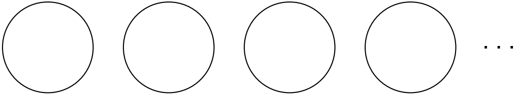
Analysis of the maximal a posteriori partition in the Gaussian Dirichlet Process Mixture Model

Łukasz Rajkowski

University of Warsaw

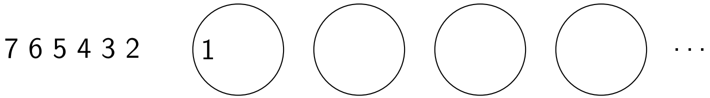
CRISM Summer School, Warwick University
July, 2018

Chinese Restaurant Process

7 6 5 4 3 2 1 

$\mathbb{P}(\text{new table}) \propto \alpha$ $\qquad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

Chinese Restaurant Process

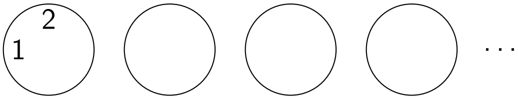


$\mathbb{P}$(new table) $\propto \alpha$ $\mathbb{P}$(join table) $\propto \#$ sitting there

$$\mathbb{P} = \frac{\alpha}{\alpha}$$

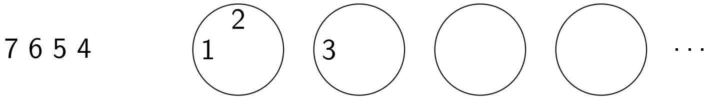
Chinese Restaurant Process

7 6 5 4 3



$\mathbb{P}$(new table) $\propto \alpha$ $\qquad$ $\mathbb{P}$(join table) $\propto$ # sitting there

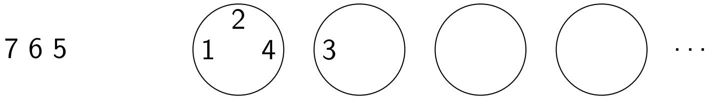
$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha}$$



$\mathbb{P}(\text{new table}) \propto \alpha$ $\qquad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha}$$

Chinese Restaurant Process

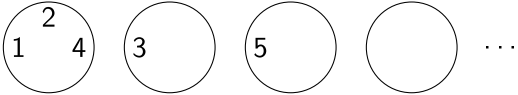


$\mathbb{P}(\text{new table}) \propto \alpha$ $\qquad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \cdot \frac{2}{3+\alpha}$$

Chinese Restaurant Process

7 6

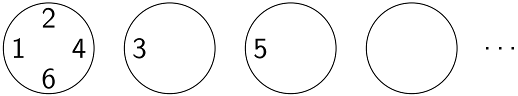


$\mathbb{P}(\text{new table}) \propto \alpha$ $\qquad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \cdot \frac{2}{3+\alpha} \cdot \frac{\alpha}{4+\alpha}$$

Chinese Restaurant Process

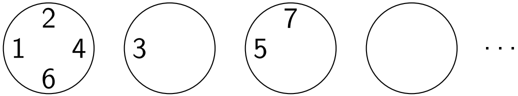
7



$\mathbb{P}(\text{new table}) \propto \alpha$ $\qquad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \cdot \frac{2}{3+\alpha} \cdot \frac{\alpha}{4+\alpha} \cdot \frac{3}{5+\alpha}$$

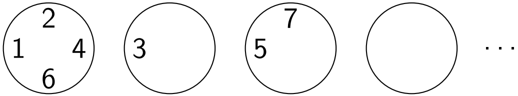
Chinese Restaurant Process



$\mathbb{P}(\text{new table}) \propto \alpha \qquad \mathbb{P}(\text{join table}) \propto \#\text{ sitting there}$

$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \cdot \frac{2}{3+\alpha} \cdot \frac{\alpha}{4+\alpha} \cdot \frac{3}{5+\alpha} \cdot \frac{1}{6+\alpha}$$

