On the Maximum a Posteriori partition in nonparametric Bayesian mixture models

Łukasz Rajkowski
University of Warsaw

Statistical Learning Seminar
Zoom, 7 May 2021

Agenda

- Introduction: definitions and notation
- Results in L.R., Bayesian Analysis 2019
- Generalisations
- Potential applications

Agenda

- Introduction: definitions and notation
- Results in L.R., Bayesian Analysis 2019
- Generalisations
- Potential applications

Estimated time $\sim$ 40 min

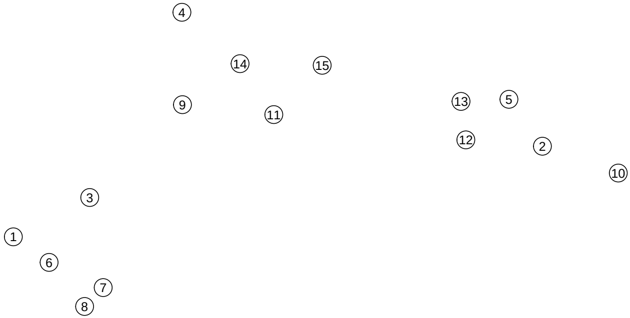
Agenda

- Introduction: definitions and notation
- Results in L.R., Bayesian Analysis 2019
- Generalisations
- Potential applications

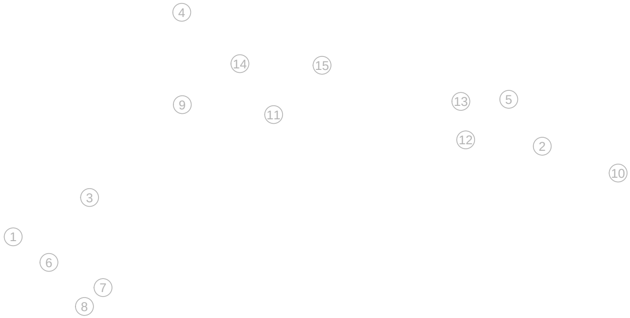
Estimated time $\sim$ 40 min

Interruptions very welcome!

Bayesian Approach to Clustering

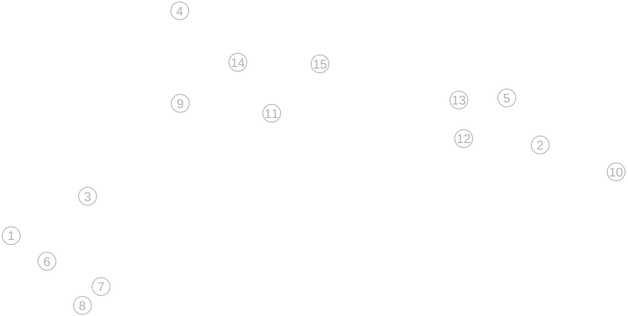


Bayesian Approach to Clustering



Bayesian model with clustering as 'parameter'

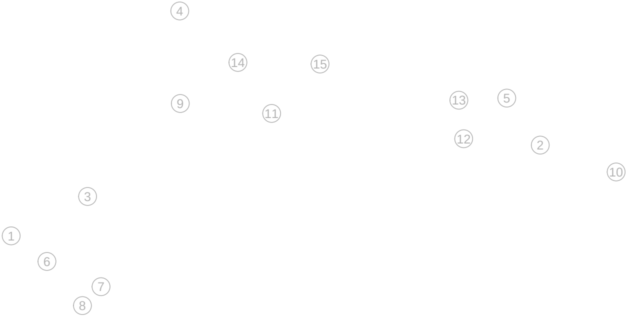
Bayesian Approach to Clustering



Bayesian model with clustering as 'parameter'

$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

Bayesian Approach to Clustering

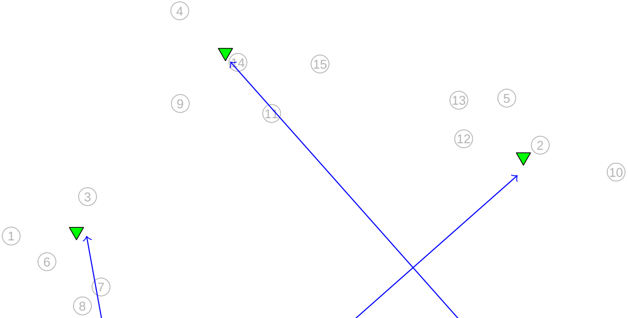


Bayesian model with clustering as 'parameter'

$$\mathbf{\Pi} = \{1,3,6,7,8\},\ \{2,5,10,12,13\},\ \{4,9,11,14,15\}$$

„observations normally distributed within clusters"

Bayesian Approach to Clustering

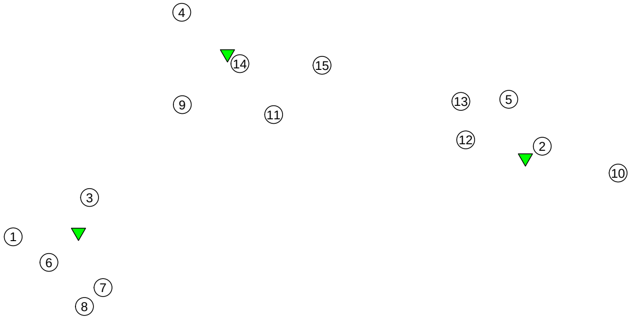


Bayesian model with clustering as ‘parameter’

$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\theta_1, \ldots, \theta_K \overset{\text{iid}}{\sim} \text{Normal}(\cdot)$$

Bayesian Approach to Clustering

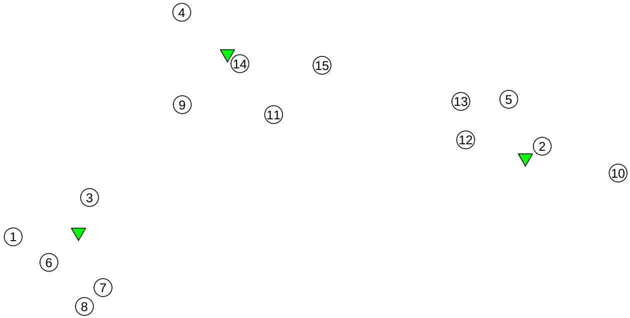


Bayesian model with clustering as 'parameter'

$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\begin{aligned} \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot) \\ (x_i)_{i \in C_k} \,|\, \theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k, \cdot) \end{aligned}$$

Bayesian Approach to Clustering



Bayesian model with clustering as 'parameter'

$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\begin{aligned} \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot) \\ (x_i)_{i \in C_k} \mid \theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k, \cdot) \end{aligned}$$

Inference based on $\mathbf{\Pi} \mid (x_i)_{i \leqslant n}$

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

Prior distribution on partitions?

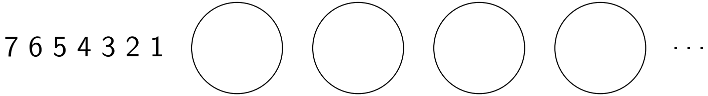
Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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

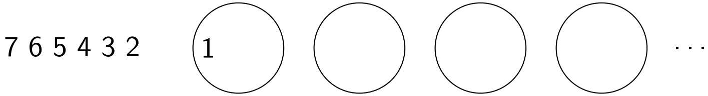


$\mathbb{P}$(new table) $\propto \alpha$ $\qquad$ $\mathbb{P}$(join table) $\propto$ # sitting there

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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



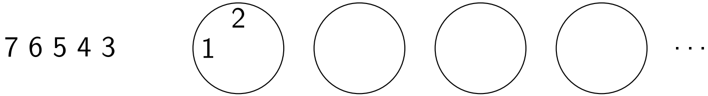
$\mathbb{P}$(new table) $\propto \alpha$ $\mathbb{P}$(join table) $\propto$ # sitting there

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

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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



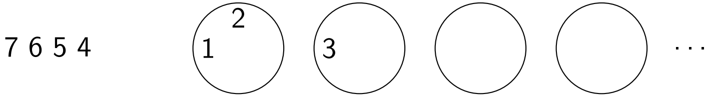
$\mathbb{P}$(new table) $\propto \alpha$ $\qquad$ $\mathbb{P}$(join table) $\propto$ # sitting there

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

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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



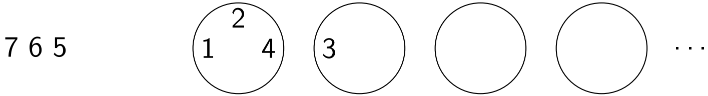
$\mathbb{P}$(new table) $\propto \alpha$ $\qquad$ $\mathbb{P}$(join table) $\propto$ # sitting there

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

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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



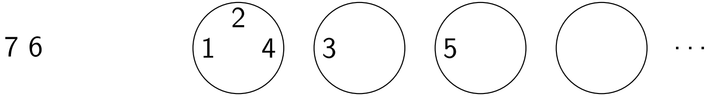
$\mathbb{P}(\text{new table}) \propto \alpha$ $\quad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

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

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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



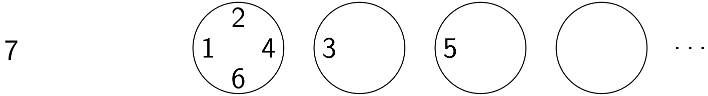
$\mathbb{P}(\text{new table}) \propto \alpha$ $\qquad$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

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

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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



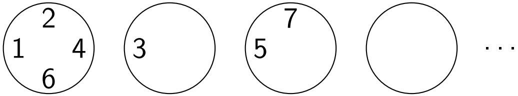
$\mathbb{P}(\text{new table}) \propto \alpha$ $\mathbb{P}(\text{join table}) \propto \#$ sitting there

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

Prior distribution on partitions?

