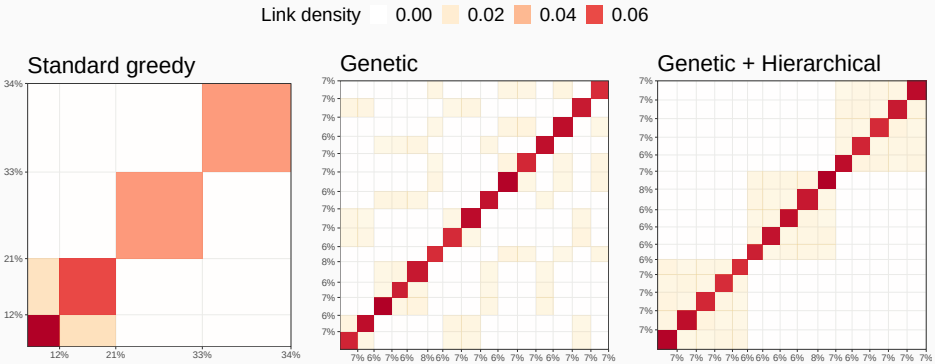
Simulation scenario: graph clustering

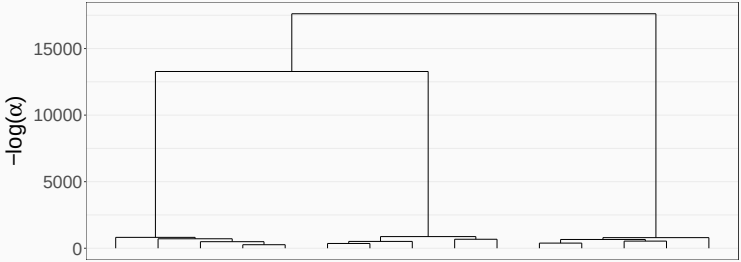
Simulate according SBM with nested structure

- $n = 1500$ nodes, $K = 15$
- 3 clusters composed of 5 smaller ones

Two-step methodology

1. Maximization of ICL_{ex} (genetic algorithm): find $\boldsymbol{Z}^{(K)}$
2. Hierarchy construction using ICL_{lin} and α


Link density
0.00
0.02
0.04
0.06
Standard greedy
Genetic
Genetic + Hierarchical
34%
33%
21%
12%
7%
6%
8%


−log(α)
15000
10000
5000
0

Conclusion

Summary of contributions

Clustering algorithms using DLVMs

- High-dimensional count data (Jouvin et al. 2020)
- High-dimensional continuous data, *preprint*, submitted to journal
- Graph data (and co-clustering), *preprint*, submitted to journal

Applications: medical data, image denoising, graph clustering

Reproducible research: R packages available

- **MoMPCA**
- **FisherEM**
- **greed** (E. Côme)

Perspectives

Clustering categorical data

- survey, census
- Mixture of Multinomial multiple correspondence analysis

$$\boldsymbol{y}_i \sim \mathcal{M}_p(1, \boldsymbol{\theta}_i), \qquad \boldsymbol{\theta}_i = \mathrm{softmax}(\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{x}_i), \qquad \boldsymbol{x}_i \sim \mathsf{GMM}(\mathbb{R}^d)$$

Extensions to BFEM

- Sparsity on $\boldsymbol{U}$ through l_1 penalty $\rightarrow$ **variable selection**
- Other formulations of orthonormal FDA

Exact ICL for GMM $\rightarrow$ handling informative prior

Thank you for your attention !