Chinese Restaurant Process

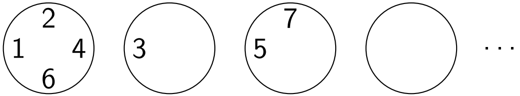


$\mathbb{P}(\text{new table}) \propto \alpha \qquad\qquad \mathbb{P}(\text{join table}) \propto \#\text{ sitting there}$

$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \cdot \frac{2}{3+\alpha} \cdot \frac{\alpha}{4+\alpha} \cdot \frac{3}{5+\alpha} \cdot \frac{1}{6+\alpha}$$

This is the probability of $\big\{\{1,2,4,6\},\{3\},\{5,7\}\big\}$

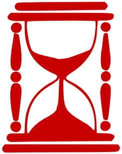
Chinese Restaurant Process



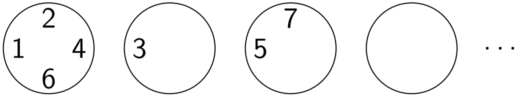
$\mathbb{P}(\text{new table}) \propto \alpha$ $\mathbb{P}(\text{join table}) \propto \#\ \text{sitting there}$

$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \cdot \frac{2}{3+\alpha} \cdot \frac{\alpha}{4+\alpha} \cdot \frac{3}{5+\alpha} \cdot \frac{1}{6+\alpha}$$

This is the probability of $\{\{1,2,4,6\},\{3\},\{5,7\}\}$



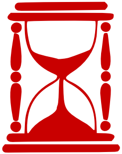
Chinese Restaurant Process



$\mathbb{P}(\text{new table}) \propto \alpha$ $\qquad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

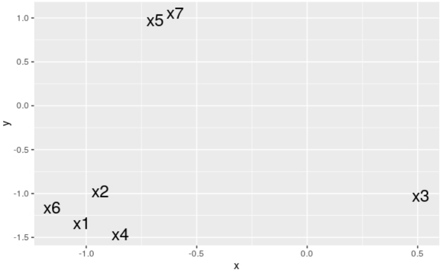
$$\mathbb{P} = \frac{\alpha}{\alpha} \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \cdot \frac{2}{3+\alpha} \cdot \frac{\alpha}{4+\alpha} \cdot \frac{3}{5+\alpha} \cdot \frac{1}{6+\alpha}$$

This is the probability of $\{\{1,2,4,6\},\{3\},\{5,7\}\}$

CRP = a nice way to sample partitions

Gaussian Dirichet Process Mixture Model



unknown number of clusters in $\mathbb{R}^d$
data spread 'normally' within each cluster

Gaussian Dirichet Process Mixture Model

This may be modelled as follows (blue=hyperparameters)

$$\mathcal{J} \sim \mathrm{CRP}(\alpha)_n$$

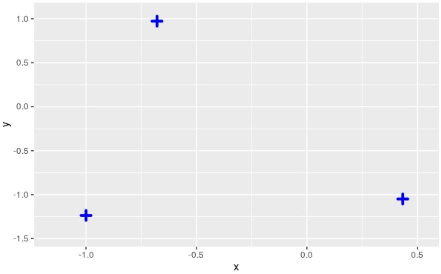
$\mathcal{J} = \big\{\{1, 2, 4, 6\}, \{3\}, \{5, 7\}\big\}$

Gaussian Dirichet Process Mixture Model

This may be modelled as follows (blue=hyperparameters)

$$
\begin{aligned}
\mathcal{J} &\sim \mathrm{CRP}(\alpha)_n \\
\boldsymbol{\theta} = (\theta_J)_{J\in\mathcal{J}} \,|\, \mathcal{J} &\overset{\text{iid}}{\sim} \mathcal{N}(\vec{\mu}, T)
\end{aligned}
$$

