A generalized Bayes framework for probabilistic clustering

Tommaso Rigon

Joint work with: Amy Herring and David Dunson

Collegio Carlo Alberto, 2020-05-29

Introduction

- There are roughly two approaches for clustering.

Introduction

- There are roughly two approaches for clustering.

- Model-based clustering often relies on mixture models, i.e.

$$\sum_{k=1}^{K} \xi_k \pi(\boldsymbol{x} \mid \boldsymbol{\theta}_k), \qquad K \geq 1,$$

with $\pi(\boldsymbol{x} \mid \boldsymbol{\theta})$ being a parametric kernel. A representative is a mixture of Gaussians model.

Introduction

- There are roughly two approaches for clustering.
- Model-based clustering often relies on mixture models, i.e.

$$\sum_{k=1}^{K} \xi_k \pi(\boldsymbol{x} \mid \boldsymbol{\theta}_k), \qquad K \geq 1,$$

  with $\pi(\boldsymbol{x} \mid \boldsymbol{\theta})$ being a parametric kernel. A representative is a mixture of Gaussians model.

- Algorithmic clustering is often based on the minimization of loss function, i.e.

$$\text{Cluster solution} = \arg\min_{\boldsymbol{c}} \ell(\boldsymbol{c}; \boldsymbol{X}).$$

  Representatives are the k-means / k-medoids algorithms and generalizations.

Model-based clustering

$$\sum_{k=1}^{K} \xi_k \pi(\boldsymbol{x} \mid \boldsymbol{\theta}_k), \qquad K \geq 1.$$

Model-based clustering

$$\sum_{k=1}^{K} \xi_k \pi(\boldsymbol{x} \mid \boldsymbol{\theta}_k), \qquad K \geq 1.$$

Pro

- Probabilistic interpretation of the partition mechanism.
- Enable uncertainty quantification e.g. within the Bayesian paradigm.

Model-based clustering

$$\sum_{k=1}^{K} \xi_k \pi(\boldsymbol{x} \mid \boldsymbol{\theta}_k), \qquad K \geq 1.$$

Pro

- Probabilistic interpretation of the partition mechanism.
- Enable uncertainty quantification e.g. within the Bayesian paradigm.

Cons

- Despite the remarkable advances, computations are still a huge bottleneck.
- Results are highly misleading if the kernel is misspecified.
- Assuming the existence of a latent partition might be unrealistic.

Loss-based algorithmic clustering

$$\text{Cluster solution} = \arg\min_{\boldsymbol{c}} \ell(\boldsymbol{c}; \boldsymbol{X})$$

Loss-based algorithmic clustering

$$\text{Cluster solution} = \arg\min_{\boldsymbol{c}} \ell(\boldsymbol{c}; \boldsymbol{X})$$

Pro

- Computational efficiency $\rightarrow$ can be used on large / massive datasets.
- Simplicity of the method $\rightarrow$ well-understood and widely used by practitioners.
- Robust algorithms are easy to design.
- Useful tools for summarizing the data.

Loss-based algorithmic clustering

$$\text{Cluster solution} = \arg\min_{\boldsymbol{c}} \ell(\boldsymbol{c}; \boldsymbol{X})$$

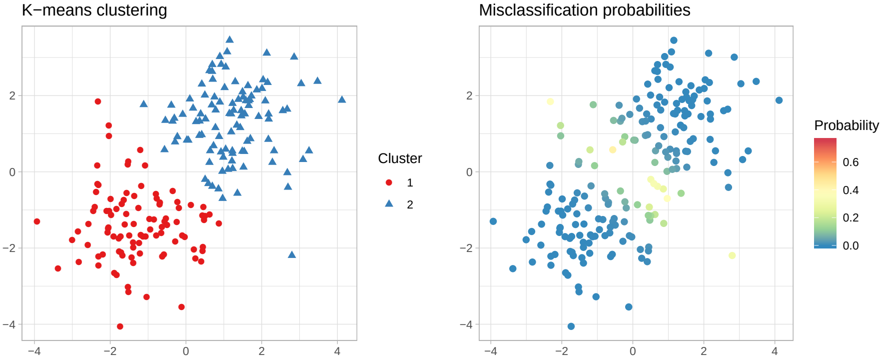
Pro

- Computational efficiency $\rightarrow$ can be used on large / massive datasets.
- Simplicity of the method $\rightarrow$ well-understood and widely used by practitioners.
- Robust algorithms are easy to design.
- Useful tools for summarizing the data.

Cons

- These methods are based on optimizations $\rightarrow$ no probabilistic interpretation.
- No uncertainty quantification.

K-means clustering



Outline of the talk

- We aim at bridging the model-based and loss-based approaches, inheriting the advantages of both.
- We rely on a generalized Bayes theorem which has a clear and coherent justification.
- We propose a large class of models closely related to product partition models.
- We provide uncertainty quantification for most loss-based clustering methods, including k-means.

Gibbs posteriors

- Bayesian inference is based on

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{X} \mid \boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})\pi(\boldsymbol{X} \mid \boldsymbol{\theta})d\boldsymbol{\theta}},$$

where $\pi(\boldsymbol{\theta})$ is the prior, $\pi(\boldsymbol{X} \mid \boldsymbol{\theta})$ is the likelihood, and $\boldsymbol{\theta}$ is a parameter.