Chinese Restaurant Process with parameter α can be viewed as a probability distribution on the space of partitions of a finite set.

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



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

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

Bayesian inference and the MAP

Bayesian inference on the clustering of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

the posterior distribution of $\mathbf{\Pi}$ given $\boldsymbol{x}$, i.e. $\mathbf{\Pi} \mid \boldsymbol{x}$

Bayesian inference and the MAP

Bayesian inference on the clustering of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

the posterior distribution of $\mathbf{\Pi}$ given $\boldsymbol{x}$, i.e. $\mathbf{\Pi} \mid \boldsymbol{x}$

- easy to compute up to the norming constant

$$\mathbb{P}(\mathbf{\Pi} = \mathcal{I} \mid \boldsymbol{x}) \propto \text{Prior} \times \text{Likelihood}$$

Bayesian inference and the MAP

Bayesian inference on the clustering of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

the posterior distribution of $\boldsymbol{\Pi}$ given $\boldsymbol{x}$, i.e. $\boldsymbol{\Pi} \,|\, \boldsymbol{x}$

- easy to compute up to the norming constant

$$\mathbb{P}(\boldsymbol{\Pi} = \mathcal{I} \,|\, \boldsymbol{x}) \propto \text{Prior} \times \text{Likelihood}$$

- not possible to compute the norming constant

Bayesian inference and the MAP

Bayesian inference on the clustering of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

the posterior distribution of $\mathbf{\Pi}$ given $\boldsymbol{x}$, i.e. $\mathbf{\Pi} \,|\, \boldsymbol{x}$

- easy to compute up to the norming constant

$$\mathbb{P}(\mathbf{\Pi} = \mathcal{I} \,|\, \boldsymbol{x}) \propto \text{Prior} \times \text{Likelihood}$$

- not possible to compute the norming constant

Definition (the Maximum A Posteriori partition)

The MAP partition of $\boldsymbol{x}$:
the partition $\hat{I}_{MAP}(\boldsymbol{x})$ that maximises $\mathbb{P}(\mathbf{\Pi} = \mathcal{I} \,|\, \boldsymbol{x})$

Normal-Normal CRP model

In R.(2019) the MAP in the following model was analysed:

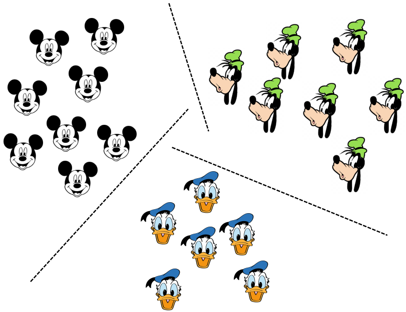
$$\begin{aligned} \mathcal{J} &\sim \mathsf{CRP}(\alpha)_n \\ \boldsymbol{\theta} = (\theta_J)_{J\in\mathcal{J}} \,|\, \mathcal{J} &\overset{\text{iid}}{\sim} \mathcal{N}(\vec{\mu}, T) \\ \boldsymbol{x}_J = (x_j)_{j\in J} \,|\, \mathcal{J}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \mathcal{N}(\theta_J,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

Normal-Normal CRP model

In R.(2019) the MAP in the following model was analysed:

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

The first result was that the clusters in the MAP partition are linearly separated.



Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

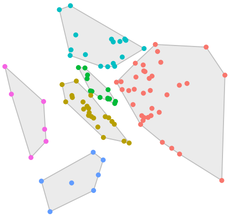
How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



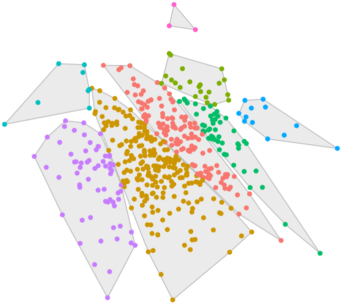
$n = 100$

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



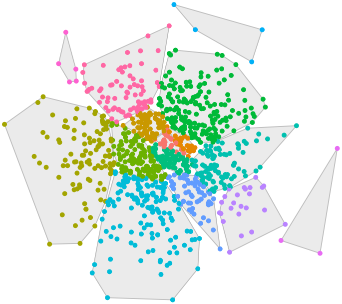
$n = 500$

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



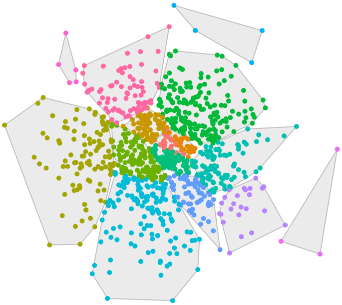
$n = 1000$

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



$n = 1000$

Question:

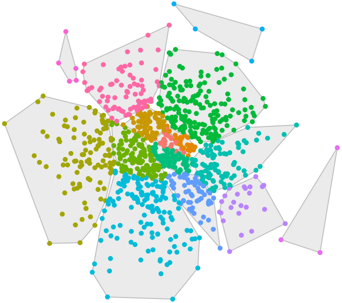
Can we control the (relative) size of the smallest cluster?

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



$n = 1000$

Question:

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

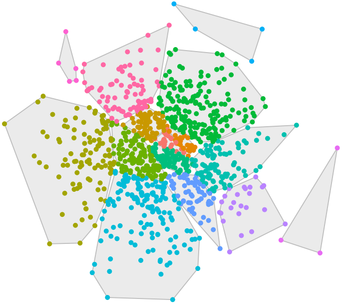
Partly...

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



$n = 1000$

Question:

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

Partly. . .

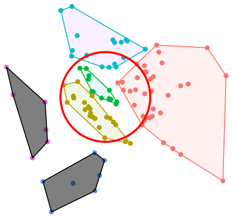
Of those clusters that intersect given ball

Normal-Normal CRP model cnt.

‘Frequentists validation of the MAP’

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



Question:

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

Partly...

Of those clusters that intersect given ball

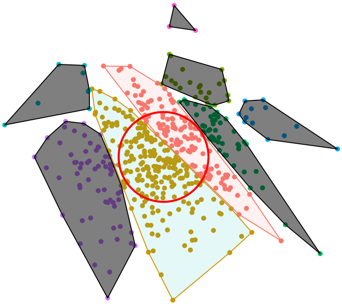
$n = 100$

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



$n = 500$

Question:

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

Partly...

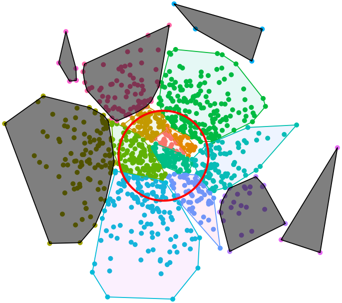
Of those clusters that intersect given ball

Normal-Normal CRP model cnt.

‘Frequentists validation of the MAP’

Let $X_1, X_2, \ldots$ be an IID sample from P.

How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?



$n = 1000$

Question:

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

Partly...

Of those clusters that intersect given ball

Normal-Normal CRP model cnt.

'Frequentists validation of the MAP'
Let $X_1, X_2, \dots$ be an IID sample from P.
How does $\hat{\mathcal{I}}_{MAP}(X_{1:n})$ behave as $n \to \infty$?

<u>Question:</u>
Can we control the (relative) size of the smallest cluster?
Partly...

Of those clusters that intersect given ball

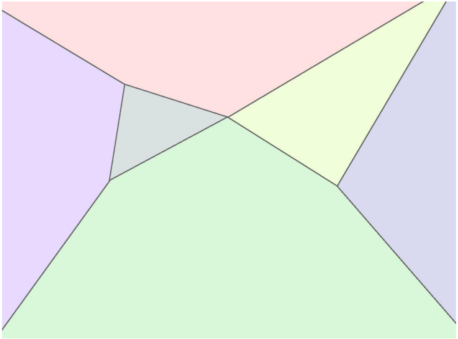
Result (size of clusters)

If $X_1, X_2, \dots \sim P$, $\mathbb{E}\,\|X\|^4 < \infty$, then a.s. for every $r > 0$

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

Induced partitions

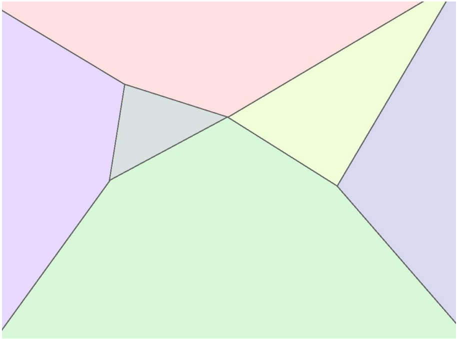
Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:



Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

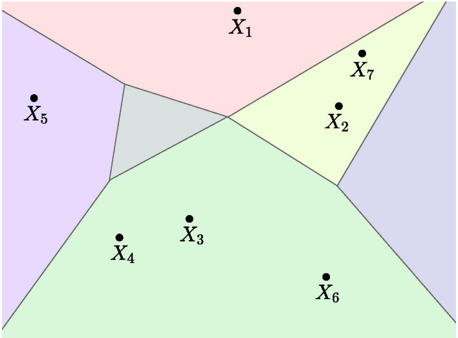
Let $X_1, X_2, \ldots, X_7 \overset{\text{iid}}{\sim} P$



Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

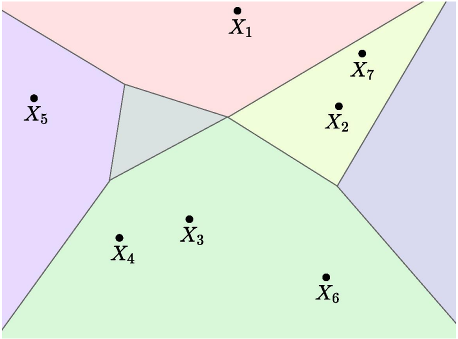
Let $X_1, X_2, \ldots, X_7 \overset{\text{iid}}{\sim} P$



Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

Let $X_1, X_2, \ldots, X_7 \overset{\text{iid}}{\sim} P$

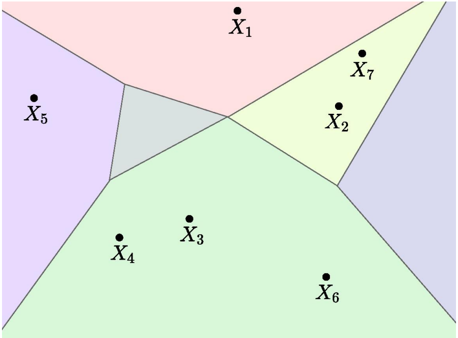


$$\mathcal{I}_7^{\mathcal{A}}(\boldsymbol{X}_{1:7}) = \{\{1\}, \{2, 7\}, \{3, 4, 6\}, \{5\}\}$$

Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

Let $X_1, X_2, \ldots, X_7 \overset{\text{iid}}{\sim} P$



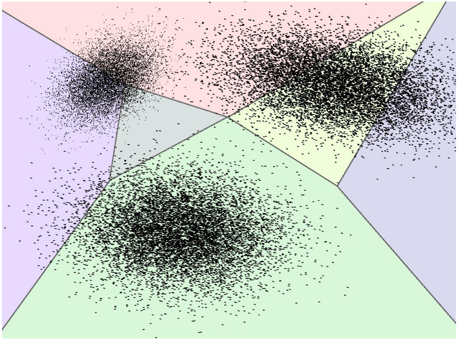
$$\mathcal{I}_7^{\mathcal{A}}(\boldsymbol{X}_{1:7}) = \{\{1\}, \{2, 7\}, \{3, 4, 6\}, \{5\}\}$$

you may compute $\mathbb{P}(\mathcal{I}_7^{\mathcal{A}}(\boldsymbol{X}_{1:7}) \mid \boldsymbol{X}_{1:7})$

Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

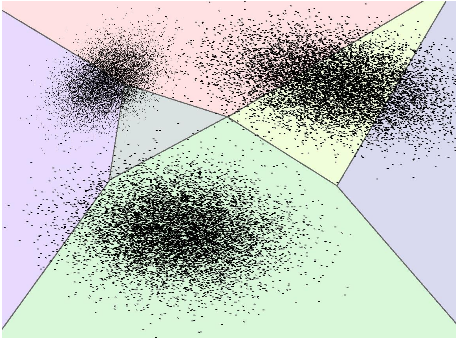
Let $X_1, X_2, \ldots, X_{10000} \stackrel{\text{iid}}{\sim} P$



Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

Let $X_1, X_2, \ldots, X_{10000} \overset{\text{iid}}{\sim} P$



$$\mathcal{I}^{\mathcal{A}}_{10000}(\boldsymbol{X}_{1:10000}) = \{\{\ldots\}, \{\ldots\}, \{\ldots\}, \{\ldots\}, \{\ldots\}\}$$
$$\mathbb{P}(\mathcal{I}^{\mathcal{A}}_{10000}(\boldsymbol{X}_{1:10000}) \mid \boldsymbol{X}_{1:10000}) \approx ???$$

Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

Let $X_1, X_2, \ldots, X_{10000} \overset{\text{iid}}{\sim} P$

Proposition

$\sqrt[n]{\mathbb{P}(\mathcal{I}_n^{\mathcal{A}}(\boldsymbol{X}_{1:n}) \mid \boldsymbol{X}_{1:n})} \overset{\text{a.s.}}{\asymp} \exp\{\Delta_P(\mathcal{A})\}$ where

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) + \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\|\mathbb{E}\left(\Sigma_0^{-1} X \mid X \in A\right)\right\|^2$$

Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

Let $X_1, X_2, \ldots, X_{10000} \overset{\text{iid}}{\sim} P$

Proposition

$\sqrt[n]{\mathbb{P}(\mathcal{I}_n^{\mathcal{A}}(\boldsymbol{X}_{1:n}) \mid \boldsymbol{X}_{1:n})} \overset{\text{a.s.}}{\asymp} \exp\{\Delta_P(\mathcal{A})\}$ where

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) + \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\| \mathbb{E}\left(\Sigma_0^{-1} X \mid X \in A\right) \right\|^2$$

$\log \sqrt[n]{\text{CRP prior}}$ $\qquad$ $\log \sqrt[n]{\text{Gaussian Likelihood}}$

Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

Let $X_1, X_2, \ldots, X_{10000} \overset{\text{iid}}{\sim} P$

Proposition

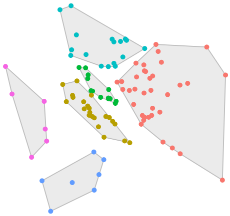
$\sqrt[n]{\mathbb{P}(\mathcal{I}_n^{\mathcal{A}}(\boldsymbol{X}_{1:n}) \mid \boldsymbol{X}_{1:n})} \overset{\text{a.s.}}{\asymp} \exp\{\Delta_P(\mathcal{A})\}$ where

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) + \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \|\mathbb{E}\left(\Sigma_0^{-1} X \mid X \in A\right)\|^2$$

$\log \sqrt[n]{\text{CRP prior}}$ $\log \sqrt[n]{\text{Gaussian Likelihood}}$

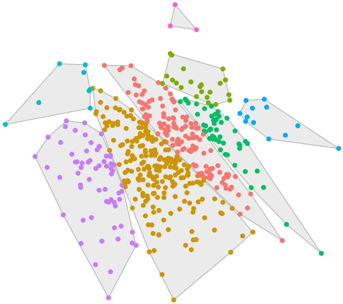
straightforward computations using SLLN

MAP limits



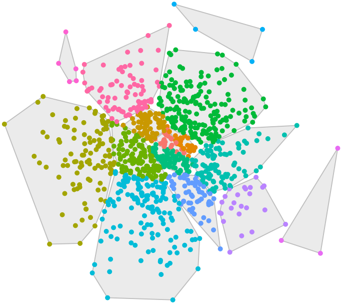
$n = 100$

MAP limits



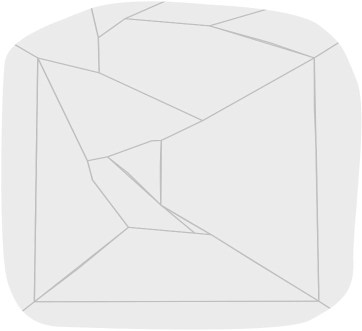
$n = 500$

MAP limits



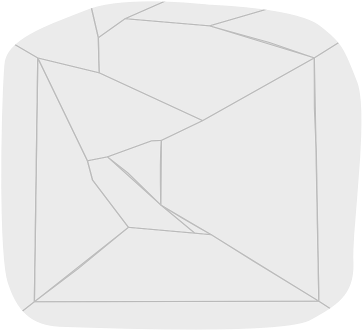
$n = 1000$

MAP limits



$n = \infty$???

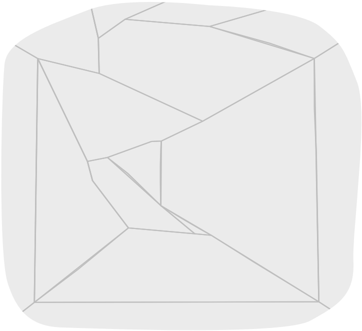
MAP limits



$n = \infty$???

If there is such limit, is it a maximiser o Δ_P?

MAP limits



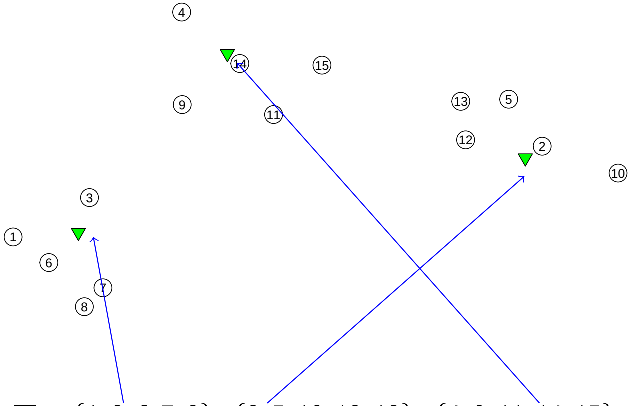
$n = \infty$???

If there is such limit, is it a maximiser o Δ_P?