$\mathcal{J} = \big\{\{1,2,4,6\},\{3\},\{5,7\}\big\}$

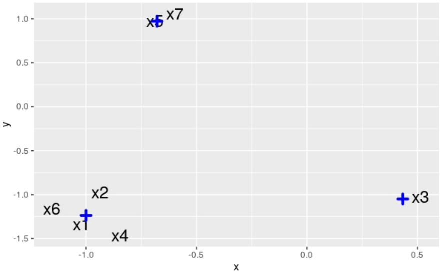


Gaussian Dirichet Process Mixture Model

This may be modelled as follows (blue=hyperparameters)

$$\begin{aligned} \mathcal{J} &\sim \mathrm{CRP}(\alpha)_n \\ \boldsymbol{\theta} = (\theta_J)_{J\in\mathcal{J}} \,|\, \mathcal{J} &\overset{\text{iid}}{\sim} \mathcal{N}(\vec{\mu}, T) \\ \boldsymbol{x}_J = (x_j)_{j\in J} \,|\, \mathcal{J}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \mathcal{N}(\theta_J, \Sigma) \quad \text{for } J \in \mathcal{J} \end{aligned}$$

$\mathcal{J} = \big\{\{1,2,4,6\},\{3\},\{5,7\}\big\}$

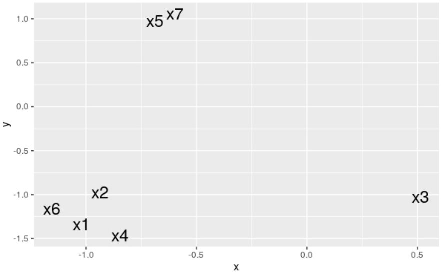


Gaussian Dirichet Process Mixture Model

This may be modelled as follows (blue=hyperparameters)

$$\begin{aligned}
\mathcal{J} &\sim \mathsf{CRP}(\alpha)_n \\
\boldsymbol{\theta} = (\theta_J)_{J\in\mathcal{J}} \,|\, \mathcal{J} &\overset{\text{iid}}{\sim} \mathcal{N}(\vec{\mu}, T) \\
\boldsymbol{x}_J = (x_j)_{j\in J} \,|\, \mathcal{J}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \mathcal{N}(\theta_J, \Sigma) \quad \text{for } J \in \mathcal{J}
\end{aligned}$$

$\cancel{\mathcal{J} = \{\{1,2,4,6\},\{3\},\{5,7\}\}}$ **The 'true' partition is not known**

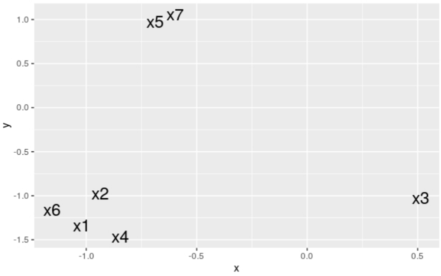


Gaussian Dirichet Process Mixture Model

This may be modelled as follows (blue=hyperparameters)

$$\begin{aligned} \mathcal{J} &\sim \mathsf{CRP}(\alpha)_n \\ \boldsymbol{\theta} = (\theta_J)_{J\in\mathcal{J}} \,|\, \mathcal{J} &\overset{\text{iid}}{\sim} \mathcal{N}(\vec{\mu}, T) \\ \boldsymbol{x}_J = (x_j)_{j\in J} \,|\, \mathcal{J}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \mathcal{N}(\theta_J, \Sigma) \quad \text{for } J \in \mathcal{J} \end{aligned}$$

$\cancel{\mathcal{J} = \{\{1,2,4,6\},\{3\},\{5,7\}\}}$ **The 'true' partition is not known**



<u>Task:</u>
Estimate the distribution of $\mathcal{J}$ provided observation $\boldsymbol{x}$

The Maximal A Posteriori Partition

- The Bayesian approach is to compute the posterior $\mathcal{J} \mid \mathbf{x}$

The Maximal A Posteriori Partition

- The Bayesian approach is to compute the posterior $\mathcal{J} \,|\, \mathbf{x}$
- Easy to compute unnormalised probability $Q_{\mathbf{x}}(\mathcal{J})$

The Maximal A Posteriori Partition

- The Bayesian approach is to compute the posterior $\mathcal{J} \mid \mathbf{x}$
- Easy to compute unnormalised probability $Q_{\mathbf{x}}(\mathcal{J})$

The MAP

The Maximal A Posteriori (MAP) is the partition defined by

$$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x}) = \operatorname{argmax}_{\mathcal{J}} \mathbb{P}(\mathcal{J} \mid \boldsymbol{x}) = \operatorname{argmax}_{\mathcal{J}} Q_{\boldsymbol{x}}(\mathcal{J})$$

How well it performs?