Gibbs posteriors

- Bayesian inference is based on

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{X} \mid \boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})\pi(\boldsymbol{X} \mid \boldsymbol{\theta}) d\boldsymbol{\theta}},$$

where $\pi(\boldsymbol{\theta})$ is the prior, $\pi(\boldsymbol{X} \mid \boldsymbol{\theta})$ is the likelihood, and $\boldsymbol{\theta}$ is a parameter.

- Generalized Bayesian inference is based on

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{\pi(\boldsymbol{\theta}) \exp\{-\lambda \ell(\boldsymbol{\theta}; \boldsymbol{X})\}}{\int \pi(\boldsymbol{\theta}) \exp\{-\lambda \ell(\boldsymbol{\theta}; \boldsymbol{X})\} d\boldsymbol{\theta}}, \qquad \lambda > 0,$$

where $\pi(\boldsymbol{\theta})$ is the prior, $\ell(\boldsymbol{\theta}; \boldsymbol{X})$ is a loss function, and $\boldsymbol{\theta}$ is a parameter.

- The latter distribution is called Gibbs posterior.

Model-based Bayesian clustering

- Let $\boldsymbol{x}_i = (x_{i1}, \dots, x_{id})^\intercal$ $i = 1, \dots, n$ be a vector of observations on $\mathbb{X} \subseteq \mathbb{R}^d$ and let $\boldsymbol{X}$ be the collection of all the data points.
- Let $\boldsymbol{C} = (C_1, \dots, C_K)$ be a cluster arrangement and $\boldsymbol{c} = (c_1, \dots, c_n)$ be the associated indicators.
- Let $\boldsymbol{X}_k = \{\boldsymbol{x}_i : i \in C_k\}$ be the observations $\boldsymbol{x}_i$ belonging to the C_k cluster.

A Bayesian mixture model is based on the standard posterior

$$\pi(\boldsymbol{c} \mid \boldsymbol{X}) \propto \pi(\boldsymbol{c}) \prod_{k=1}^{K} \left[\int_{\Theta} \prod_{i \in C_k} \pi(\boldsymbol{x}_i \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right].$$

Generalized Bayes product partition models (GB-PPM)

A Generalized Bayes product partition model is based on the Gibbs posterior

$$\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \pi(\boldsymbol{c}) \prod_{k=1}^{K} \rho(C_k; \lambda, \boldsymbol{X}_k) \propto \prod_{k=1}^{K} \exp\left\{-\lambda \sum_{i \in C_k} \mathcal{D}(\boldsymbol{x}_i; \boldsymbol{X}_k)\right\},$$

with $\boldsymbol{c} : |\boldsymbol{C}| = K$ and $\lambda > 0$.

Generalized Bayes product partition models (GB-PPM)

A Generalized Bayes product partition model is based on the Gibbs posterior

$$\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \pi(\boldsymbol{c}) \prod_{k=1}^{K} \rho(C_k; \lambda, \boldsymbol{X}_k) \propto \prod_{k=1}^{K} \exp\left\{-\lambda \sum_{i \in C_k} \mathcal{D}(\boldsymbol{x}_i; \boldsymbol{X}_k)\right\},$$

with $\boldsymbol{c} : |\boldsymbol{C}| = K$ and $\lambda > 0$.

- The term $\rho(C_k; \lambda, \boldsymbol{X}_k)$ is the cohesion associated to the kth cluster.
- The function $\mathcal{D}(\boldsymbol{x}_i; \boldsymbol{X}_k)$ measures the discrepancy of the ith unit from the kth cluster.
- The uniform prior $\pi(\boldsymbol{c}) \propto 1$ is employed. This is a proper prior, since the partition space is finite.

Foundations of Gibbs posteriors

- *"Isn't this a Bayesian heresy?"* — A colleague.
- Gibbs posteriors have been widely used since the late 90's.
- They were mainly motivated by the PAC-Bayesian approach, which partially clarifies their interpretation.

Foundations of Gibbs posteriors

- *"Isn't this a Bayesian heresy?"* — A colleague.
- Gibbs posteriors have been widely used since the late 90's.
- They were mainly motivated by the PAC-Bayesian approach, which partially clarifies their interpretation.

- The rigorous foundations of Gibbs posteriors have been recently discussed in Bissiri, Holmes, & Walker (2016). JRSS-B.

The target of a GB-PPM

- The target of a GB-PPM is the optimal partition

$$\boldsymbol{c}_{\mathrm{OPT}} = \arg\min_{\boldsymbol{c}} \mathbb{E}_{\pi_0}\{\ell(\boldsymbol{c};\boldsymbol{X})\} = \arg\min_{\boldsymbol{c}:|\boldsymbol{C}|=K} \sum_{k=1}^{K}\sum_{i\in C_k} \mathbb{E}_{\pi_0}\left\{\mathcal{D}(\boldsymbol{x}_i;\boldsymbol{X}_k)\right\},$$

where $\pi_0(\boldsymbol{X})$ is the unknown data generating process.

The target of a GB-PPM

- The target of a GB-PPM is the optimal partition

$$\boldsymbol{c}_{\text{OPT}} = \arg\min_{\boldsymbol{c}} \mathbb{E}_{\pi_0}\{\ell(\boldsymbol{c}; \boldsymbol{X})\} = \arg\min_{\boldsymbol{c}:|\boldsymbol{C}|=K} \sum_{k=1}^{K} \sum_{i \in C_k} \mathbb{E}_{\pi_0}\{\mathcal{D}(\boldsymbol{x}_i; \boldsymbol{X}_k)\},$$

where $\pi_0(\boldsymbol{X})$ is the unknown data generating process.