Questions

Appendix menu

1. Appendix MoMPCA Go to

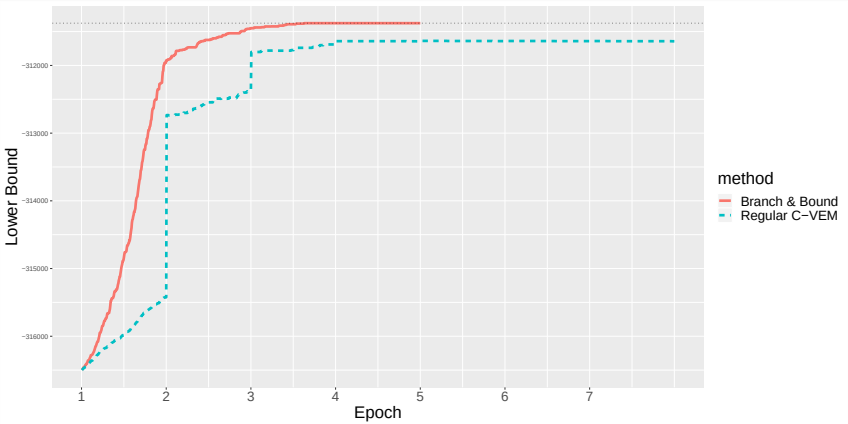
2. Appendix BFEM Go to

3. Appendix HC-ICL Go to

MoMPCA (Appendix)

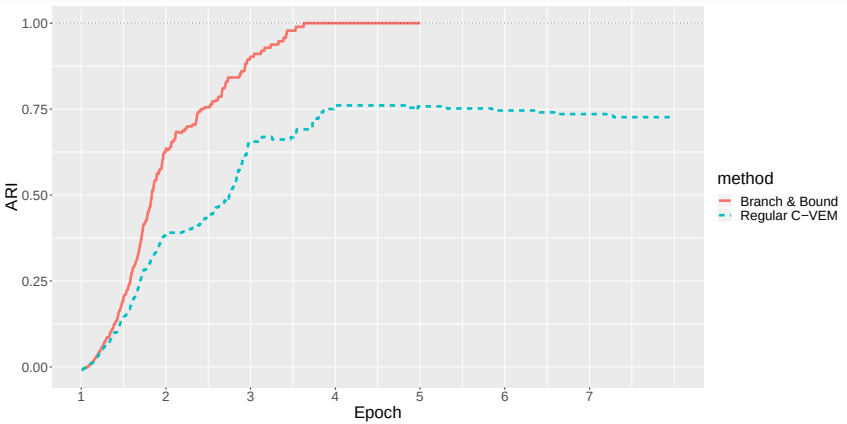
Branch & Bound VS standard C-VEM

Standard C-VEM: no variational inference step after a swap



Branch & Bound VS standard C-VEM

Standard C-VEM: no variational inference step after a swap



MoMPCA: two simulation scenarios

Fixed setting:

$$p = 1000, \qquad K = 6, \qquad d = 4, \qquad \boldsymbol{U}^\star, \qquad \boldsymbol{x}^\star, \qquad \forall i, c_i = 400$$

Scenario 1: noisy structure Goto $n = 400$

$$\boldsymbol{x}_{\epsilon,k} = (1-\epsilon)\boldsymbol{x}_k^\star + \frac{\epsilon}{d}\underbrace{(1,\ldots,1)}_{d}^\top, \quad \epsilon \in [0,1]$$

- $\epsilon = 0 \rightarrow \boldsymbol{x}_{0,k} = \boldsymbol{x}_k^\star$ distribution across topics
- $\epsilon = 1 \rightarrow \boldsymbol{x}_{1,k}$ uniform across topics (no cluster structure)

Scenario 2: small-sample size Goto $\epsilon = 0.2$

$$n = r \times p, \quad r \in [0,1]$$

Metric: *adjusted Rand index* (ARI) the higher, the better

	Topic1	Topic2	Topic3	Topic4	Topic5
$\boldsymbol{x}_1$	0.00	0.01	**0.98**	0.00	0.00
$\boldsymbol{x}_2$	0.19	0.11	0.04	**0.38**	**0.29**
$\boldsymbol{x}_3$	0.13	0.09	0.01	**0.76**	0.00
$\boldsymbol{x}_4$	0.01	0.00	0.01	0.01	**0.97**
$\boldsymbol{x}_5$	0.00	**1.00**	0.00	0.00	0.00
$\boldsymbol{x}_6$	0.05	**0.65**	0.03	0.26	0.01
$\boldsymbol{x}_7$	**0.74**	0.12	0.03	0.11	0.00