Theorem (R. 2019)

Every limit point of the sequence of convex hulls of the MAP partitions is a maximiser of Δ_P. **(in Gaussian CRP model**$+P$ **bounded&continuous)**

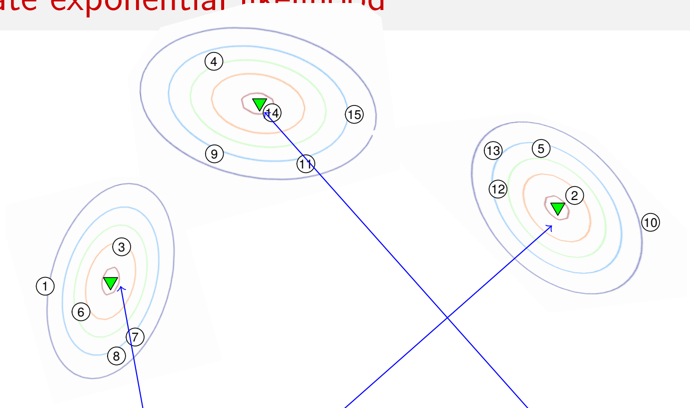
Conjugate exponential likelihood



$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\begin{aligned} \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot) \\ (x_i)_{i \in C_k} \mid \theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k, \Sigma_0) \end{aligned}$$

Conjugate exponential likelihood

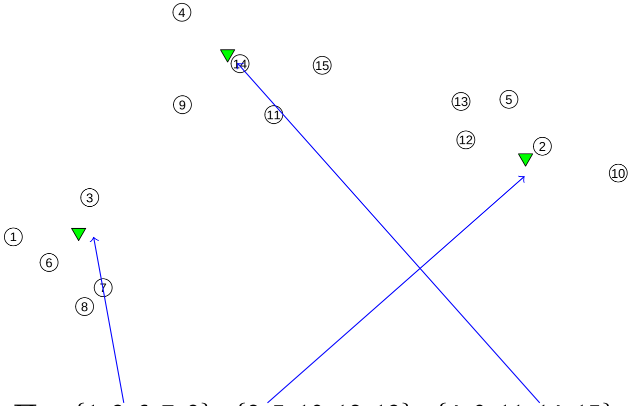


$$\mathbf{\Pi} = \{1,3,6,7,8\},\ \{2,5,10,12,13\},\ \{4,9,11,14,15\}$$

$$\theta_1,\dots,\theta_K \overset{\text{iid}}{\sim} \mathrm{Normal}(\cdot)\times\mathrm{Wishart}(\cdot)$$

$$(x_i)_{i\in C_k} \mid \theta\text{'s} \overset{\text{iid}}{\sim} \mathrm{Normal}(\theta_k^{\mu},\theta_k^{\Sigma})$$

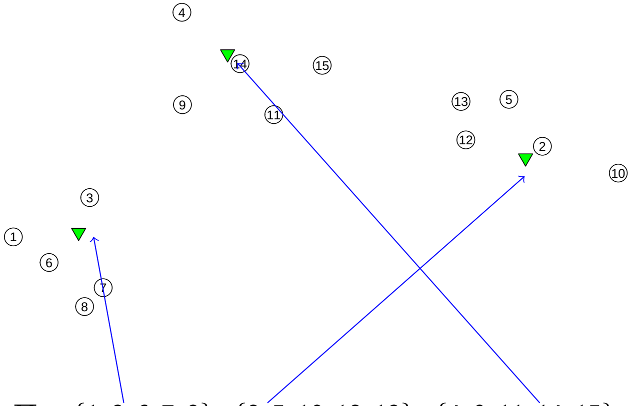
Conjugate exponential likelihood



$$\boldsymbol{\Pi} = \{1,3,6,7,8\},\ \{2,5,10,12,13\},\ \{4,9,11,14,15\}$$

$$\begin{aligned}\theta_1,\dots,\theta_K &\overset{\text{iid}}{\sim} H(\theta)\exp\{\chi\cdot\theta - \tau C(\theta) - A(\chi,\tau)\}\\ (x_i)_{i\in C_k}\mid\theta\text{'s} &\overset{\text{iid}}{\sim} h(x)\exp\{T(x)\cdot\theta - C(\theta)\}\end{aligned}$$

Conjugate exponential likelihood



$$\mathbf{\Pi} = \{1,3,6,7,8\},\ \{2,5,10,12,13\},\ \{4,9,11,14,15\}$$

$$\begin{aligned} \theta_1,\ldots,\theta_K &\overset{\text{iid}}{\sim} H(\theta)\exp\{\chi\cdot\theta - \tau \mathsf{C}(\theta) - \underbrace{\mathsf{A}}_{\text{par. log-partition}}(\chi,\tau)\} \\ (x_i)_{i\in C_k} \mid \theta\text{'s} &\overset{\text{iid}}{\sim} h(x)\exp\{\underbrace{\mathsf{T}}_{\text{sufficient stat.}}(x)\cdot\theta - \underbrace{\mathsf{C}}_{\text{obs. log-partition}}(\theta)\} \end{aligned}$$

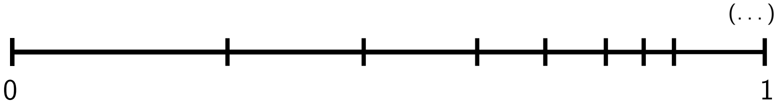
Exchangeable Random Partition (ERP) prior on partitions

Exchangeable Random Partition (ERP) prior on partitions

1. Sample a division of $[0, 1]$ into segments (possibly ∞):

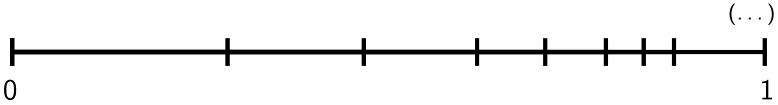
Exchangeable Random Partition (ERP) prior on partitions

1. Sample a division of $[0, 1]$ into segments (possibly ∞):



Exchangeable Random Partition (ERP) prior on partitions

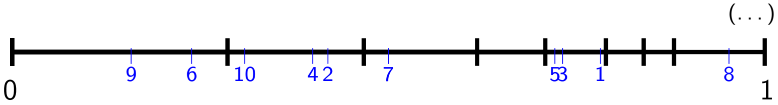
1. Sample a division of $[0, 1]$ into segments (possibly ∞):



2. Sample $X_1, \dots, X_n \sim \mathrm{Unif}([0, 1])$

Exchangeable Random Partition (ERP) prior on partitions

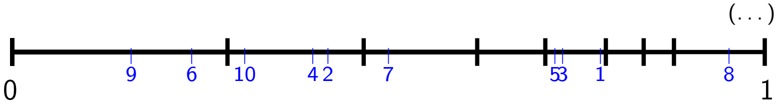
1. Sample a division of $[0, 1]$ into segments (possibly ∞):



2. Sample $X_1, \ldots, X_n \sim \text{Unif}([0, 1])$

Exchangeable Random Partition (ERP) prior on partitions

1. Sample a division of $[0, 1]$ into segments (possibly ∞):

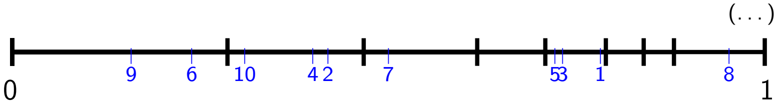


2. Sample $X_1, \ldots, X_n \sim \text{Unif}([0, 1])$

3. Read the partition:

Exchangeable Random Partition (ERP) prior on partitions

1. Sample a division of $[0,1]$ into segments (possibly ∞):



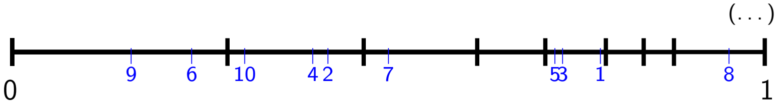
2. Sample $X_1, \ldots, X_n \sim \mathrm{Unif}([0,1])$

3. Read the partition:

$$\{1,3,5\},\ \{2,4,10\},\ \{6,9\},\ \{7\},\ \{8\}$$

Exchangeable Random Partition (ERP) prior on partitions

1. Sample a division of $[0, 1]$ into segments (possibly ∞):



2. Sample $X_1, \ldots, X_n \sim \mathrm{Unif}([0, 1])$

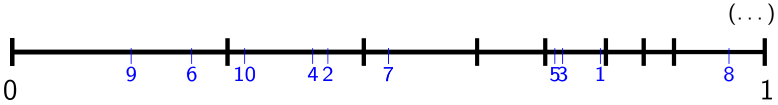
3. Read the partition:

$$\{1, 3, 5\},\ \{2, 4, 10\},\ \{6, 9\},\ \{7\},\ \{8\}$$

Exchangeable **R**andom **P**artition

Exchangeable Random Partition (ERP) prior on partitions

1. Sample a division of $[0, 1]$ into segments (possibly ∞):



2. Sample $X_1, \ldots, X_n \sim \mathrm{Unif}([0, 1])$

3. Read the partition:

$$\{1, 3, 5\}, \ \{2, 4, 10\}, \ \{6, 9\}, \ \{7\}, \ \{8\}$$

Exchangeable **R**andom **P**artition

e.g. the Chinese Restaurant Process, Pitman-Yor Process

Separability of the MAP

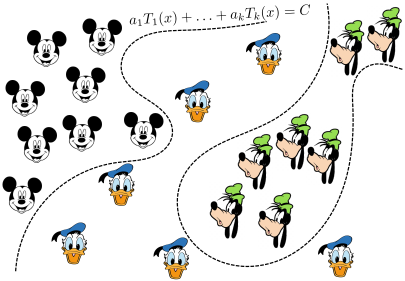
Theorem

For every pairwise distinct $x_1, \ldots, x_n \in \mathbb{R}^d$ and ex. part. $\mathbf{\Pi}$
the clusters of MAP in general exponential scheme are separated by the contour lines of linear functionals of T.

Separability of the MAP

THEOREM

For every pairwise distinct $x_1, \ldots, x_n \in \mathbb{R}^d$ and ex. part. $\mathbf{\Pi}$
the clusters of MAP in general exponential scheme are separated by the contour lines of linear functionals of T.



Separability of the MAP

Theorem

For every pairwise distinct $x_1, \ldots, x_n \in \mathbb{R}^d$ and ex. part. $\mathbf{\Pi}$ the clusters of MAP in general exponential scheme are separated by the contour lines of linear functionals of T.