- assume that the data comes from an iid sample from given distribution P on $\mathbb{R}^d$, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$. How would my Bayesian machinery behave as n grows infinitely?

How well it performs?

- assume that the data comes from an iid sample from given distribution P on $\mathbb{R}^d$, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$. How would my Bayesian machinery behave as n grows infinitely?
- Jeffrey Miller and Matthew Harrison. "Inconsistency of Pitman-Yor process mixtures for the number of components." JMLR (2014).
  Corollary: *In a very general family of conjugate models with CRP as a prior on partitions then if P is a mixture of t distributions from the model, then*

$$\limsup_{n\to\infty} \mathbb{P}(T_n = t \mid X_{1:n}) < 1,$$

  *so the posterior is **not consistent** for the number of clusters.*

How well it performs?

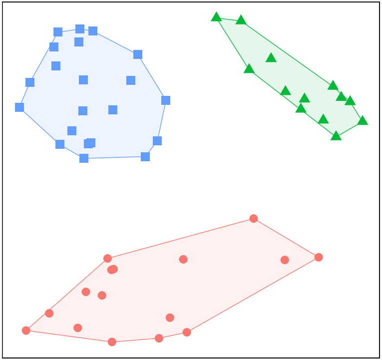
- assume that the data comes from an iid sample from given distribution P on $\mathbb{R}^d$, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$. How would my Bayesian machinery behave as n grows infinitely?
- Jeffrey Miller and Matthew Harrison. "Inconsistency of Pitman-Yor process mixtures for the number of components." JMLR (2014).
  Corollary: *In a very general family of conjugate models with CRP as a prior on partitions then if P is a mixture of t distributions from the model, then*
  $$\limsup_{n\to\infty} \mathbb{P}(T_n = t \mid X_{1:n}) < 1,$$
  *so the posterior is **not consistent** for the number of clusters.*
- **Goal:** Perform similar analysis for the MAP in Gaussian model.

Main results

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

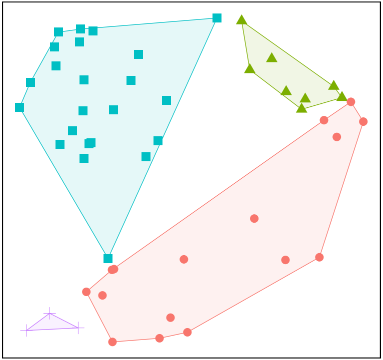


Convex and lovely

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

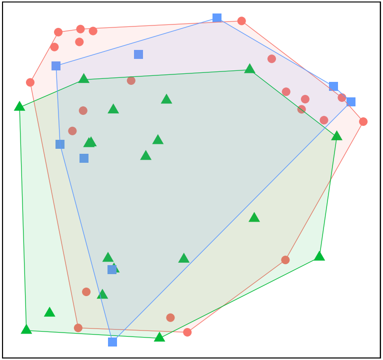


Convex but not lovely

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.



Not convex and disastrous

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

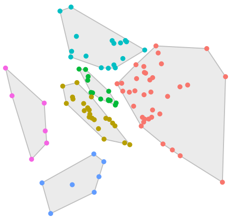
Infinite sequence of observations, the MAP on prefixes (a movie).

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



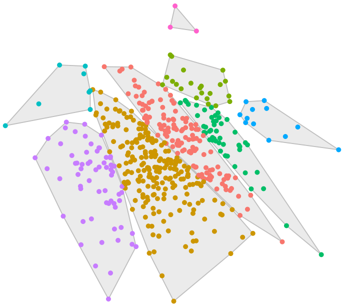
$n = 100$

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



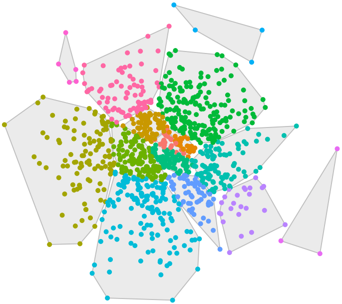
$n = 500$

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



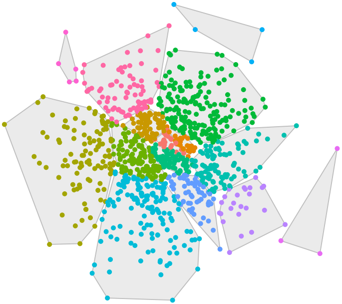
$n = 1000$

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



$n = 1000$

Question:

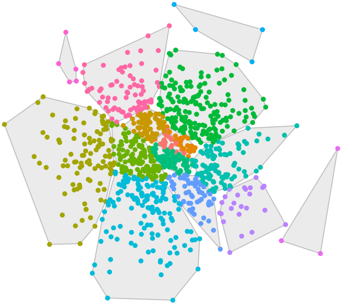
Can we control the (relative) size of the smallest cluster?

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



$n = 1000$

Question:

Can we control the (relative) size of the smallest cluster?

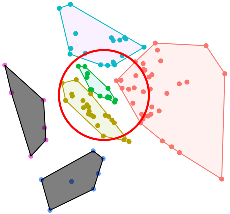
Partly...

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



$n = 100$

Question:

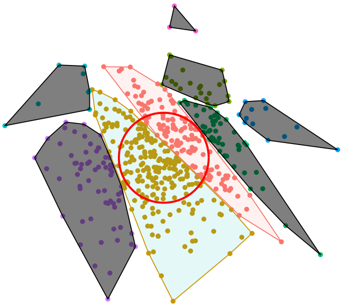
Can we control the (relative) size of the smallest cluster?

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



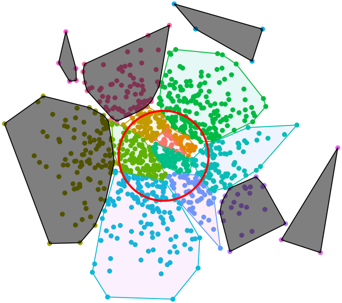
$n = 500$

Question:
Can we control the (relative) size of the smallest cluster?

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Infinite sequence of observations, the MAP on prefixes (a movie).



$n = 1000$

Question:

Can we control the (relative) size of the smallest cluster?

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Result 2 (size of clusters)

If $\sup_n \frac{1}{n} \sum_{i=1}^{n} \|x_n\|^2 < \infty$ then for every $r > 0$

$$\liminf_{n\to\infty} \min\{|J| \colon J \in \hat{\mathcal{J}}_{MAP}(\boldsymbol{x}_{1:n}), \exists_{j\in J} \|x_j\| < r\}/n := \gamma > 0.$$

Main results

Result 1 (convexity)

$\hat{\mathcal{J}}_{MAP}(\boldsymbol{x})$ is a convex partition with respect to $\boldsymbol{x}$.

Result 2 (size of clusters)

If $\sup_n \frac{1}{n}\sum_{i=1}^{n} \|x_n\|^2 < \infty$ then for every $r > 0$

$$\liminf_{n\to\infty} \min\{|J| \colon J \in \hat{\mathcal{J}}_{MAP}(\boldsymbol{x}_{1:n}), \exists_{j\in J}\|x_j\| < r\}/n := \gamma > 0.$$

Result 3 (behaviour in the limit)

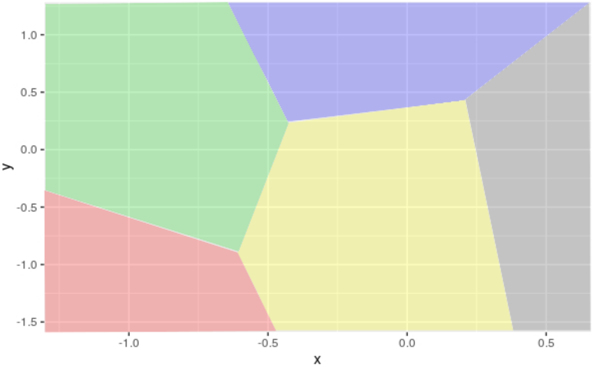
If $X_1, X_2, \ldots \sim P$ then $\hat{\mathcal{J}}_{MAP}(\boldsymbol{X}_{1:n})$ 'concentrates' around 'partitions' of R^d that maximise some given functional Δ **(P bounded and continous).**

Induced partitions

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).

