



DE CASTRO
STATISTICS

Collegio Carlo Alberto
UNIVERSITÀ DEGLI STUDI DI TORINO

From infinity to here: a Bayesian nonparametric perspective of finite mixture models

Raffaele Argiento

ESOMAS Department University of Torino and Collegio Carlo Alberto

Wien, May 17th – 2019

joint with Maria de Iorio (Yale-NUS Singapore)



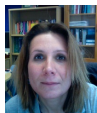
DE CASTRO
STATISTICS

Collegio Carlo Alberto
UNIVERSITÀ DEGLI STUDI DI TORINO

From infinity to here: a Bayesian nonparametric perspective of finite mixture models

Raffaele Argiento

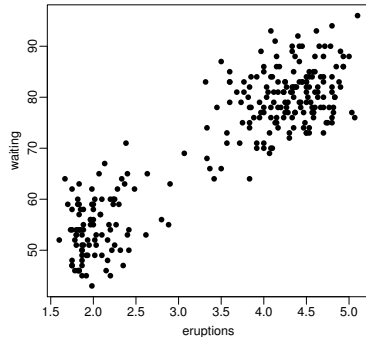
ESOMAS Department University of Torino and Collegio Carlo Alberto



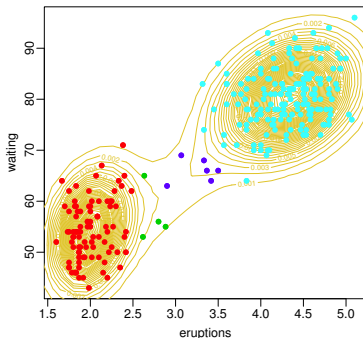
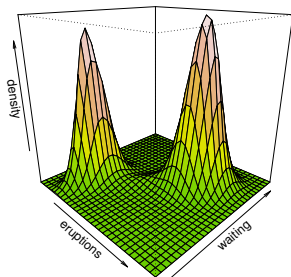
Wien, May 17th – 2019

joint with Maria de Iorio (Yale-NUS Singapore)

- Mixture models are a very powerful and natural statistical tool to model data from heterogeneous populations.
- Observations are assumed to have arisen from one of M (finite or infinite) groups, each group being suitably modelled by a density typically from a parametric family.
- The density of each group is referred to as a component of the mixture and is weighted by the relative frequency (weight) of the group in the population.



- Mixture models are a very powerful and natural statistical tool to model data from heterogeneous populations.
- Observations are assumed to have arisen from one of M (finite or infinite) groups, each group being suitably modelled by a density typically from a parametric family.
- The density of each group is referred to as a component of the mixture and is weighted by the relative frequency (weight) of the group in the population.
- The statistical goals are density estimation and cluster analysis (see Fruhwirth-Schnatter et al. 2019).



Hierarchical representation

$$X_1, \dots, X_n \mid \mathbf{w}, \boldsymbol{\tau} \stackrel{iid}{\sim} \sum_{h=1}^M w_h f(x \mid \tau_h)$$

$$\mathbf{w} \mid M \sim \text{Dirichlet}_M(\gamma, \dots, \gamma)$$

$$\tau_h \mid M \stackrel{iid}{\sim} P_0(d\tau), \quad M \sim q_M$$

Hierarchical representation

$$X_1, \dots, X_n \mid j_1, \dots, j_n \stackrel{iid}{\sim} f(x \mid \tau_{j_i})$$

$$j_1, \dots, j_n \mid \mathbf{w} \stackrel{iid}{\sim} \text{Multinomial}_M(1, w_1, \dots, w_M)$$

$$\mathbf{w} \mid M \sim \text{Dirichlet}_M(\gamma, \dots, \gamma)$$

$$\tau_h \mid M \stackrel{iid}{\sim} P_0(d\tau_h), \quad M \sim q_M$$

Hierarchical representation

$$\begin{aligned}
 X_1, \dots, X_n \mid \theta_1, \dots, \theta_n &\stackrel{\text{ind}}{\sim} f(x_i \mid \theta_i), \quad \theta_i = \tau_{j_i} \\
 \theta_1, \dots, \theta_n \mid P &\stackrel{\text{iid}}{\sim} P, \quad P(\cdot) \stackrel{d}{=} \sum_{h=1}^M w_h \delta_{\tau_h}(\cdot) \\
 \mathbf{w} \mid M &\sim \text{Dirichlet}_M(\gamma, \dots, \gamma) \\
 \tau_h \mid M &\stackrel{\text{iid}}{\sim} P_0(d\tau_h), \quad M \sim q_M
 \end{aligned}$$