Cluster 2 contains micro-calcifications and peaked towards

- Topic4: vocabulary of *in-situ* lesions
- Topic5: vocabulary of benign lesions

Posterior explanation: all samples came from macro-biopsy exams

Bayesian Fisher EM (appendix)

BFEM: two simulation scenarios

Fix $K^\star = 3$, $d^\star = 2$, $n = 900$, $\boldsymbol{\pi}^\star$, $\boldsymbol{\Sigma}_k^\star = \boldsymbol{\Sigma}^\star$

Scenario 1: increasing dimension Goto Fix $\beta_k = \beta = 1$, increase p

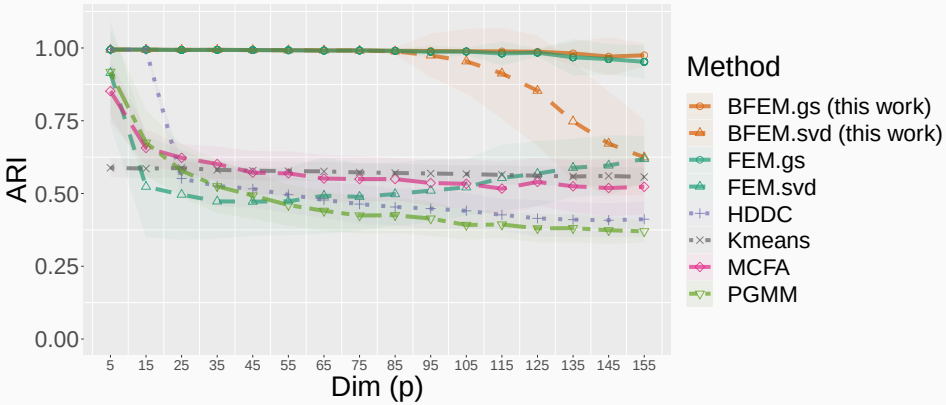
Scenario 2: signal-to-noise Goto Fix $p = 150$, increase β

$$SNR = 10 \times \log_{10}\left(\frac{\mathrm{Tr}\left[\boldsymbol{\Sigma}\right]}{\beta}\right)$$

- $SNR >> 0$: Noiseless regime
- $SNR << 0$: No signal

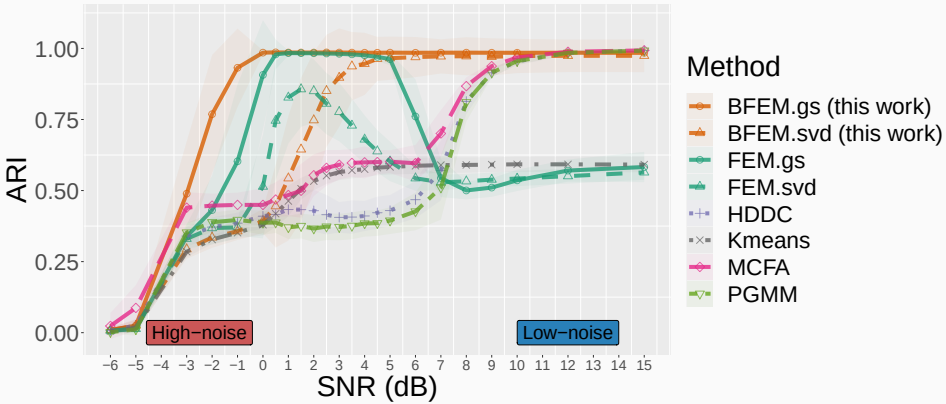
BFEM scenario 1: graph with SVD method

$n = 900, \qquad d = 2, \qquad \beta_k = 1, \qquad p \in \{5, 15, \ldots, 155\}$



BFEM scenario 2: graph with SVD method

$n = 900, \qquad d = 2, \qquad p = 150, \qquad \beta \in \{8, \ldots, 0.8, \ldots, 0.08\}$



BFEM performances on real clustering datasets

		Non-HD models		HD models				
Dataset	p	k-means	Mclust	HDDC	MCFA	PGMM	FEM	BFEM
Iris	4	0.73	0.90	0.90	0.92	**0.94**	0.88	0.90
Wine 27	27	0.90	0.93	0.95	0.96	**0.98**	0.93	0.93
Satellite	**36**	0.53	0.36	0.45	0.43	0.56	0.53	**0.64**
USPS358	**256**	0.64	0	0.35	0.28	0.38	0.66	**0.76**

