Bayesian Non-Parametric Parsimonious Clustering

F. Chamroukhi, M. Bartcus and H. Glotin

Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France

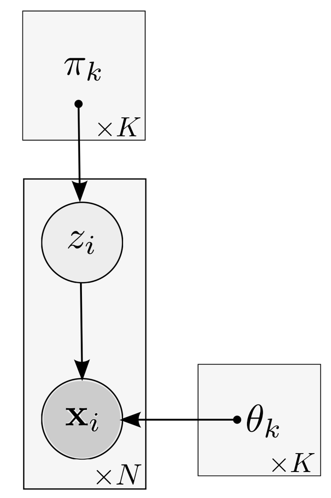
Context and Objectives

- Context: Mixture model-based clustering
- Objectives: Provide a principled approach to:
  - learn the mixture parameters and simultaneously infer the number of clusters from the data
  - provide flexible clusters adapted for different group shapes, orientations, volumes
- Data: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$: n i.i.d observations in $\mathbb{R}^d$
- $\mathbf{z} = (z_1, \ldots, z_n)$: unknown cluster labels; $z_i \in \{1, \ldots, K\}$
- K possibly unknown number of clusters

Finite Gaussian Mixture

- Model: $p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}_k(\mathbf{x}_i|\theta_k)$
- Generative model:

$$\begin{aligned} z_i|\boldsymbol{\pi} &\sim \text{Mult}(\boldsymbol{\pi}) \\ \mathbf{x}_i|\boldsymbol{\theta}_{z_i} &\sim p(.|\boldsymbol{\theta}_{z_i}) \end{aligned}$$

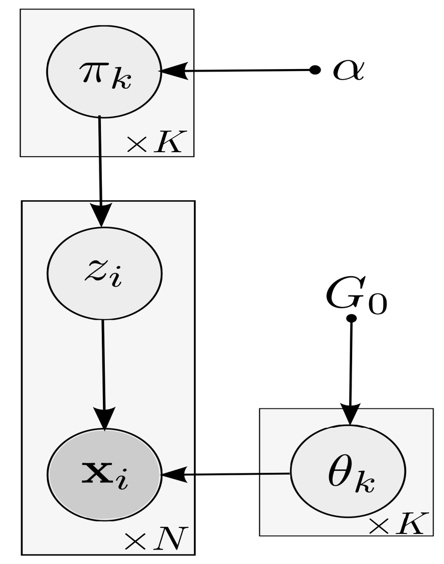


- Likelihood: $p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \, \mathcal{N}_k(\mathbf{x}_i|\theta_k)$
- Learning: e.g, (X)EM - BIC, AIC, ICL ...

Bayesian Gaussian Mixture Model

- Prior: $p(\boldsymbol{\theta})$ (e.g conjugate prior)
- Generative model:

$$\begin{aligned} \boldsymbol{\pi}|\alpha &\sim \text{Dir}(\alpha) \\ z_i|\boldsymbol{\pi} &\sim \text{Mult}(\boldsymbol{\pi}) \\ \boldsymbol{\theta}_{z_i} &\sim G_0 \\ \mathbf{x}_i|\boldsymbol{\theta}_{z_i} &\sim p(.|\boldsymbol{\theta}_{z_i}) \end{aligned}$$

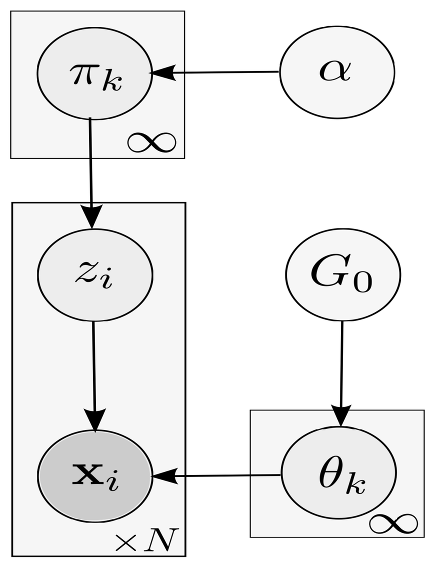


- Posterior (MAP): $p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})$
- Learning: Gibbs - Bayes Factor, ...