Hierarchical representation

$$X_1, \dots, X_n \mid \theta_1, \dots, \theta_n \stackrel{\text{ind}}{\sim} f(x_i \mid \theta_i)$$

$$\theta_1, \dots, \theta_n \mid P \stackrel{\text{iid}}{\sim} P, \quad P(\cdot) \stackrel{d}{=} \sum_{h=1}^M w_h \delta_{\tau_h}(\cdot)$$

$$P \sim FDP$$

Hierarchical representation

$$\begin{aligned}
 X_1, \dots, X_n \mid \theta_1, \dots, \theta_n &\stackrel{ind}{\sim} f(x_i \mid \theta_i) \\
 \theta_1, \dots, \theta_n \mid P &\stackrel{iid}{\sim} P, \quad P(\cdot) \stackrel{d}{=} \sum_{h=1}^M w_h \delta_{\tau_h}(\cdot) \\
 P &\sim FDP
 \end{aligned}$$

- ✓ The density f_P of the population variable X is **random**.
- ✓ The law of this random density is assigned by a mixture model:

$$X|P \sim f_P(x) = \int_{\Theta} f(x; \theta) P(d\theta) = \sum_{h=1}^M w_h f(x, \tau_h)$$

Targets:

★ **Density estimation:** $\mathcal{L}(f_P | X_1, \dots, X_n)$

★ **Cluster analysis:** $\mathcal{L}(\rho | X_1, \dots, X_n)$

where ρ is the random partition induced by P .

Hierarchical representation

$$\begin{aligned}
 X_1, \dots, X_n \mid \theta_1, \dots, \theta_n &\stackrel{\text{ind}}{\sim} f(x_i \mid \theta_i) \\
 \theta_1, \dots, \theta_n \mid P &\stackrel{\text{iid}}{\sim} P, \quad P(\cdot) \stackrel{d}{=} \sum_{h=1}^M w_h \delta_{\tau_h}(\cdot) \\
 P &\sim \text{Norm} - \text{IFPP}
 \end{aligned}$$

- ✓ The density f_P of the population variable X is **random**.
- ✓ The law of this random density is assigned by a mixture model:

$$X|P \sim f_P(x) = \int_{\Theta} f(x; \theta) P(d\theta) = \sum_{h=1}^M w_h f(x, \tau_h)$$

Targets:

★ **Density estimation:** $\mathcal{L}(f_P | X_1, \dots, X_n)$

★ **Cluster analysis:** $\mathcal{L}(\rho | X_1, \dots, X_n)$

where ρ is the random partition induced by P .

In this work:

- (a) we introduce a general class of prior for P
- (b) we set up a easy blocked Gibbs sampler.

Prior for the mixing distribution:

$$P(\cdot) = \sum_{j=1}^M w_j \delta_{\tau_j}(\cdot) \quad \text{M-FDP}$$

then $(w_1, \dots, w_M) \sim \text{Dirichlet}_M(\gamma, \dots, \gamma)$, $\gamma > 0$, $(\tau_1, \dots, \tau_M) \stackrel{iid}{\sim} P_0$.

- M is fixed: one fits several mixture models for $M = 1, 2, \dots, M^*$ then choose the best M according to some goodness of fit index.
- M is random: we need MCMCs that allow transitions across dimensions of the state space
 - ✓ **Reversible jump** ([Richardson and Green, 1997]).
 - ✓ **Point processes** representation of the posteriors distribution ([Stephens, 2000]).
 - ✓ Borrowing notation from nonparametric literature: **Marginal Gibbs sampler** ([Miller and Harrison, 2018]).

Prior for the mixing distribution:

$P \sim$ Dirichlet Process, $P \sim$ Normalized CRM, $P \sim$ Stick-breaking Priors.