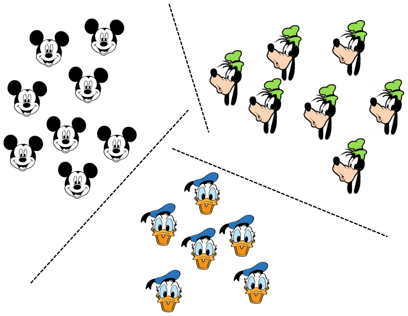
- the MAP in Normal-location scheme yields linear separability

Separability of the MAP

Theorem

For every pairwise distinct $x_1, \ldots, x_n \in \mathbb{R}^d$ and ex. part. $\mathbf{\Pi}$ the clusters of MAP in general exponential scheme are separated by the contour lines of linear functionals of T.

- the MAP in Normal-location scheme yields linear separability



Separability of the MAP

Theorem

For every pairwise distinct $x_1, \ldots, x_n \in \mathbb{R}^d$ and ex. part. $\mathbf{\Pi}$ the clusters of MAP in general exponential scheme are separated by the contour lines of linear functionals of T.

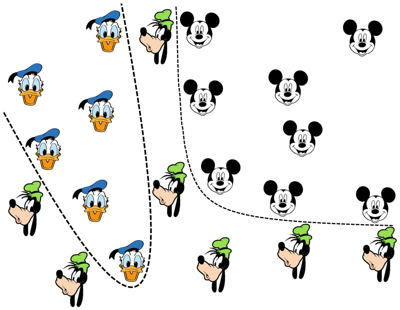
- the MAP in Normal-location scheme yields linear separability
- the MAP in Normal-location-scale scheme yields quadratic separability

Separability of the MAP

Theorem

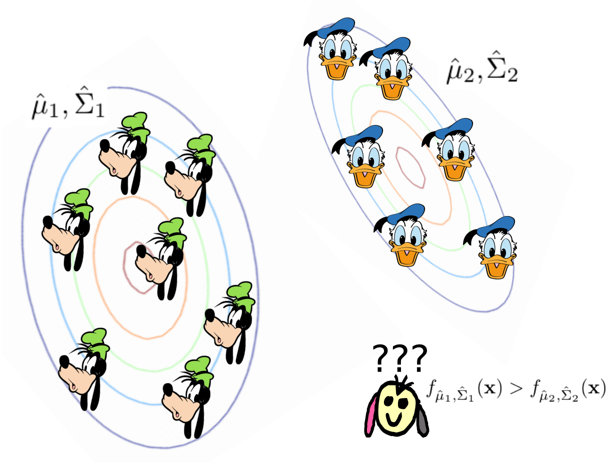
For every pairwise distinct $x_1, \ldots, x_n \in \mathbb{R}^d$ and ex. part. $\mathbf{\Pi}$ the clusters of MAP in general exponential scheme are separated by the contour lines of linear functionals of T.

- the MAP in Normal-location scheme yields linear separability
- the MAP in Normal-location-scale scheme yields quadratic separability



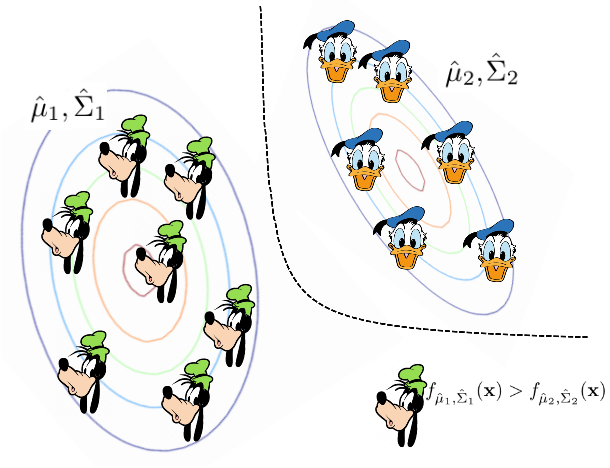
Analogy to Fisher Discriminant Analysis

Fisher Discriminant Analysis is a technique for **supervised** learning (classification, not clustering)



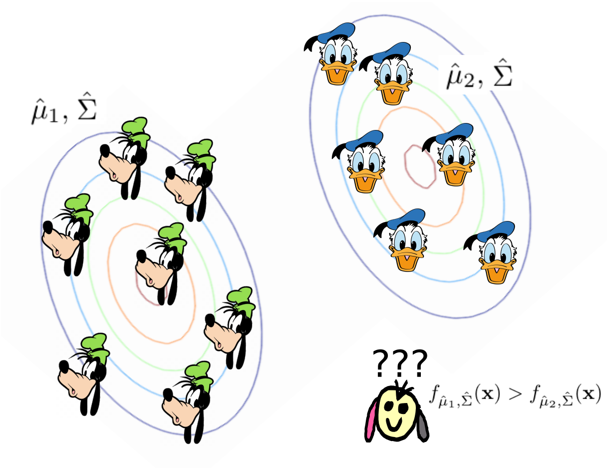
Analogy to Fisher Discriminant Analysis

Fisher Discriminant Analysis is a technique for **supervised** learning (classification, not clustering)



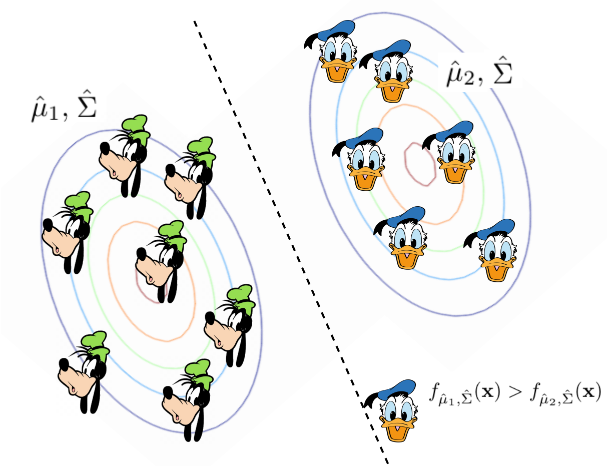
Analogy to Fisher Discriminant Analysis

Fisher Discriminant Analysis is a technique for **supervised** learning (classification, not clustering)



Analogy to Fisher Discriminant Analysis

Fisher Discriminant Analysis is a technique for **supervised** learning (classification, not clustering)



The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\big(\sum_{i \in I} z_i\big)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\} \setminus \hat{I}}$ are lin. sep.

The only proof in this presentation

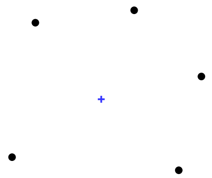
Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \dots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \dots, m\}$ maximises $h\left(\sum_{i \in I} z_i\right)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\dots,m\}\setminus\hat{I}}$ are lin. sep.

Example:

$m = 5$

$k = 2$

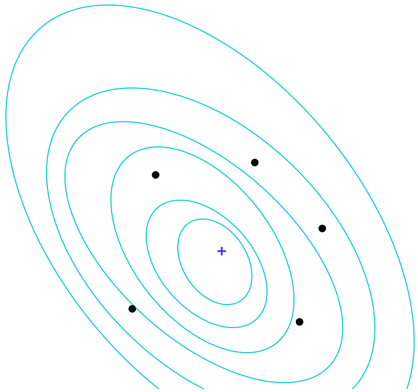


The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\big(\sum_{i \in I} z_i\big)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

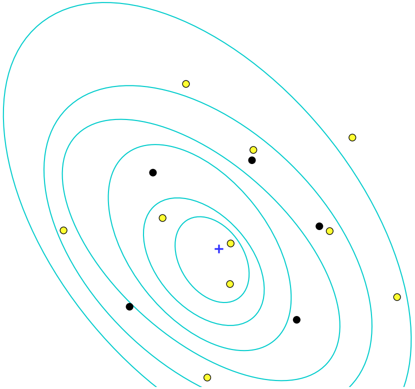


The only proof in this presentation

Lemma

If $h \colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\left(\sum_{i \in I} z_i\right)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

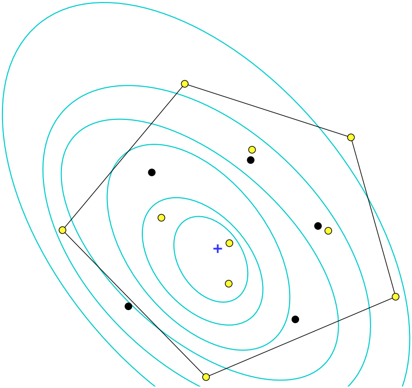


The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\big(\sum_{i \in I} z_i\big)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\} \setminus \hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

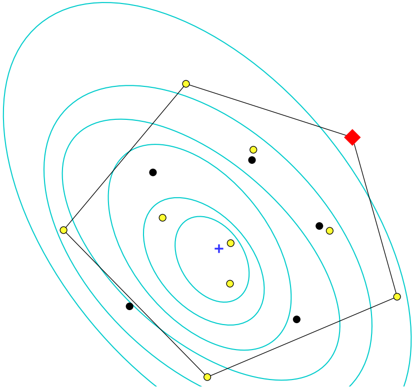


The only proof in this presentation

Lemma

If $h \colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\big(\sum_{i \in I} z_i\big)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

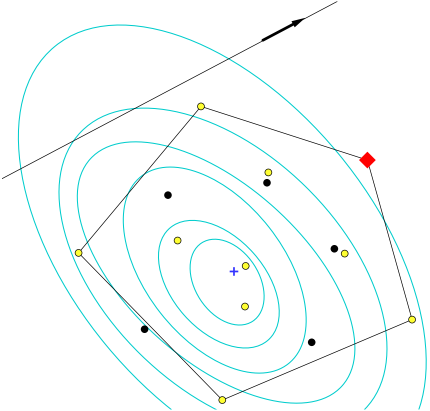


The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\big(\sum_{i \in I} z_i\big)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