Bayesian Non-Parametric Parsimonious GMM

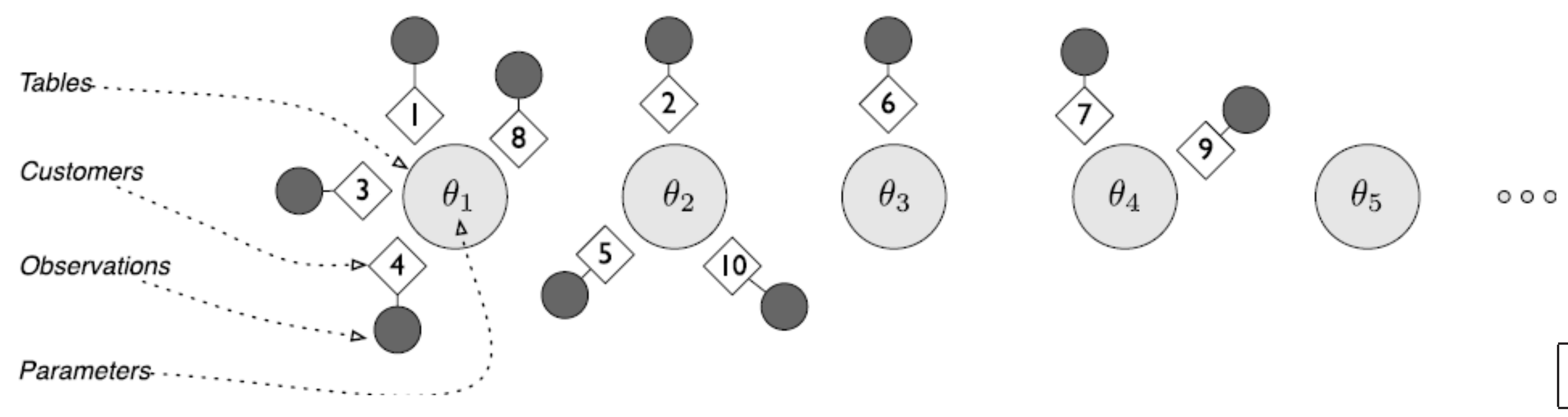
- Infinite GMM: $p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_k \, \mathcal{N}_k(\mathbf{x}_i|\theta_k)$
- Parameters: $\boldsymbol{\theta} = \{\pi_k, \ \boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \ \boldsymbol{\Sigma}_k)\}_{k=1}^{\infty}$
- Prior: add a distribution over the parameters distribution: a Dirichlet Process
- Generative model:

$$\begin{aligned} G|\alpha, G_0 &\sim \text{DP}(\alpha, G_0) \\ \boldsymbol{\theta}_i|G &\sim G \\ \mathbf{x}_i|\boldsymbol{\theta}_i &\sim p(.|\boldsymbol{\theta}_i) \end{aligned}$$



- Learning algorithms: Gibbs, Collapsed Gibbs...

Chinese Restaurant Process (CRP):



- The CRP provides a distribution on the infinite partitions of the data: $p(\mathbf{z}) = p(z_1)p(z_2|z_1)\ldots p(z_n|z_{n-1})$:

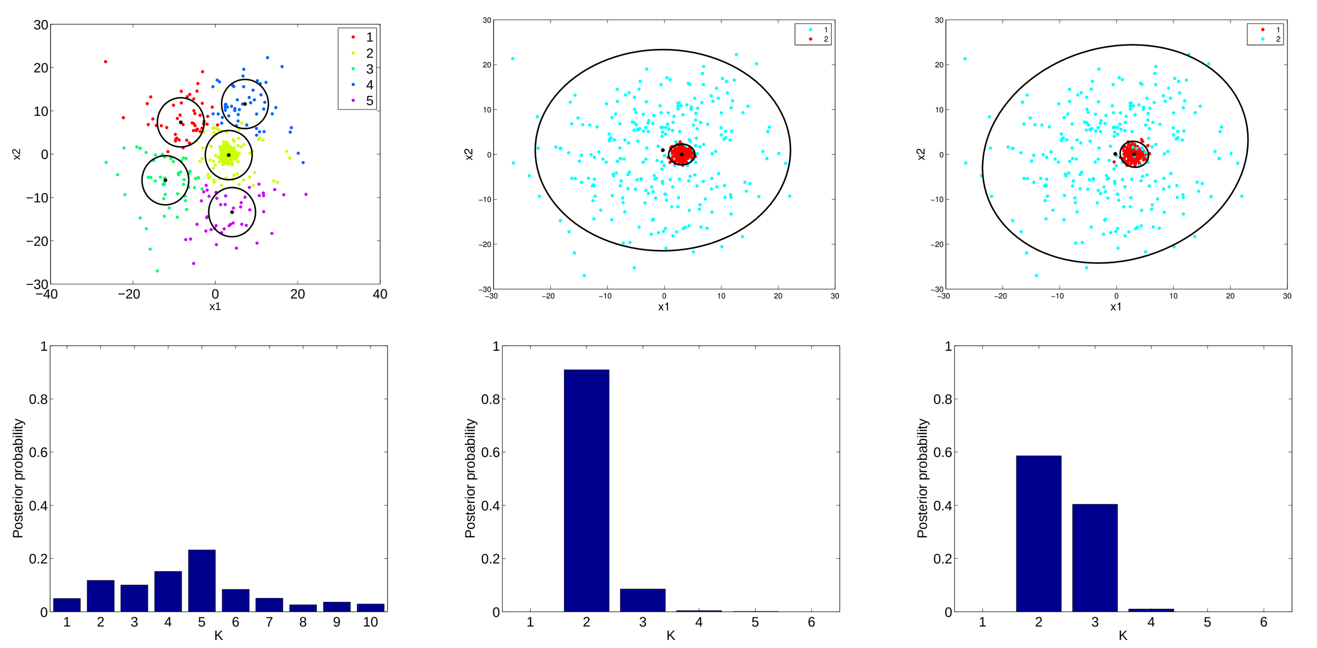
$$\begin{aligned} p(z_i = k|z_1, ..., z_{i-1}) &= \text{CRP}(z_1, ..., z_{i-1}; \alpha) \\ &= \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if} \quad k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & \text{if} \quad k > K_+ \end{cases} \end{aligned}$$

- Parametrization of cov. matrix: $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$
  - λ_k scalar that defines the volume of cluster k
  - $\mathbf{D}_k$ orthogonal matrix which defines the orientation
  - $\mathbf{A}_k$ diagonal matrix with det. 1 defines the shape.

Decomposition	Type	Prior (G_0)	Applied to
$\lambda\mathbf{I}$	Sph.	$\mathcal{IG}$	λ
$\lambda_k\mathbf{I}$	Sph.	$\mathcal{IG}$	λ_k
$\lambda\mathbf{A}$	Diag.	$\mathcal{IG}$	diag. elmnts of $\lambda\mathbf{A}$
$\lambda_k\mathbf{A}$	Diag.	$\mathcal{IG}$	diag. elmnts of $\lambda_k\mathbf{A}$
$\lambda\mathbf{DAD}^T$	Gen.	$\mathcal{IW}$	$\boldsymbol{\Sigma} = \lambda\mathbf{DAD}^T$
$\lambda_k\mathbf{DAD}^T$	Gen.	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\boldsymbol{\Sigma}_0 = \mathbf{DAD}^T$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	Gen.	$\mathcal{IW}$	$\boldsymbol{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