Critical issues, infinite dimensional parameter $P = \sum_{i=1}^{\infty} w_i \delta_{\tau_i}$

Marginal Gibbs sampler algorithms [Neal, 2000] [Favaro e Teh, 2013]

- ✓ **Integrate out** P and resort to generalized Polya urn schemes
- ✓ Inference is limited to the point estimates: **predictive** $f_{X_{n+1}}(\cdot | X_1, \dots, X_n)$

Conditional methods

- ✓ Use some *tricks* to build a **Gibbs sampler** whose state space encompasses P .
- ✓ **Full Bayesian** posterior analysis.

For instance:

- ✓ **Slice sampler** [Kalli et al. 2009]
- ✓ **Retrospective methods** [Papaspiliopoulos et al., 2008]
- ✓ **Truncation** (either a-priori or a-posteriori) of the infinite sum defining the r.p.m. P [Argiento et al., 2010, Argiento et al., 2015a]

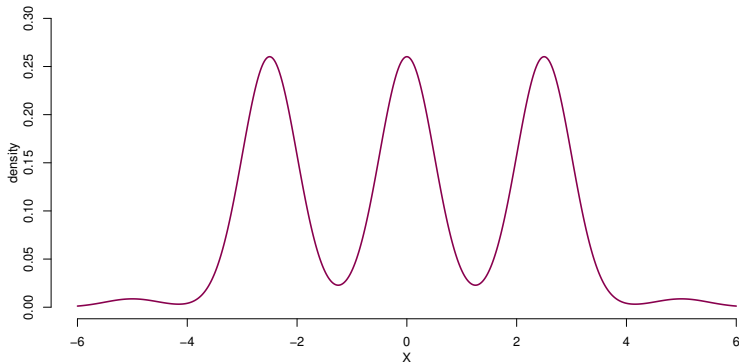
The number of components and the number of clusters

- ✎ It is important to stress the difference between components and clusters (Nobile, 2004; Rousseau and Mengersen, 2011; Frühwirth-Schnatter and Malsiner-Walli 2019).

The number of components and the number of clusters

✎ It is important to stress the difference between components and clusters (Nobile, 2004; Rousseau and Mengersen, 2011; Frühwirth-Schnatter and Malsiner-Walli 2019).

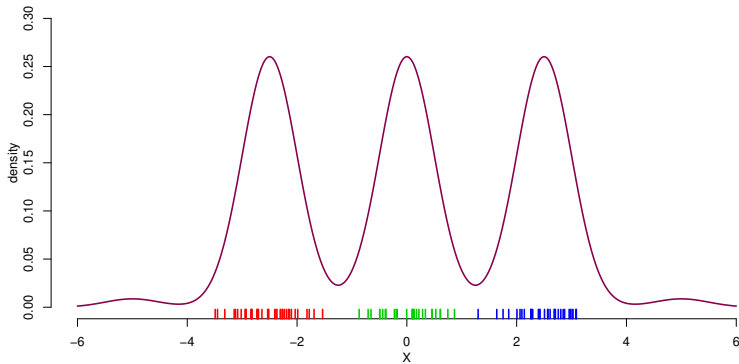
✓ This is a plot of a mixture density with $M = 5$ five components.



The number of components and the number of clusters

✎ It is important to stress the difference between components and clusters (Nobile, 2004; Rousseau and Mengersen, 2011; Frühwirth-Schnatter and Malsiner-Walli 2019).

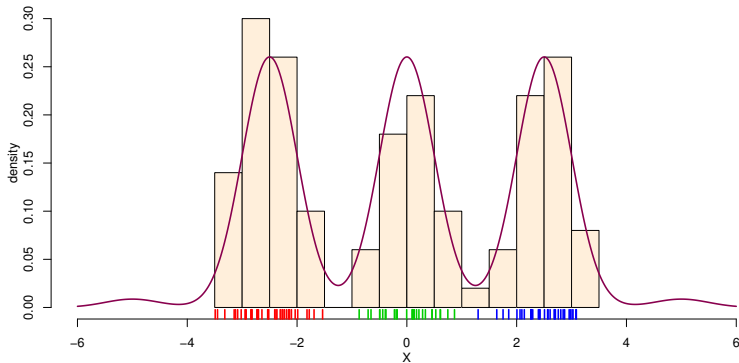
✓ I draw a sample of size 500 from the mixture



The number of components and the number of clusters

✎ It is important to stress the difference between components and clusters (Nobile, 2004; Rousseau and Mengersen, 2011; Frühwirth-Schnatter and Malsiner-Walli 2019).