Application to single image denoising

Context: Observe I_0 blurred with Gaussian white noise

$$I = I_0 + N, \qquad N \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

Patch denoising: decompose I into $f \times f$ sub-images $\boldsymbol{y}_i \in \mathbb{R}^{f^2}$

- GMM prior on the true patch $\boldsymbol{t}_i$

$$\boldsymbol{y}_i = \boldsymbol{t}_i + \epsilon_i \sim \textstyle\sum_k \pi_k \mathcal{N}_{f^2}(\boldsymbol{m}_k, \boldsymbol{\phi}_k + \sigma^2 \boldsymbol{I}_{f^2})$$

- Optimal estimate of $\boldsymbol{t}_i$ w.r.t. quadratic risk

$$\hat{\boldsymbol{t}}_i = \mathbb{E}\left[\boldsymbol{t}_i \mid \boldsymbol{y}_i\right] = \sum_{k=1}^{K} \tau_{ik} \, \mathbb{E}\left[\boldsymbol{t}_i \mid \boldsymbol{y}, z_{ik} = 1, \pi_k, \boldsymbol{\phi}_k, \sigma^2\right]$$

Use BFEM to estimate $\boldsymbol{\pi}$, τ_{ik} and $\boldsymbol{\phi}_k = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{U}^\top$

Clustering and hierarchical clustering in DLVMs (appendix)