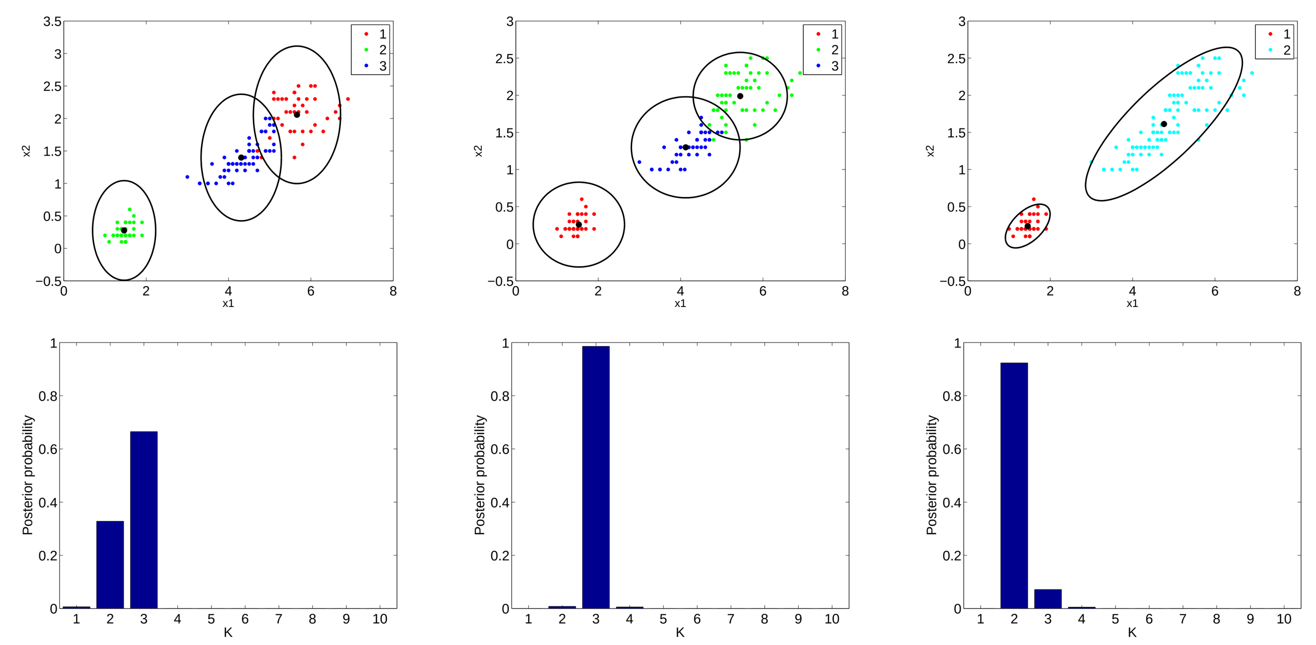
- Learning: Gibbs sampler

Illustration on simulated data



$\lambda\mathbf{I}$ $\lambda_k\mathbf{I}$ $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

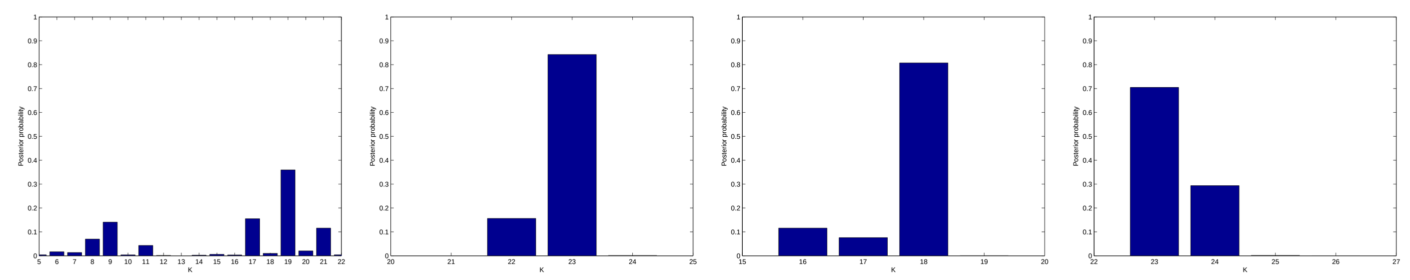
Benchmarks



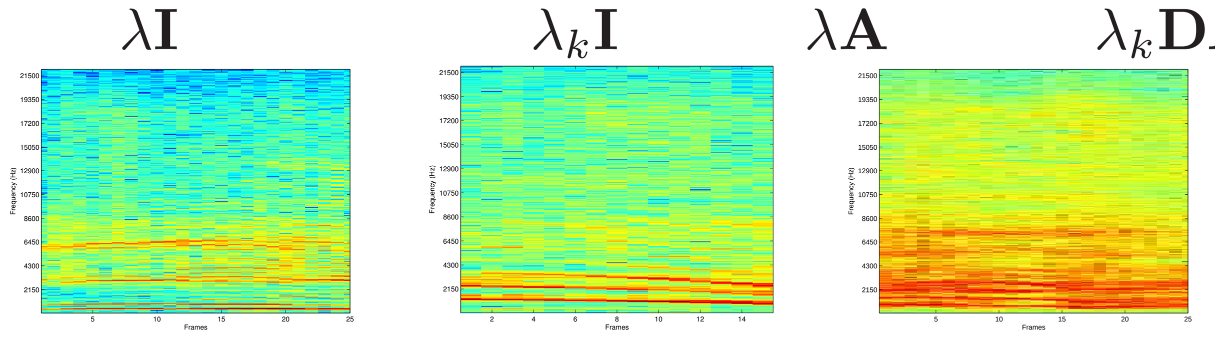
$\lambda_k\mathbf{I}$ $\lambda_k\mathbf{A}$ $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

Whale song decomposition

Learned on 8 minutes of whale signals:



$\lambda\mathbf{I}$ $\lambda_k\mathbf{I}$ $\lambda\mathbf{A}$ $\lambda_k\mathbf{DAD}^T$



song unit 1 song unit 9 song unit 14