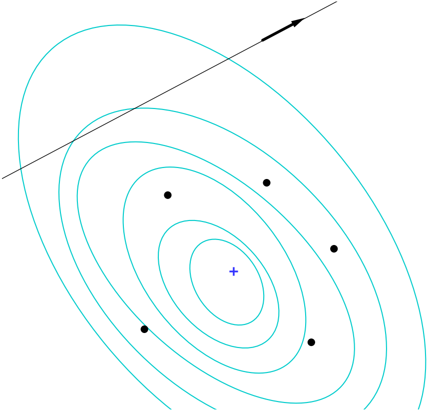


The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\big(\sum_{i \in I} z_i\big)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

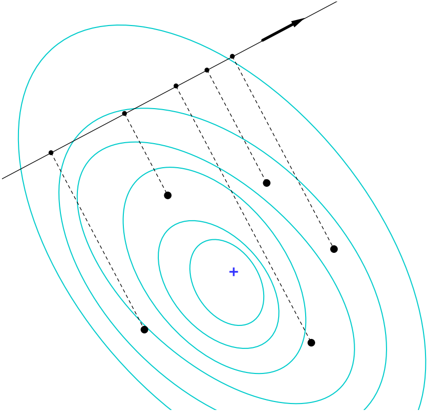


The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\left(\sum_{i \in I} z_i\right)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

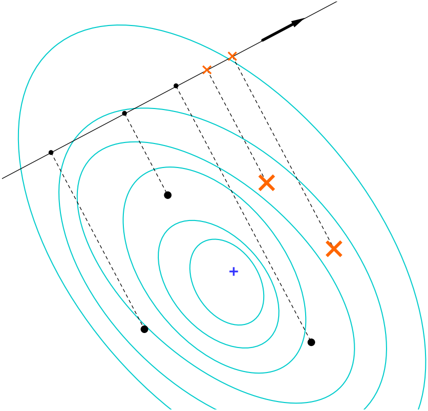


The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\big(\sum_{i \in I} z_i\big)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

Example:
$m = 5$
$k = 2$

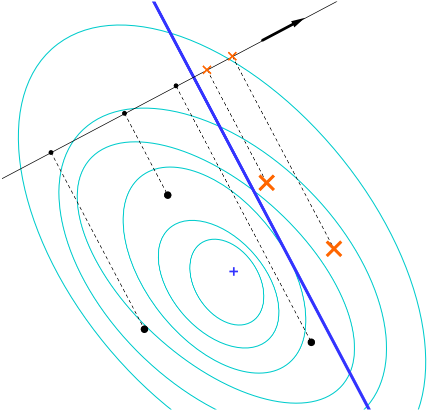


The only proof in this presentation

Lemma

If $h\colon \mathbb{R}^D \to \mathbb{R}$ is convex $z_1, \ldots, z_m \in \mathbb{R}^D$, $k \leqslant m$ and $\hat{I} \subset \{1, \ldots, m\}$ maximises $h\left(\sum_{i \in I} z_i\right)$ over $|I| = k$ then $\boldsymbol{z}_{\hat{I}}$ and $\boldsymbol{z}_{\{1,\ldots,m\}\setminus\hat{I}}$ are lin. sep.

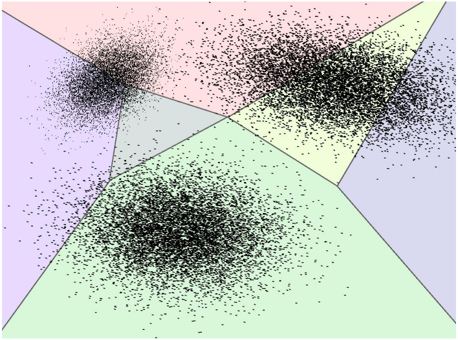
Example:
$m = 5$
$k = 2$



Induced partitions

Let $\mathcal{A}$ be a **fixed** partition of $\mathbb{R}^d$:

Let $X_1, X_2, \ldots, X_{1000} \overset{\text{iid}}{\sim} P$



$$\mathcal{I}^{\mathcal{A}}_{10000}(\boldsymbol{X}_{1:10000}) = \{\{\ldots\}, \{\ldots\}, \{\ldots\}, \{\ldots\}, \{\ldots\}\}$$
$$\mathbb{P}(\mathcal{I}^{\mathcal{A}}_{10000}(\boldsymbol{X}_{1:10000}) \mid \boldsymbol{X}_{1:10000}) \approx ???$$

Induced partitions

Proposition (in previous Gaussian CRP model)

$\sqrt[n]{\mathbb{P}(\mathcal{I}_n^{\mathcal{A}}(\boldsymbol{X}_{1:n}) \mid \boldsymbol{X}_{1:n})} \overset{\text{a.s.}}{\asymp} \exp\{\Delta_P(\mathcal{A})\}$ where

$$\Delta_P(\mathcal{A}) = \underbrace{\sum_{A \in \mathcal{A}} P(A) \ln P(A)}_{\log \sqrt[n]{\text{CRP prior}}} + \underbrace{\frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \|\mathbb{E}\left(\Sigma_0^{-1} X \mid X \in A\right)\|^2}_{\log \sqrt[n]{\text{Gaussian Likelihood}}}$$

straightforward computations using SLLN

Induced partitions

Proposition (in general exponential ERP model)

$\sqrt[n]{\mathbb{P}(\mathcal{I}_{\boldsymbol{n}}^{\mathcal{A}}(\boldsymbol{X}_{1:n}) \mid \boldsymbol{X}_{1:n})} \overset{\text{a.s.}}{\asymp} \exp\{\Delta_P(\mathcal{A})\}$ where

$$\Delta_P(\mathcal{A}) = \underbrace{\sum_{A \in \mathcal{A}} P(A) \ln P(A)}_{\log \sqrt[n]{\text{ERP prior}}} + \underbrace{\sum_{A \in \mathcal{A}} P(A) \cdot \mathrm{C}^* \Big(\mathbb{E}\left(T(X) \mid X \in A \right) \Big)}_{\log \sqrt[n]{\text{Exponential Likelihood}}}$$

$C^*(t) = \sup_\theta \big(t \cdot \theta - C(\theta) \big)$

not that straightforward analysis

Induced partitions

Proposition (in general exponential ERP model)

$\sqrt[n]{\mathbb{P}(\mathcal{I}_n^{\mathcal{A}}(\boldsymbol{X}_{1:n}) \mid \boldsymbol{X}_{1:n})} \overset{\text{a.s.}}{\asymp} \exp\{\Delta_P(\mathcal{A})\}$ where

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) + \sum_{A \in \mathcal{A}} P(A) \cdot \mathrm{C}^*\Big(\mathbb{E}\,(T(X) \mid X \in A)\Big)$$

$\log \sqrt[n]{\text{ERP prior}}$

$\log \sqrt[n]{\text{Exponential Likelihood}}$

$C^*(t) = \sup_\theta \big(t \cdot \theta - C(\theta)\big)$

not that straightforward analysis

Both limits can be expressed as $\|f_n\|_{L^n(\mathcal{X},\mu)} \to \|f\|_{L^\infty(\mathcal{X},\mu)}$,
(where $f_n \to f$ pointwise)

Uniform example

(A) Normal, fixed covariance:

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) + \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \|\mathbb{E}\,(\Sigma_0^{-1} X \mid X \in A)\|^2$$

(B) Normal, random (Wishart) covariance

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) - \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \ln \det \big(\mathsf{V}(X \mid X \in A)\big)$$

Uniform example

(A) Normal, fixed covariance:

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) + \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\| \mathbb{E}\left(\Sigma_0^{-1} X \mid X \in A\right) \right\|^2$$

(B) Normal, random (Wishart) covariance

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) - \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \ln\det\left(\mathrm{V}(X \mid X \in A)\right)$$

What are the maximisers for $P = \mathrm{Unif}([0,1])$?

Uniform example

(A) Normal, fixed covariance:

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) + \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \left\| \mathbb{E}\left(\Sigma_0^{-1} X \mid X \in A\right) \right\|^2$$

(B) Normal, random (Wishart) covariance

$$\Delta_P(\mathcal{A}) = \sum_{A \in \mathcal{A}} P(A) \ln P(A) - \frac{1}{2} \sum_{A \in \mathcal{A}} P(A) \cdot \ln \det \left(\mathrm{V}(X \mid X \in A)\right)$$

What are the maximisers for $P = \mathrm{Unif}([0, 1])$? 

- they are divisions into subsegments 

Uniform example

(A) Normal, fixed covariance:

$$\Delta(p_1, \ldots, p_n) = \sum_{i \leqslant n} p_i \ln p_i + \frac{1}{2\Sigma_0} \sum_i p_i \cdot \Big(\frac{p_i}{2} \sum_{j<i} p_j\Big)^2$$

(B) Normal, random (Wishart) covariance

$$\Delta(p_1, \ldots, p_n) = \sum_{i \leqslant n} p_i \ln p_i - \frac{1}{2} \sum_{i \leqslant n} p_i \cdot \ln \frac{{p_i}^2}{12}$$

What are the maximisers for $P = \mathrm{Unif}([0,1])$? 

- they are divisions into subsegments 

Uniform example

(A) Normal, fixed covariance:

$$\Delta(p_1, \ldots, p_n) = \sum_{i \leqslant n} p_i \ln p_i + \frac{1}{2\Sigma_0} \sum_i p_i \cdot \Big(\frac{p_i}{2} \sum_{j<i} p_j \Big)^2$$

(B) Normal, random (Wishart) covariance

$$\Delta(p_1, \ldots, p_n) = \sum_{i \leqslant n} p_i \ln p_i - \frac{1}{2} \sum_{i \leqslant n} p_i \cdot \ln \frac{{p_i}^2}{12}$$

What are the maximisers for $P = \mathrm{Unif}([0, 1])$? 