Induced partitions

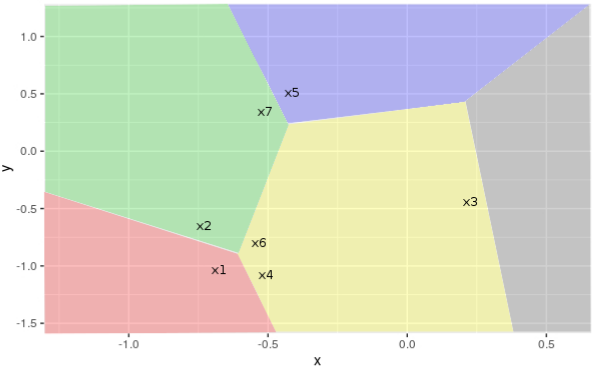
Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).
$\mathcal{A}$ is a **fixed** partition of $\mathbb{R}^d$;



$\mathcal{A}$

Induced partitions

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).
$\mathcal{A}$ is a **fixed** partition of $\mathbb{R}^d$; $\quad \mathcal{J}_n^{\mathcal{A}} = \{\{i \leqslant n \colon X_i \in A\} \colon A \in \mathcal{A}\}$.

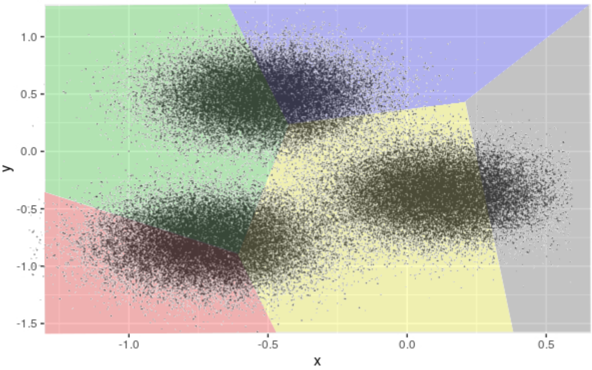


$$\mathcal{J}_7^{\mathcal{A}} = \{\{1\}, \{2, 7\}, \{3, 4, 6\}, \{5\}\}$$

you may compute $Q_{X_{1:7}}(\mathcal{J}_7^{\mathcal{A}})$

Induced partitions

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).
$\mathcal{A}$ is a **fixed** partition of $\mathbb{R}^d$; $\quad \mathcal{J}_n^{\mathcal{A}} = \{\{i \leqslant n \colon X_i \in A\} \colon A \in \mathcal{A}\}$.



$$\mathcal{J}_{10000}^{\mathcal{A}} = \{\{\ldots\}, \{\ldots\}, \{\ldots\}, \{\ldots\}, \{\ldots\}\}$$
$$Q_{X_{1:10000}}(\mathcal{J}_{10000}^{\mathcal{A}}) \approx ???$$

Induced partitions

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).
$\mathcal{A}$ is a **fixed** partition of $\mathbb{R}^d$; $\quad \mathcal{J}_n^{\mathcal{A}} = \{\{i \leqslant n \colon X_i \in A\} \colon A \in \mathcal{A}\}$.

Proposition

$\sqrt[n]{Q_{\boldsymbol{x}_{1:n}}(\mathcal{J}_n^{\mathcal{A}})} \overset{\text{a.s.}}{\approx} \frac{n}{e} \exp\{\Delta(\mathcal{A})\}$, where

$$\Delta(\mathcal{A}) = \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\|\mathbb{E}\left(\Sigma^{-2} X \mid A\right)\right\|^2 + \sum_{A \in \mathcal{A}} P(A) \ln P(A)$$

Induced partitions

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).
$\mathcal{A}$ is a **fixed** partition of $\mathbb{R}^d$; $\quad \mathcal{J}_n^{\mathcal{A}} = \{\{i \leqslant n \colon X_i \in A\} \colon A \in \mathcal{A}\}$.