Key concepts

- Gibbs posteriors quantify the uncertainty about the optimal and unknown $\boldsymbol{c}_{\text{OPT}}$.
- We are not assuming the existence of a latent partition in the generating mechanism.
- $\boldsymbol{c}_{\text{OPT}}$ represent an optimal summary of the data.

Derivation of Gibbs posteriors

- A posterior ν_1 is a better candidate than ν_2 if $\mathscr{L}(\nu_1) \le \mathscr{L}(\nu_2)$, with

$$\mathscr{L}\{\nu(\boldsymbol{c})\} = \lambda \mathbb{E}_\nu \left\{\ell(\boldsymbol{c}; \boldsymbol{X})\right\} + \mathrm{KL}\{\nu(\boldsymbol{c}) \mid\mid \pi(\boldsymbol{c})\},$$

  being a loss function on the space of conditional distributions.

- The optimal posterior is the one minimizing the loss $\mathscr{L}$.

Derivation of Gibbs posteriors

- A posterior ν_1 is a better candidate than ν_2 if $\mathscr{L}(\nu_1) \leq \mathscr{L}(\nu_2)$, with

$$\mathscr{L}\{\nu(\boldsymbol{c})\} = \lambda \mathbb{E}_\nu\{\ell(\boldsymbol{c}; \boldsymbol{X})\} + \mathrm{KL}\{\nu(\boldsymbol{c}) \,||\, \pi(\boldsymbol{c})\},$$

  being a loss function on the space of conditional distributions.

- The optimal posterior is the one minimizing the loss $\mathscr{L}$.

- The loss $\mathscr{L}$ balances the proximity to the data and the closeness to the prior.

- When $\lambda \to \infty$ the minimizer of $\mathscr{L}$ is the point mass $\delta_{\hat{\boldsymbol{c}}_{\mathrm{OPT}}}$, where

$$\hat{\boldsymbol{c}}_{\mathrm{OPT}} = \arg\min_{\boldsymbol{c}} \ell(\boldsymbol{c}; \boldsymbol{X}),$$

  is the empirical version of the optimal partition $\boldsymbol{c}_{\mathrm{OPT}}$.

- When $\lambda \to 0$ the minimizer of $\mathscr{L}$ coincides with the prior distribution.

Derivation of Gibbs posteriors (cont'd)

- Key result 1: Our GB-PPM minimize the loss $\mathscr{L}$ for general values of $\lambda > 0$, that is
$$\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) = \arg\min_{\nu} \mathscr{L}\{\nu(\boldsymbol{c})\}.$$
Hence, $\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ is the best posterior for quantifying the uncertainty about optimal partition $\boldsymbol{c}_{\text{OPT}}$.

Derivation of Gibbs posteriors (cont'd)

- Key result 1: Our GB-PPM minimize the loss $\mathscr{L}$ for general values of $\lambda > 0$, that is
$$\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) = \arg\min_{\nu} \mathscr{L}\{\nu(\boldsymbol{c})\}.$$
Hence, $\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ is the best posterior for quantifying the uncertainty about optimal partition $\boldsymbol{c}_{\text{OPT}}$.

- Key result 2: The loss $\mathscr{L}$ is not arbitrary, because is the only one satisfying natural coherency conditions (Bissiri et al., 2016).

Derivation of Gibbs posteriors (cont'd)

- Key result 1: Our GB-PPM minimize the loss $\mathscr{L}$ for general values of $\lambda > 0$, that is
$$\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) = \arg\min_{\nu} \mathscr{L}\{\nu(\boldsymbol{c})\}.$$
Hence, $\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ is the best posterior for quantifying the uncertainty about optimal partition $\boldsymbol{c}_{\mathrm{OPT}}$.

- Key result 2: The loss $\mathscr{L}$ is not arbitrary, because is the only one satisfying natural coherency conditions (Bissiri et al., 2016).

- Remark: Gibbs posteriors are not a pseudo-Bayes approach nor an approximate Bayesian procedure. They are coherent Bayesian updates.

Point estimation

- Although several alternative exist, the MAP is a sensible point estimate.

Point estimation

- Although several alternative exist, the MAP is a sensible point estimate.

A (trivial) Proposition

Let $\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM. Then,

$$\hat{\boldsymbol{c}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{c}} \pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) = \arg \min_{\boldsymbol{c}\,:\,|\boldsymbol{C}|=K} \ell(\boldsymbol{c}; \boldsymbol{X}).$$

- The $\hat{\boldsymbol{c}}_{\mathrm{MAP}}$ is the value minimizing a loss.
- Well-known algorithms can be used for finding the MAP, such as k-means.
- Note that the estimate $\hat{\boldsymbol{c}}_{\mathrm{MAP}}$ does not depend on λ. This is not the case for general point estimates.

Posterior inference

- Posterior inference is conducted through a Gibbs sampling.

Posterior inference

- Posterior inference is conducted through a Gibbs sampling.