- they are divisions into subsegments 

(A) division into segments of equal length, such that the within cluster variance is Σ_0

Uniform example

(A) Normal, fixed covariance:

$$\Delta(p_1, \dots, p_n) = \sum_{i \leqslant n} p_i \ln p_i + \frac{1}{2\Sigma_0} \sum_i p_i \cdot \left(\frac{p_i}{2} \sum_{j<i} p_j \right)^2$$

(B) Normal, random (Wishart) covariance

$$\Delta(p_1, \dots, p_n) = \sum_{i \leqslant n} p_i \ln p_i - \frac{1}{2} \sum_{i \leqslant n} p_i \cdot \ln \frac{{p_i}^2}{12}$$

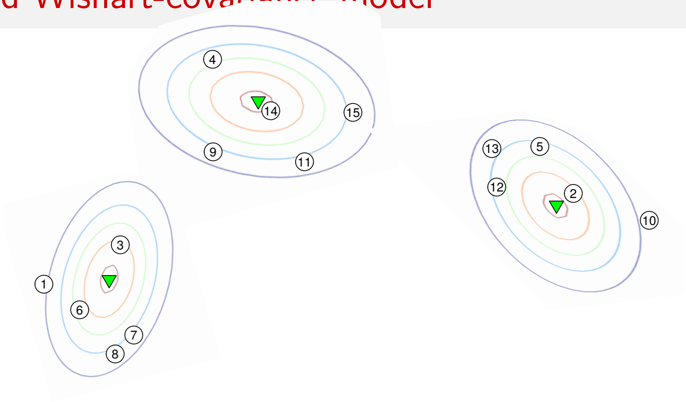
What are the maximisers for $P = \mathrm{Unif}([0,1])$? 

- they are divisions into subsegments 

(A) division into segments of equal length, such that the within cluster variance is Σ_0

(B) **every division into subsegments gives the same (maximum) score!**

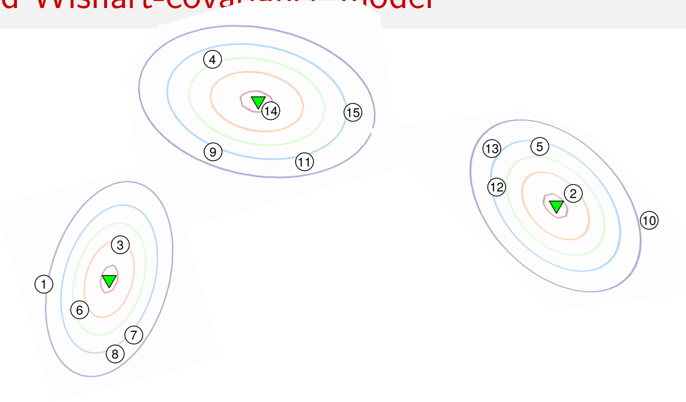
Adjusted Wishart-covariance model



$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\begin{aligned} \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot) \times \text{Wishart}(\cdot) \\ (x_i)_{i \in C_k} \,|\, \theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k^{\mu}, \theta_k^{\Sigma}) \end{aligned}$$

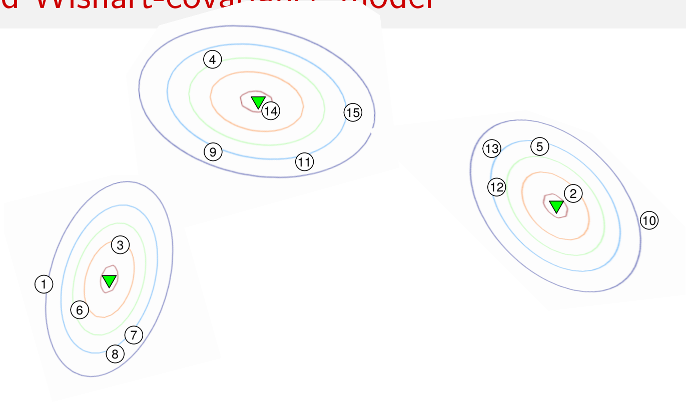
Adjusted Wishart-covariance model



$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\begin{aligned} \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot) \times \text{Wishart}(\Sigma_0, \eta_0) \\ (x_i)_{i \in C_k} \,|\, \theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k^{\mu}, \theta_k^{\Sigma}) \end{aligned}$$

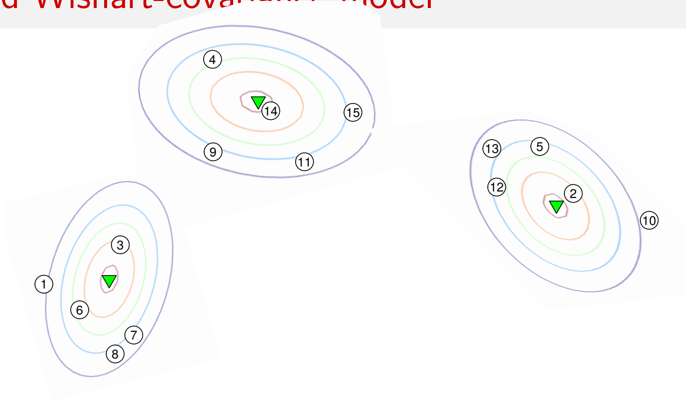
Adjusted Wishart-covariance model



$$\mathbf{\Pi} = \{1,3,6,7,8\},\ \{2,5,10,12,13\},\ \{4,9,11,14,15\}$$

$$\begin{aligned} \theta_1,\ldots,\theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot)\times\text{Wishart}(\Sigma_0,\eta_0) \\ (x_i)_{i\in C_k}\,|\,\theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k^{\mu},\theta_k^{\Sigma}) \end{aligned}$$

Adjusted Wishart-covariance model



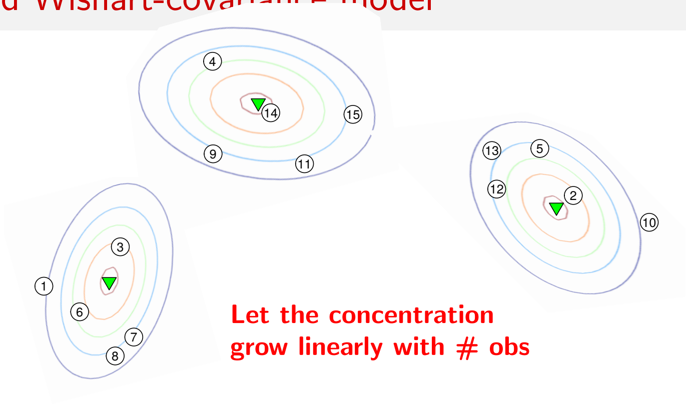
$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\begin{aligned} \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot) \times \text{Wishart}(\Sigma_0, \eta_0) \\ (x_i)_{i \in C_k} \mid \theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k^\mu, \theta_k^\Sigma) \end{aligned}$$

concentration

expected covariance

Adjusted Wishart-covariance model



Let the concentration grow linearly with # obs

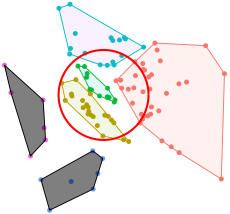
$$\mathbf{\Pi} = \{1, 3, 6, 7, 8\},\ \{2, 5, 10, 12, 13\},\ \{4, 9, 11, 14, 15\}$$

$$\begin{aligned} \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} \text{Normal}(\cdot) \times \text{Wishart}(\Sigma_0, \boldsymbol{\lambda n}) \\ (x_i)_{i \in C_k} \mid \theta\text{'s} &\overset{\text{iid}}{\sim} \text{Normal}(\theta_k^{\mu}, \theta_k^{\Sigma}) \end{aligned}$$

concentration

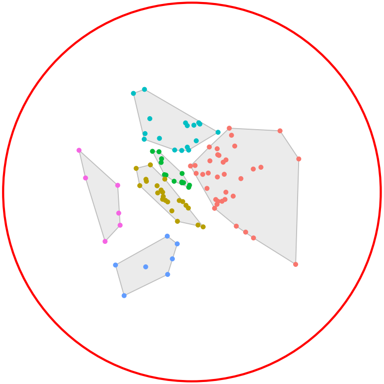
expected covariance

Linear growth of clusters for adjusted model



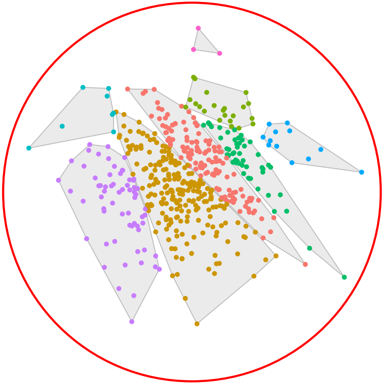
$n = 100$

Linear growth of clusters for adjusted model



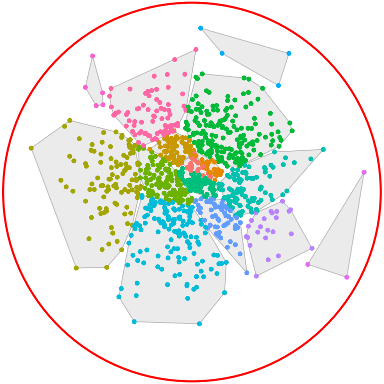
$n = 100$

Linear growth of clusters for adjusted model



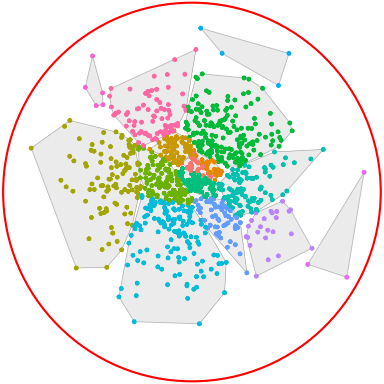
$n = 500$

Linear growth of clusters for adjusted model



$n = 1000$

Linear growth of clusters for adjusted model



$n = 1000$

Result for adjusted model and CRP prior

If $X_1, X_2, \ldots \sim P$, where P has a bounded support, then

$$\liminf_{n \to \infty} \min_{J \in \hat{\mathcal{I}}_{MAP}(\boldsymbol{X}_{1:n})} |J|/n > 0.$$