Proposition

$\sqrt[n]{Q_{\boldsymbol{x}_{1:n}}(\mathcal{J}_n^{\mathcal{A}})} \overset{\text{a.s.}}{\approx} \frac{n}{e} \exp\{\Delta(\mathcal{A})\}$, where

$$\Delta(\mathcal{A}) = \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\| \mathbb{E}\left(\Sigma^{-2} X \mid A\right) \right\|^2 + \sum_{A \in \mathcal{A}} P(A) \ln P(A)$$

- nice interpretation of Δ (variance of CEV vs entropy)

Induced partitions

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).
$\mathcal{A}$ is a **fixed** partition of $\mathbb{R}^d$; $\mathcal{J}_n^{\mathcal{A}} = \{\{i \leqslant n \colon X_i \in A\} \colon A \in \mathcal{A}\}$.

Proposition

$\sqrt[n]{Q_{\boldsymbol{x}_{1:n}}(\mathcal{J}_n^{\mathcal{A}})} \overset{\text{a.s.}}{\approx} \frac{n}{e} \exp\{\Delta(\mathcal{A})\}$, where

$$\Delta(\mathcal{A}) = \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\| \mathbb{E}\left(\Sigma^{-2} X \mid A\right) \right\|^2 + \sum_{A \in \mathcal{A}} P(A) \ln P(A)$$

- nice interpretation of Δ (variance of CEV vs entropy)
- for P bounded you can do something similar for the MAP and hence prove Result 3

Induced partitions

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$ (e.g. a mixture of three gaussians).
$\mathcal{A}$ is a **fixed** partition of $\mathbb{R}^d$; $\mathcal{J}_n^{\mathcal{A}} = \{\{i \leqslant n \colon X_i \in A\} \colon A \in \mathcal{A}\}$.

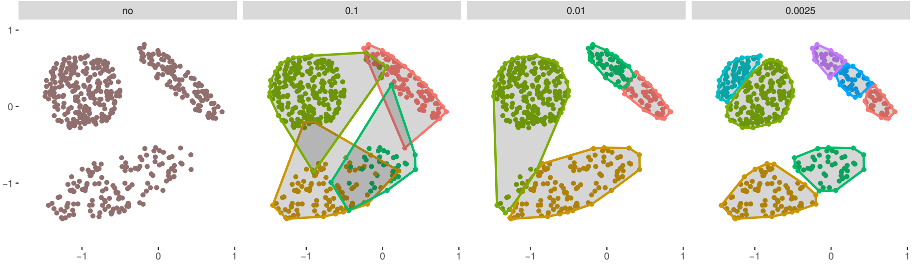
Proposition

$\sqrt[n]{Q_{\boldsymbol{x}_{1:n}}(\mathcal{J}_n^{\mathcal{A}})} \overset{\text{a.s.}}{\approx} \frac{n}{e} \exp\{\Delta(\mathcal{A})\}$, where

$$\Delta(\mathcal{A}) = \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\| \mathbb{E}\left(\Sigma^{-2} X \mid A\right) \right\|^2 + \sum_{A \in \mathcal{A}} P(A) \ln P(A)$$

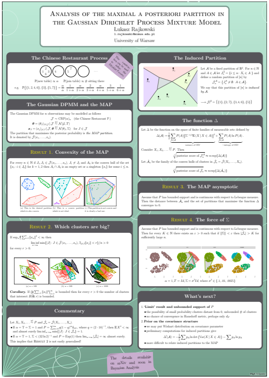
- nice interpretation of Δ (variance of CEV vs entropy)
- for P bounded you can do something similar for the MAP and hence prove Result 3
- depends only on within-group covariance Σ^2 – 'inconsistency'!

Illustration of the last point



Interested in details?

- *Analysis of the maximal posterior partition in the Dirichlet Process Gaussian Mixture Model* available on `arXiv.org` and accepted to Bayesian Analysis
- Poster:



Thank you for your attention