Theorem

Let $\pi(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM. Then, the conditional distribution of c_i given $\boldsymbol{c}_{-i}$ is

$$\mathbb{P}(c_i = k \mid \boldsymbol{c}_{-i}, \lambda, \boldsymbol{X}) \propto \frac{\rho(C_k; \lambda, \boldsymbol{X}_k)}{\rho(C_{k,-i}; \lambda, \boldsymbol{X}_{k,-i})}$$
$$\propto \exp\left\{-\lambda\left[\sum_{i' \in C_k} \mathcal{D}(\boldsymbol{x}_{i'}; \boldsymbol{X}_k) - \sum_{i' \in C_{k,-i}} \mathcal{D}(\boldsymbol{x}_{i'}; \boldsymbol{X}_{k,-i})\right]\right\},$$

for $k = 1, \dots, K$ and for any partition $\boldsymbol{c} : |\boldsymbol{C}| = K$.

- The ith unit is likely to be allocated in the kth cluster if the cohesion of the newly created cluster is higher than the old cohesion.

GB-PPMs with Bregman cohesions

GB-PPM with Bregman cohesions

Bregman divergence

Let $\varphi : \mathbb{X} \to \mathbb{R}$ be a strictly convex function defined on a convex set $\mathbb{X} \subseteq \mathbb{R}^d$, such that φ is differentiable on the relative interior of $\mathbb{X}$. Then

$$\mathcal{D}_\varphi(\boldsymbol{x}; \boldsymbol{\mu}) = \varphi(\boldsymbol{x}) - [\varphi(\boldsymbol{\mu}) + (\boldsymbol{x} - \boldsymbol{\mu})^\intercal \nabla \varphi(\boldsymbol{\mu})],$$

is a Bregman divergence, for any $\boldsymbol{x} \in \mathbb{X}$ and any $\boldsymbol{\mu}$ in the relative interior of $\mathbb{X}$.

GB-PPM with Bregman cohesions

Bregman divergence

Let $\varphi : \mathbb{X} \to \mathbb{R}$ be a strictly convex function defined on a convex set $\mathbb{X} \subseteq \mathbb{R}^d$, such that φ is differentiable on the relative interior of $\mathbb{X}$. Then

$$\mathcal{D}_\varphi(\boldsymbol{x}; \boldsymbol{\mu}) = \varphi(\boldsymbol{x}) - [\varphi(\boldsymbol{\mu}) + (\boldsymbol{x} - \boldsymbol{\mu})^\intercal \nabla \varphi(\boldsymbol{\mu})],$$

is a Bregman divergence, for any $\boldsymbol{x} \in \mathbb{X}$ and any $\boldsymbol{\mu}$ in the relative interior of $\mathbb{X}$.

- A Bregman divergence $\mathcal{D}_\varphi(\boldsymbol{x}; \boldsymbol{\mu})$ is non-negative.
- The discrepancy between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ is measured as the difference between $\varphi(\boldsymbol{x})$ and the value of its tangent hyperplane at $\boldsymbol{\mu}$, evaluated at $\boldsymbol{x}$.
- The squared Euclidean distance (k-means), the Mahalanobis distance, and the KL are instances of Bregman divergences.

GB-PPM with Bregman cohesions (cont'd)

Let $\pi_\varphi(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM. We will say it has Bregman cohesions if

$$\pi_\varphi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \rho(C_k; \lambda, \boldsymbol{X}_k) = \prod_{k=1}^{K} \exp\left\{-\lambda \sum_{i \in C_k} \mathcal{D}_\varphi(\boldsymbol{x}_i; \bar{\boldsymbol{x}}_k)\right\},$$

$\boldsymbol{c} : |\boldsymbol{C}| = K$, where $\mathcal{D}_\varphi(\boldsymbol{x}; \boldsymbol{\mu})$ is a Bregman divergence.

GB-PPM with Bregman cohesions (cont'd)

Let $\pi_\varphi(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM. We will say it has Bregman cohesions if

$$\pi_\varphi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \rho(C_k; \lambda, \boldsymbol{X}_k) = \prod_{k=1}^{K} \exp\left\{-\lambda \sum_{i \in C_k} \mathcal{D}_\varphi(\boldsymbol{x}_i; \bar{\boldsymbol{x}}_k)\right\},$$

$\boldsymbol{c} : |\boldsymbol{C}| = K$, where $\mathcal{D}_\varphi(\boldsymbol{x}; \boldsymbol{\mu})$ is a Bregman divergence.

- The arithmetic mean $\bar{\boldsymbol{x}}_k$ is not an arbitrary choice, because

$$\bar{\boldsymbol{x}}_k = \arg\max_{\boldsymbol{\mu}} \exp\left\{-\lambda \sum_{i \in C_k} \mathcal{D}_\varphi(\boldsymbol{x}_i; \boldsymbol{\mu})\right\},$$

  i.e. is the value maximizing the cohesion.

- The Bregman divergence $\mathcal{D}_\varphi(\boldsymbol{x}_i; \bar{\boldsymbol{x}}_k)$ evaluated at $\bar{\boldsymbol{x}}_k$ is not always well-defined, but there are easy fixes to this issue.

The Bregman k-means algorithm

Bregman k-means (Banerjee et al., 2005)

Choose K and a set of initial centroids $\boldsymbol{m}_1, \dots, \boldsymbol{m}_K$.

Until the centroids stabilize:

for $i = 1, \dots, n$ **do**

Set the cluster indicator c_i equal to k, so that $\mathcal{D}_\varphi(\boldsymbol{x}_i; \boldsymbol{m}_k)$ is minimum.

for $k = 1, \dots, K$ **do**

Let $\boldsymbol{m}_k$ be equal to the arithmetic mean $\bar{\boldsymbol{x}}_k$ of the subjects belonging to group k.

return $\hat{\boldsymbol{c}}_{\text{MAP}} = (c_1, \dots, c_n)$.