Δ_P function for the adjusted model

$$\begin{aligned}\Delta_{P,\lambda}(\mathcal{A}) = & \frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\ & - \frac{1}{2}\sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log\left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda}\mathsf{V}_P(X \mid X \in A)\right| + \\ & + \sum_{A\in\mathcal{A}} P(A)\log P(A)\end{aligned}$$

Δ_P function for the adjusted model

$$\begin{aligned}
\Delta_{P,\lambda}(\mathcal{A}) = &\frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\
&- \frac{1}{2}\sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log\left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda}\mathsf{V}_P(X \mid X\in A)\right| + \\
&+ \sum_{A\in\mathcal{A}} P(A)\log P(A)
\end{aligned}$$

Δ_P function for the adjusted model

$$\begin{aligned}\Delta_{P,\lambda}(\mathcal{A}) = &\frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\ &- \frac{1}{2}\sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log\left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda}\mathsf{V}_P(X \mid X\in A)\right| + \\ &+ \sum_{A\in\mathcal{A}} P(A)\log P(A)\end{aligned}$$

Δ_P function for the adjusted model

$$\Delta_{P,\lambda}(\mathcal{A}) = \frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} -$$
$$- \frac{1}{2}\sum_{A \in \mathcal{A}} \big(P(A) + \lambda\big) \log \left| \frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda} \mathsf{V}_P(X \mid X \in A) \right| +$$
$$+ \sum_{A \in \mathcal{A}} P(A) \log P(A)$$

Δ_P function for the adjusted model

$$
\begin{aligned}
\Delta_{P,\lambda}(\mathcal{A}) = {} & \frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\
& - \frac{1}{2} \sum_{A \in \mathcal{A}} \big(P(A) + \lambda\big) \log\left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda} \mathsf{V}_P(X \mid X \in A)\right| + \\
& + \sum_{A \in \mathcal{A}} P(A) \log P(A)
\end{aligned}
$$

Δ_P function for the adjusted model

$$\begin{aligned}\Delta_{P,\lambda}(\mathcal{A}) = {} & \frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\ & - \frac{1}{2}\sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log\left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda}\mathsf{V}_P(X \mid X \in A)\right| + \\ & + \sum_{A\in\mathcal{A}} P(A)\log P(A)\end{aligned}$$

- as $\lambda \to 0$, this approaches random-covariance Δ

Δ_P function for the adjusted model

$$\Delta_{P,\lambda}(\mathcal{A}) = \frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\ - \frac{1}{2}\sum_{A\in\mathcal{A}} \big(P(A)+\lambda\big) \log\left|\frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda}\mathsf{V}_P(X \mid X \in A)\right| + \\ + \sum_{A\in\mathcal{A}} P(A)\log P(A)$$

- as $\lambda \to 0$, this approaches random-covariance Δ
- as $\lambda \to \infty$, this approaches fixed-covariance Δ

Δ_P function for the adjusted model

$$\begin{aligned}\Delta_{P,\lambda}(\mathcal{A}) = {} & \frac{1}{2}|\mathcal{A}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\ & - \frac{1}{2}\sum_{A \in \mathcal{A}} \big(P(A) + \lambda\big) \log \left| \frac{\lambda}{P(A)+\lambda}\Sigma_0 + \frac{P(A)}{P(A)+\lambda} \mathrm{V}_P(X \mid X \in A) \right| + \\ & + \sum_{A \in \mathcal{A}} P(A) \log P(A)\end{aligned}$$

- as $\lambda \to 0$, this approaches random-covariance Δ
- as $\lambda \to \infty$, this approaches fixed-covariance Δ

Maybe use its empirical equivalent to „score" clustering proposals?

Δ_P function for the adjusted model

$$\widehat{\Delta}_{P,\lambda}(\mathcal{J}) = \frac{1}{2}|\mathcal{J}| \cdot \lambda \log|\Sigma_0| - \frac{d}{2} - \\ - \frac{1}{2} \sum_{J \in \mathcal{J}} \left(\hat{p}_J + \lambda\right) \log\left|\frac{\lambda}{\hat{p}_J + \lambda}\Sigma_0 + \frac{\hat{p}_J}{\hat{p}_J + \lambda}\hat{\mathrm{V}}_J\right| + \\ + \sum_{J \in \mathcal{J}} \hat{p}_J \log \hat{p}_J$$

- as $\lambda \to 0$, this approaches random-covariance Δ
- as $\lambda \to \infty$, this approaches fixed-covariance Δ

Maybe use its empirical equivalent to „score" clustering proposals?

Δ_P function for the adjusted model

$$
\begin{aligned}
\widehat{\Delta}_{P,\lambda}(\mathcal{J}) = & \frac{1}{2}|\mathcal{J}| \cdot \lambda \log |\Sigma_0| - \frac{d}{2} - \\
& - \frac{1}{2} \sum_{J \in \mathcal{J}} \left(\hat{p}_J + \lambda\right) \log \left| \frac{\lambda}{\hat{p}_J + \lambda} \Sigma_0 + \frac{\hat{p}_J}{\hat{p}_J + \lambda} \hat{V}_J \right| + \\
& + \sum_{J \in \mathcal{J}} \hat{p}_J \log \hat{p}_J
\end{aligned}
$$

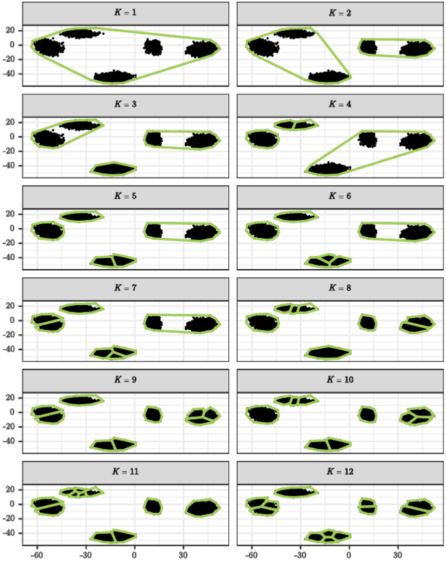
- as $\lambda \to 0$, this approaches random-covariance Δ
- as $\lambda \to \infty$, this approaches fixed-covariance Δ

Maybe use its empirical equivalent to „score" clustering proposals?

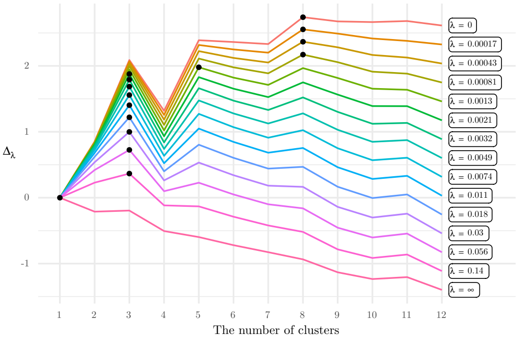
We choose Σ_0 to be the total covariance matrix

- ...its a natural upper bound for Σ_0
- ...then the value for $\mathcal{J} = \{[n]\}$ is the same for every λ

K-means divisions of 5 Gaussian-clusters dataset

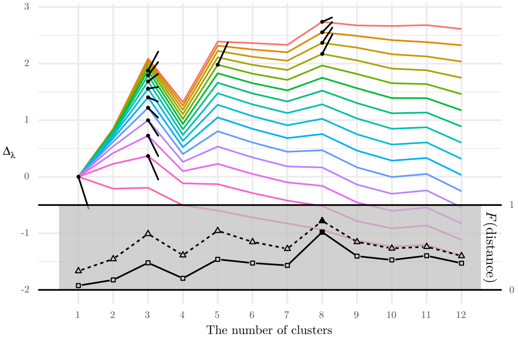


Scoring the divisions using $\widehat{\Delta}_\lambda$



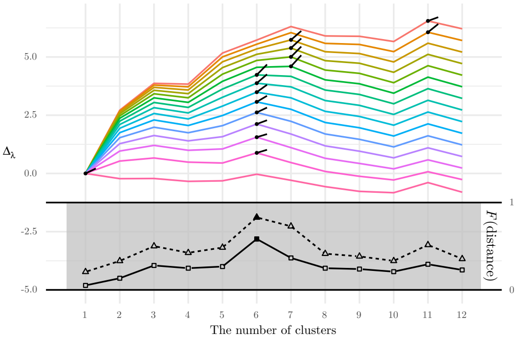
Black dots denote maximums

K-means divisions of 5 Gaussian-clusters dataset



Two curves on gray area represent the distance to the 'true' clustering.

4 dimensional example of 7 clusters



Quite representative; there is a range of λ's for which we have a good choice.

Summary

- Introduction: definitions and notation
  Bayesian models for clustering, MAP
- Results of R. (2019), for Gaussian CRP model with fixed covariance
  - linear separability of clusters
  - linear growth of clusters (interesecting a fixed ball)
  - limit formula for induced partitions
  - converge result for convex hulls of clusters
- Generalisations
  - separability of clusters in general exponential ERP
  - limit formula for induced partitions in general exponential ERP
  - linear growth of clusters for adjusted Wishart-covariance model and bounded input
- Applications
  Using empirical version of adjusted Wishart-covariance Δ to score clustering proposals

Thank you for your attention