Genetic algorithm

Works with ICL_{ex}, α is fixed *e.g.* to 1 or $1/2$

Existing works: find local maxima of ICL_{ex} with greedy swaps and merge

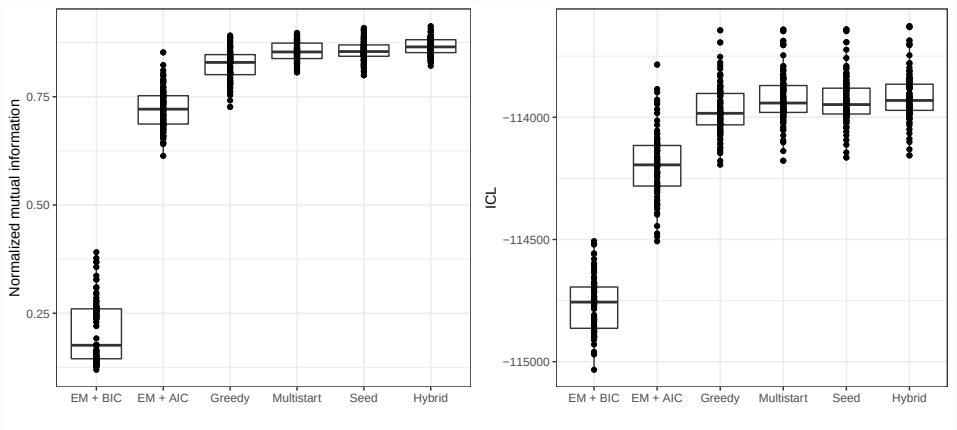
Problem: too many spurious local maxima

Contribution: genetic algorithm to improve exploration

- Recombination: cross-partition operator
- Mutation: cluster split

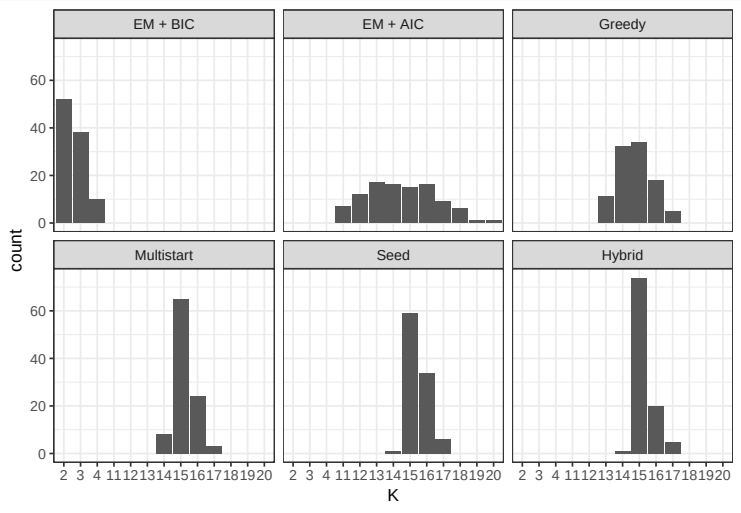
Genetic algorithm: medium-scale MoM

$n = 500, \qquad p = 100, \qquad K = 15, \qquad \boldsymbol{\theta}_k$ peaked toward 10 variables



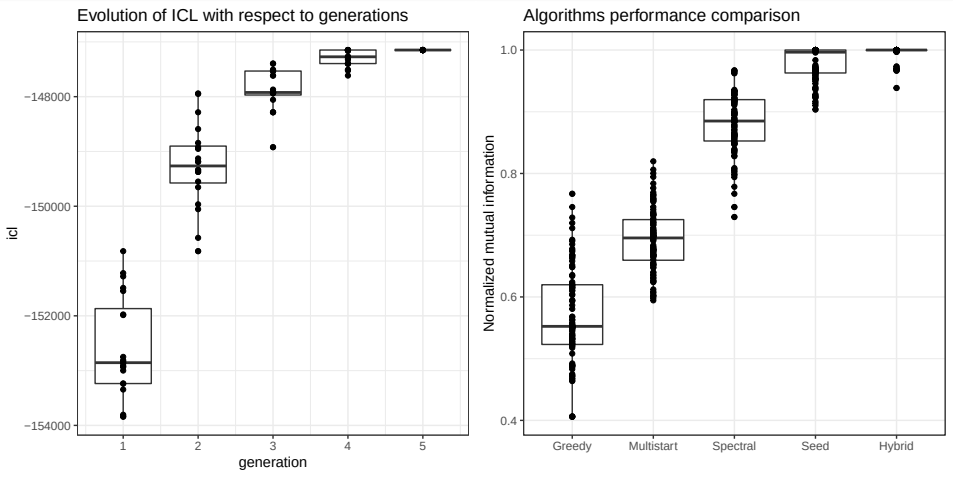
Genetic algorithm: medium-scale MoM

$n = 500, \qquad p = 100, \qquad K = 15, \qquad \boldsymbol{\theta}_k$ peaked toward 10 variables



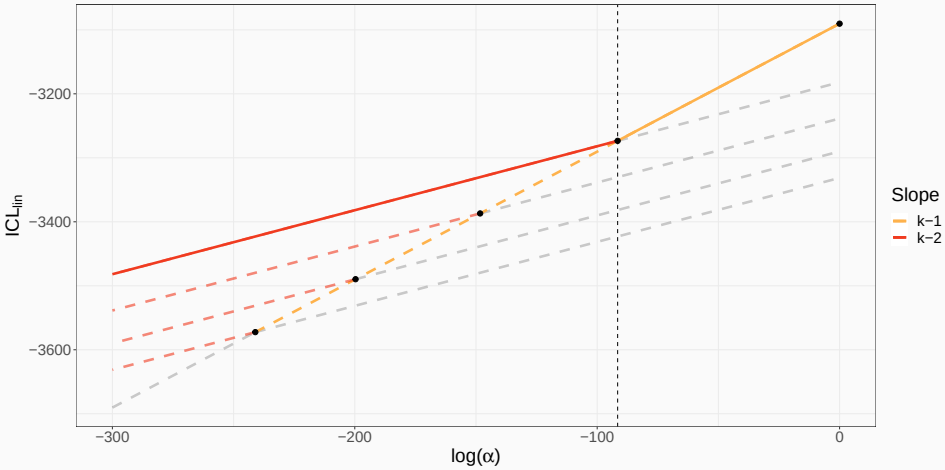
Genetic algorithm: medium-scale SBM

Hierarchical nested SBM with $K = 15$ and $n = 1500$



Choosing best fusion at stage (k)

$$\forall g \neq h,\ \Delta_{g \cup h}(\alpha) > 0 \iff \log(\alpha) < I(\boldsymbol{Z}_{g \cup h}) - I(\boldsymbol{Z}^{(k)})$$