- The Bregman k-means monotonically decreases the loss function, and it reaches a local optimum in a finite number of steps.

Connection with exponential dispersion families

Exponential dispersion family (Jørgensen, 1987)

Let $\pi(\boldsymbol{x} \mid \lambda)$ be a density function on $\mathbb{X} \subseteq \mathbb{R}^d$ indexed by $\lambda > 0$. Then, the class of densities

$$\pi_{\text{ED}}(\boldsymbol{x} \mid \boldsymbol{\theta}, \lambda) = \pi(\boldsymbol{x} \mid \lambda) e^{\lambda[\boldsymbol{\theta}^\intercal \boldsymbol{x} - \kappa(\boldsymbol{\theta})]}, \qquad \boldsymbol{\theta} \in \Theta, \quad \lambda \in \Lambda,$$

is called exponential dispersion family.

- If $\boldsymbol{x} \sim \pi_{\text{ED}}(\boldsymbol{x} \mid \boldsymbol{\theta}, \lambda)$, then

$$\mathbb{E}(\boldsymbol{x}) = \mu(\boldsymbol{\theta}), \qquad \text{Var}(\boldsymbol{x}) = \frac{1}{\lambda} \boldsymbol{V}.$$

- The function $\mu(\cdot)$ is injective and $\boldsymbol{V}$ is a $d \times d$ matrix not depending on λ.

- There is a one-to-one correspondence between the natural parametrization $\boldsymbol{\theta}$ and the mean parametrization $\boldsymbol{\mu} = \mu(\boldsymbol{\theta})$.

Connection with exponential dispersion families (cont'd)

Theorem

Let $\pi_{\mathrm{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM of the form

$$\pi_{\mathrm{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \pi(\boldsymbol{x}_i \mid \lambda) \exp\left\{\lambda[\hat{\boldsymbol{\theta}}_k^{\intercal} \boldsymbol{x}_i - \kappa(\hat{\boldsymbol{\theta}}_k)]\right\},$$

where $\hat{\boldsymbol{\theta}}_k = \theta(\bar{\boldsymbol{x}}_k) = \arg\max_{\boldsymbol{\theta}_k} \prod_{i \in C_k} \pi_{\mathrm{ED}}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k, \lambda)$.

Connection with exponential dispersion families (cont'd)

Theorem

Let $\pi_{\text{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM of the form

$$\pi_{\text{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \pi(\boldsymbol{x}_i \mid \lambda) \exp\left\{\lambda[\hat{\boldsymbol{\theta}}_k^\intercal \boldsymbol{x}_i - \kappa(\hat{\boldsymbol{\theta}}_k)]\right\},$$

where $\hat{\boldsymbol{\theta}}_k = \theta(\bar{\boldsymbol{x}}_k) = \arg\max_{\boldsymbol{\theta}_k} \prod_{i \in C_k} \pi_{\text{ED}}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k, \lambda)$. Then, there exists a GB-PPM with Bregman cohesion such that

$$\pi_{\text{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) = \pi_{\varphi}(\boldsymbol{c} \mid \lambda, \boldsymbol{X}), \qquad \boldsymbol{c} : |\boldsymbol{C}| = K.$$

Connection with exponential dispersion families (cont'd)

Theorem

Let $\pi_{\text{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM of the form

$$\pi_{\text{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \pi(\boldsymbol{x}_i \mid \lambda) \exp\left\{\lambda[\hat{\boldsymbol{\theta}}_k^{\intercal}\boldsymbol{x}_i - \kappa(\hat{\boldsymbol{\theta}}_k)]\right\},$$

where $\hat{\boldsymbol{\theta}}_k = \theta(\bar{\boldsymbol{x}}_k) = \arg\max_{\boldsymbol{\theta}_k} \prod_{i \in C_k} \pi_{\text{ED}}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k, \lambda)$. Then, there exists a GB-PPM with Bregman cohesion such that

$$\pi_{\text{ED}}(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) = \pi_{\varphi}(\boldsymbol{c} \mid \lambda, \boldsymbol{X}), \qquad \boldsymbol{c} : |\boldsymbol{C}| = K.$$

- The λ parameter is proportional to the within-cluster precision.
- Key result: this probabilistic interpretation simplifies the estimation / elicitation of λ.
- The GB-PPM $\pi_{\varphi}(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ can be also regarded as the Bayesian update of a profile likelihood.

GB-PPMs with pairwise dissimilarities

GB-PPM with pairwise dissimilarities

- Let $\mathbb{X} = \mathbb{R}^d$ and let $||\boldsymbol{x}||_p = (|x_1|^p + \cdots + |x_d|^p)^{1/p}$ be the L^p norm.
- A general measure of dissimilarity is

$$\gamma(||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_p^p), \qquad \boldsymbol{x}_i, \boldsymbol{x}_{i'} \in \mathbb{R}^d, \quad p \geq 1,$$

  for some increasing function $\gamma(\cdot)$ such that $\gamma(0) = 0$.

Let $\pi_\gamma(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM with covariate space $\mathbb{X} = \mathbb{R}^d$. We will say it has average dissimilarity cohesions if

$$\pi_\gamma(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^K \exp\left\{-\frac{\lambda}{2}\sum_{i \in C_k} \frac{1}{n_k} \sum_{i' \in C_k} \gamma(||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_p^p)\right\}, \qquad \boldsymbol{c} : |\boldsymbol{C}| = K,$$

with $p \geq 1$ and with $\gamma(\cdot)$ being an increasing function such that $\gamma(0) = 0$.

The k-dissimilarities algorithm

K-dissimilarities

Randomly allocate the indicators $c_1, \dots, c_n$ into K sets.

Until the partition stabilizes:

for $i = 1, \dots, n$ **do**

Allocate the indicator c_i, given the others $\boldsymbol{c}_{-i}$, to the k cluster, so that

$$\sum_{i' \in C_k} \mathcal{D}_\gamma(\boldsymbol{x}_{i'}; \boldsymbol{X}_k) - \sum_{i' \in C_{k,-i}} \mathcal{D}_\gamma(\boldsymbol{x}_{i'}; \boldsymbol{X}_{k,-i})$$

is minimum. Recursive formulas are available.

return $\hat{\boldsymbol{c}}_{\mathrm{MAP}} = (c_1, \dots, c_n)$.

- The k-dissimilarities monotonically decreases the loss function, and it reaches a local optimum in a finite number of steps.

Connection with L^p spherical distributions

L^p spherical distributions (Gupta & Song, 1997)

A random vector $\boldsymbol{x} \in \mathbb{R}^d$ follows a L^p spherical distribution if its density function can be written as

$$\pi_{\mathrm{SP}}(\boldsymbol{x}) = g(||\boldsymbol{x}||_p^p),$$

for some measurable function $g : \mathbb{R}_+ \to \mathbb{R}_+$.

- The class of L^p spherical distributions includes e.g. the multivariate Gaussian, the multivariate Laplace and the multivariate Student's t.
- The family is indexed by the function g, which is sometimes called density generator.

Connection with L^p spherical distributions (cont'd)

Theorem

Let $\pi_\gamma(\boldsymbol{c} \mid \lambda, \boldsymbol{X})$ be a GB-PPM with average dissimilarities. If

$$\int_{\mathbb{R}_+} r^{d-1} \exp\left\{-\frac{\lambda}{2}\gamma(r^p)\right\} \mathrm{d}r < \infty,$$

then there exists an L^p spherical distribution on $\mathbb{R}^d$ such that

$$\pi_\gamma(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \left[\prod_{i' \in C_k} \pi_{\mathrm{SP}}(\boldsymbol{x}_i - \boldsymbol{x}_{i'} \mid \lambda)\right]^{1/n_k},$$

where $\pi_{\mathrm{SP}}(\boldsymbol{x}_i - \boldsymbol{x}_{i'} \mid \lambda) \propto \exp\left\{-\lambda/2\gamma(||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_p^p)\right\}$ for any $i \in C_k$ and $i' \in C_k$.

- Key result: as before, this probabilistic interpretation simplifies the estimation / elicitation of λ.

Connection with composite likelihoods

- A GB-PPM with average dissimilarities can be interpreted as the Bayesian update of a pairwise difference likelihood (Varin et al., 2011).

Connection with composite likelihoods

- A GB-PPM with average dissimilarities can be interpreted as the Bayesian update of a pairwise difference likelihood (Varin et al., 2011).

- Suppose the observations follow some location family of distributions

$$(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \lambda, c_i = k) \overset{\text{iid}}{\sim} \pi\left(\boldsymbol{x} - \boldsymbol{\mu}_k \mid \lambda\right), \qquad i \in C_k, \quad k = 1, \dots, K,$$

  where $\boldsymbol{\mu}_k \in \mathbb{R}^d$.

Connection with composite likelihoods

- A GB-PPM with average dissimilarities can be interpreted as the Bayesian update of a pairwise difference likelihood (Varin et al., 2011).

- Suppose the observations follow some location family of distributions

$$(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \lambda, c_i = k) \stackrel{\text{iid}}{\sim} \pi\left(\boldsymbol{x} - \boldsymbol{\mu}_k \mid \lambda\right), \qquad i \in C_k, \quad k = 1, \dots, K,$$

  where $\boldsymbol{\mu}_k \in \mathbb{R}^d$.

- We model the within-cluster differences $\boldsymbol{x}_i - \boldsymbol{x}_{i'}$ with L^p spherical distributions, which are symmetric around 0. The location parameter $\boldsymbol{\mu}_k$ simplifies.

Connection with composite likelihoods

- A GB-PPM with average dissimilarities can be interpreted as the Bayesian update of a pairwise difference likelihood (Varin et al., 2011).

- Suppose the observations follow some location family of distributions

$$(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \lambda, c_i = k) \overset{\text{iid}}{\sim} \pi\left(\boldsymbol{x} - \boldsymbol{\mu}_k \mid \lambda\right), \qquad i \in C_k, \quad k = 1, \dots, K,$$

  where $\boldsymbol{\mu}_k \in \mathbb{R}^d$.

- We model the within-cluster differences $\boldsymbol{x}_i - \boldsymbol{x}_{i'}$ with L^p spherical distributions, which are symmetric around 0. The location parameter $\boldsymbol{\mu}_k$ simplifies.

- The associated pairwise difference likelihood is proportional to

$$\pi_{\text{DIFF}}(\boldsymbol{X} \mid \boldsymbol{c}, \lambda) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \left[\prod_{i' \in C_k} \pi_{\text{SP}}(\boldsymbol{x}_i - \boldsymbol{x}_{i'} \mid \lambda)\right]^{1/n_k},$$

  where the exponent $1/n_k$ is a correction that deflates the likelihood.