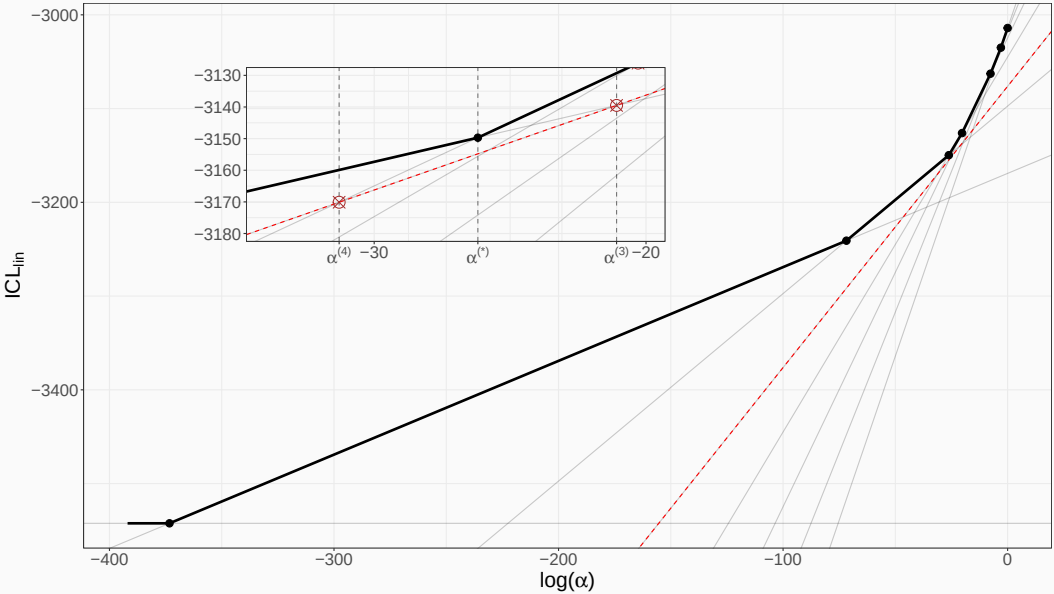
Backtracking step

Question: do we have $\alpha^{(1)} \leq \ldots \leq \alpha^{(K)}$?

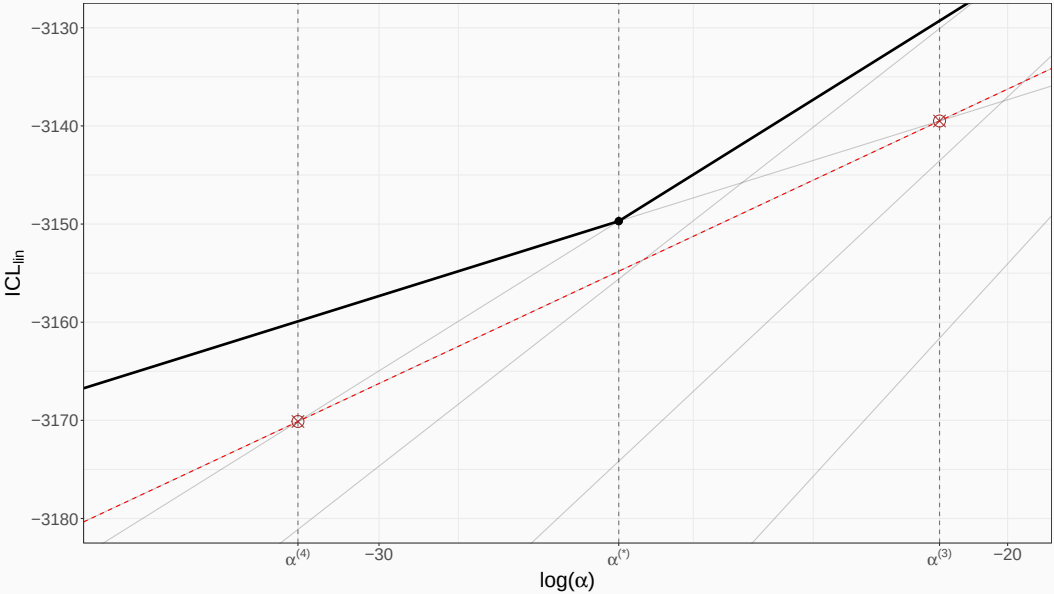
Answer:

- not necessarily, some $\boldsymbol{Z}^{(k)}$ can be nowhere dominant
- easily tracked: corresponds to $\alpha^{(k-1)} \geq \alpha^{(k)}$
- happens when several fusions are possible at once
- compute the new sequence to get the dendrogram representation

Backtrack nowhere dominant partitions



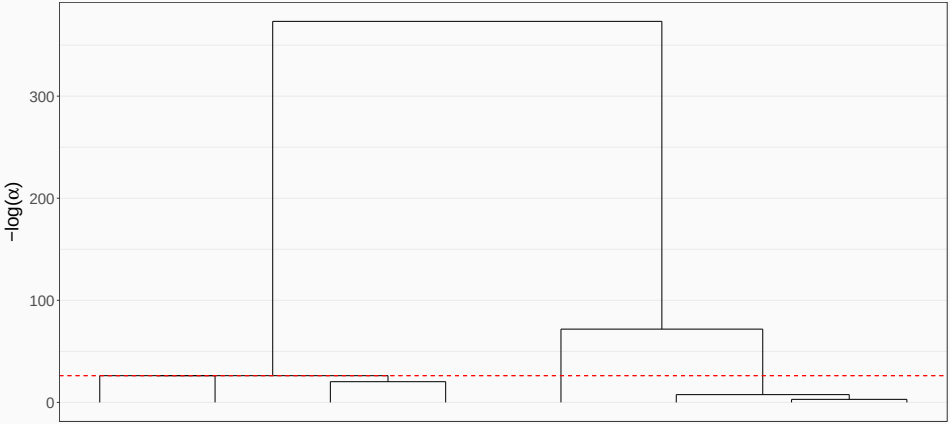
Backtrack nowhere dominant partitions: fusion $5 \to 3$



Dendrogram representation: several fusions at $\alpha^\star$

The partition $\boldsymbol{Z}^{(4)}$ is not in the Pareto front

Remove and compute the intersection $\alpha^\star$ between $\boldsymbol{Z}^{(3)}$ and $\boldsymbol{Z}^{(5)}$.



References

- Bertoletti, Marco, Nial Friel, and Riccardo Rastelli (Aug. 2015). "Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion". In: *METRON* 73.2, pp. 177–199.
- Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.
- Biernacki, Christophe, Gilles Celeux, and Gerard Govaert (2010). "Exact and monte carlo calculations of integrated likelihoods for the latent class model". In: *Journal of Statistical Planning and Inference* 140, pp. 2991–3002.
- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

- Bouveyron, Charles and Camille Brunet (2012). "Simultaneous model-based clustering and visualization in the Fisher discriminative subspace". In: *Statistics and Computing* 22.1, pp. 301–324.
- Buntine, Wray (2002). "Variational extensions to EM and multinomial PCA". In: *European Conference on Machine Learning*. Springer, pp. 23–34.
- Carel, Léna and Pierre Alquier (2017). "Simultaneous dimension reduction and clustering via the NMF-EM algorithm". In: *Advances in Data Analysis and Classification*, pp. 1–30.
- Côme, Etienne and Pierre Latouche (2015). "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood". In: *Statistical Modelling* 15.6, pp. 564–589.
- Fisher, Ronald A (1936). "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2, pp. 179–188.
- Houdard, Antoine, Charles Bouveyron, and Julie Delon (2018). "High-dimensional mixture models for unsupervised image denoising (HDMI)". In: *SIAM Journal on Imaging Sciences* 11.4, pp. 2815–2846.

Jouvin, Nicolas et al. (2020). “Greedy clustering of count data through a mixture of multinomial PCA”. In: *Computational Statistics*.

Yu, Shipeng et al. (2005). “A probabilistic clustering-projection model for discrete data”. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 417–428.