Two notable examples

Squared Euclidean GB-PPM

Bregman-divergence representation

$$\pi_\varphi(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \exp\left\{-\lambda \sum_{i \in C_k} ||\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k||_2^2\right\}, \qquad \boldsymbol{c} : |\boldsymbol{C}| = K.$$

Pairwise dissimilarity representation

$$\pi_\gamma(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} \exp\left\{-\frac{\lambda}{2} \sum_{i \in C_k} \frac{1}{n_k} \sum_{i' \in C_k} ||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_2^2\right\}, \qquad \boldsymbol{c} : |\boldsymbol{C}| = K.$$

- In both cases, this is consistent with

$$(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \lambda, c_i = k) \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{\mu}_k, (2\lambda)^{-1} I_d), \qquad i \in C_k, \quad k = 1, \dots, K.$$

Squared Euclidean GB-PPM, estimation of λ

- The parameter λ is proportional to the within-cluster precision.
- A possibility is to estimate λ from the data by considering the joint model

$$\pi(\boldsymbol{c}, \lambda \mid \boldsymbol{X}) \propto \pi(\lambda)\lambda^{nd/2} \prod_{k=1}^{K} \exp\left\{-\lambda \sum_{i \in C_k} ||\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k||_2^2\right\}, \qquad \boldsymbol{c} : |\boldsymbol{C}| = K.$$

- Note that the term $\lambda^{nd/2}$ follows from our probabilistic interpretation. Without our Theorems the estimation of λ would be much more problematic.
- This constitutes a reasonable and simple default strategy for the estimation of λ, which is otherwise a difficult problem.
- If we let $\lambda \sim \textsc{Gamma}(a_\lambda, b_\lambda)$ a priori, then the full conditional is conjugate.

Minkowski dissimilarities GB-PPM

Let $\gamma(||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_p^p) = ||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_p$ be the Minkowski distance. The associated GB-PPM is

$$\pi_\gamma(\boldsymbol{c} \mid \lambda, \boldsymbol{X}) \propto \prod_{k=1}^K \exp\left\{-\frac{\lambda}{2}\sum_{i\in C_k}\frac{1}{n_k}\sum_{i'\in C_k}||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_p\right\}, \qquad \boldsymbol{c} : |\boldsymbol{C}| = K.$$

- The L^p spherical distribution associated to the pairs $\boldsymbol{x}_i - \boldsymbol{x}_{i'}$ has density

$$\pi_{\mathrm{SP}}(\boldsymbol{x}_i - \boldsymbol{x}_{i'} \mid \lambda) = \frac{p^{d-1}}{2^d\Gamma(1/p)^d}\frac{\Gamma(d/p)}{\Gamma(d)}\left(\frac{\lambda}{2}\right)^d \exp\left\{-\frac{\lambda}{2}||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_p\right\}.$$

- λ is therefore a scale parameter and can be estimated paralleling the steps of the k-means case. The availability of the term λ^d is crucial.
- The Manhattan distance case ($p = 1$) has appealing robustness properties.

Illustrations

Synthetic dataset I

- In this experiment we consider $n = 200$ observations evenly divided in $K = 4$ clusters, each having $n_1 = \cdots = n_4 = 50$ data points.

- We simulate the data as follows

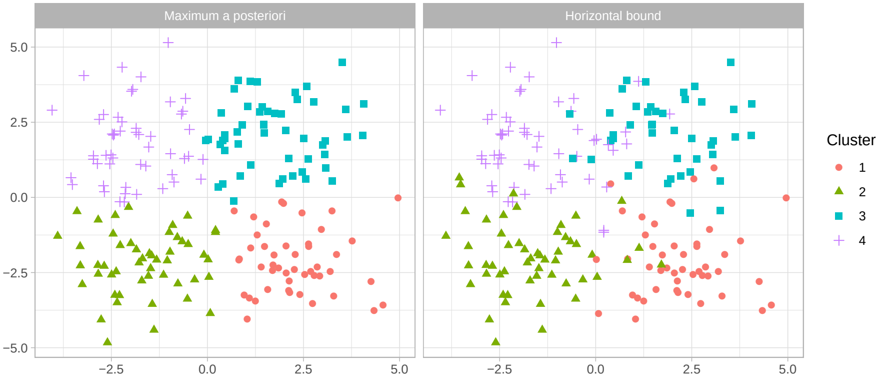
$$(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \sigma^2, c_i = k) \overset{\text{iid}}{\sim} \mathrm{N}\left(\boldsymbol{\mu}_k, \sigma^2 I_2\right), \qquad i \in C_k, \quad k = 1, \ldots, K,$$

  with $\boldsymbol{\mu}_1 = (-2, -2)$, $\boldsymbol{\mu}_2 = (-2, 2)$, $\boldsymbol{\mu}_3 = (2, -2)$, $\boldsymbol{\mu}_4 = (2, 2)$, and $\sigma^2 = 1.5$.

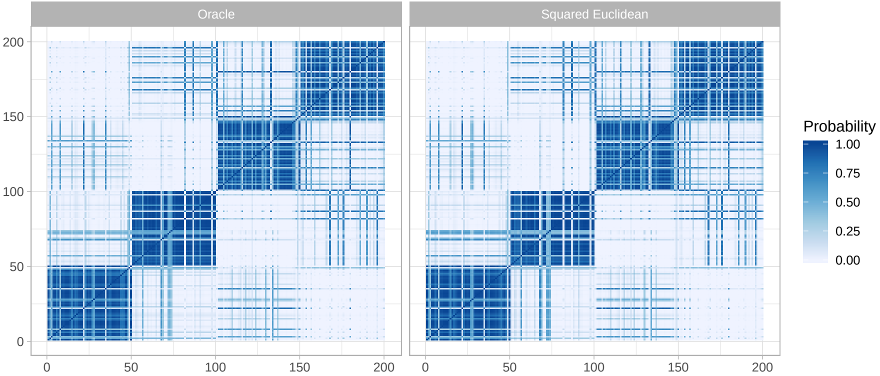
- We aim at comparing the uncertainty quantification of a GB-PPM with that of an oracle distribution, i.e. with

$$\pi_{\text{ORACLE}}(\boldsymbol{c} \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sigma^2, \boldsymbol{X}) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \mathrm{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \sigma^2 I_2)^{\mathbb{1}(c_i = k)}.$$

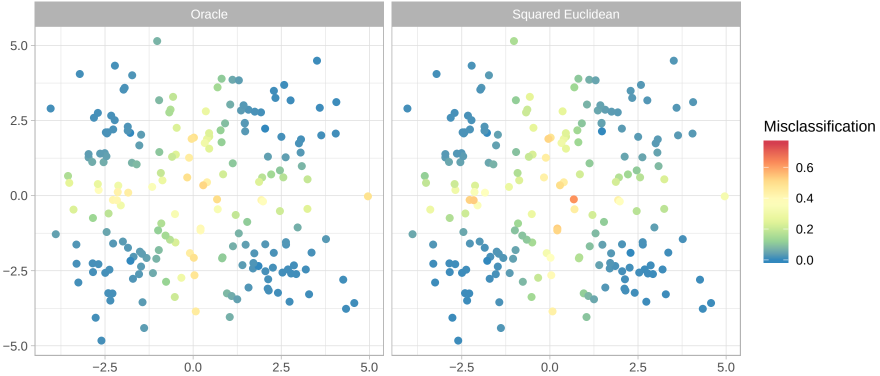
Synthetic dataset I (cont'd)



Synthetic dataset I (cont'd)



Synthetic dataset I (cont'd)



Synthetic dataset II

- We consider $n = 200$ observations evenly divided in $K = 4$ clusters, each having $n_1 = \cdots = n_4 = 50$ data points.

- We simulate the data from

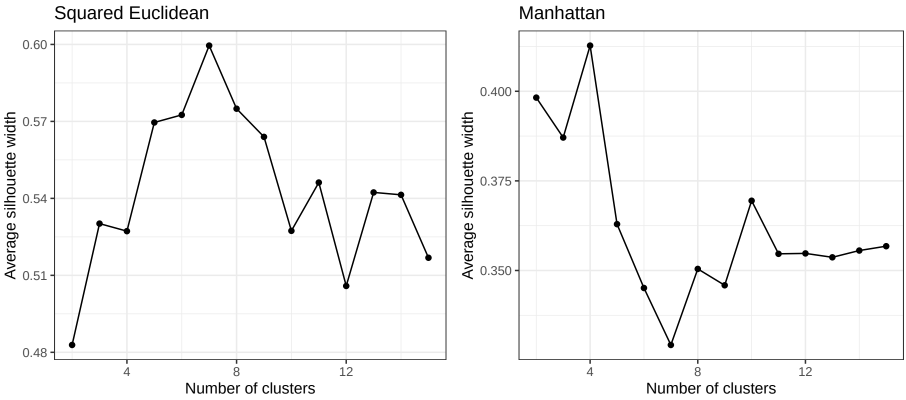
$$(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \sigma^2, c_i = k) \overset{\text{iid}}{\sim} t_2\left(\boldsymbol{\mu}_k, \sigma^2 I_2\right), \qquad i \in C_k, \quad k = 1, \dots, K,$$

  where $t_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Student's t-distribution with location $\boldsymbol{\mu}$, scale $\boldsymbol{\Sigma}$, and 2 degrees of freedom.

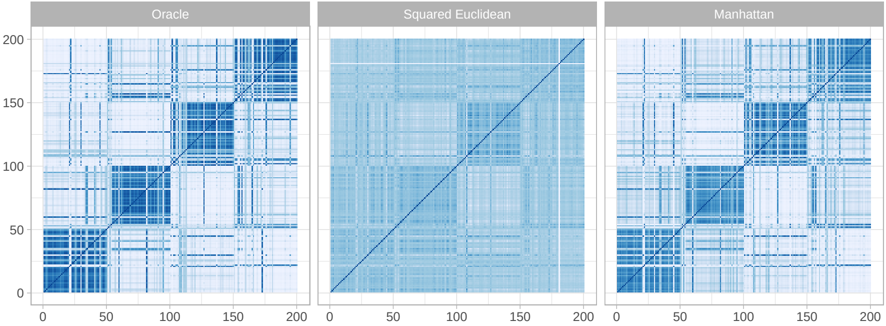
- Some "outliers" expected, because a t_2 distribution has infinite variance.

- We compare our estimates with the oracle distribution also in this case.

Synthetic dataset II (cont'd)



Synthetic dataset II (cont'd)



Thanks!

- We introduced a generalized Bayes modeling framework for clustering.
- We studied its general properties and presented two broad classes of tractable models.
- The manuscript is forthcoming on ArXiv!