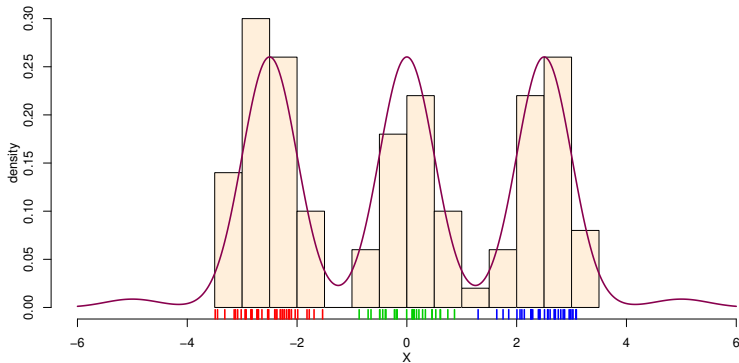
✓ The number of clusters are the allocated components, they are $K := M^{(a)} = 3$



The number of components and the number of clusters

✎ It is important to stress the difference between components and clusters (Nobile, 2004; Rousseau and Mengersen, 2011; Frühwirth-Schnatter and Malsiner-Walli 2019).

✓ The non-allocated components (empty) are $M^{(na)} := M - M^{(a)} = 2$.



- Normalized Independent Finite Point Processes (Norm-IFPP)
- Clustering induced by Norm-IFPP and posterior characterization
- Norm-IFPP mixtures
- Conditional Algorithm for Norm-IFPP
- Illustrative Example (Galaxy Data)

- A **finite point process** $\mathcal{S} = \{S_1, \dots, S_M\}$ is a random set of **unordered** points in a metric space \mathcal{S} [see Daley and Vere-Jones (2003)].
- The law of a finite point process is identified by:

- A **finite point process** $S = \{S_1, \dots, S_M\}$ is a random set of **unordered** points in a metric space \mathcal{S} [see Daley and Vere-Jones (2003)].
- The law of a finite point process is identified by:
 - ✓ $\{q_m, m = 0, 1, \dots\}$ A discrete probability density determining the law of the total number M of points of the process.
 - ✓ $H_m(\cdot)$ For each integer $m \geq 1$ this is a probability distribution on \mathcal{S}^m that determines the joint law of the positions of the points of the process, given that their total number is m .
- Since S is unordered, $H_m(\cdot)$ should be symmetric,

An alternative notation to identify the law of S , which has some advantages in simplifying combinatorial formulae, utilizes the nonprobability **Janossy measure**:

$$\mathbb{J}_m(A_1 \times \cdots \times A_m) = q_m \sum_{perm} H_m(A_{i_1} \times \cdots \times A_{i_m}) = m!q_m H_m(A_1 \times \cdots \times A_m).$$

for each $m \geq 0$.

Interpretation: if $\mathcal{X} = \mathbb{R}^d$ and $s_i \neq s_j$ for $i \neq j$, then

$$\mathbb{J}_m(ds_1, \dots, ds_m) = \mathbb{P}(\text{there are exactly } m \text{ points in the process, one in each of the distinct infinitesimal regions } (s_i, s_i + ds_i)).$$

- Janossy densities plays a fundamental role in the study of finite point processes and spatial point patterns, we refer to [see Daley and Vere-Jones (2003)] for more details.

$$P \sim \text{Norm-IFPP}(h, \{q_m\}, P_0), \text{ on } \Theta \subset \mathbb{R}^s$$

$$P \sim \text{Norm-IFPP}(h, \{q_m\}, P_0), \text{ on } \Theta \subset \mathbb{R}^s$$

Constructive definition: Normalization of a finite point process

$$P(\cdot) = \sum_{j \in \mathcal{J}} w_j \delta_{\tau_j}(\cdot) \stackrel{d}{=} \sum_{j \in \mathcal{J}} \frac{S_j}{T} \delta_{\tau_j}(\cdot), \quad (1)$$

where $\mathcal{J} = \{1, \dots, M\}$ and $0 < T = \sum_{j \in \mathcal{J}} S_j < \infty$.

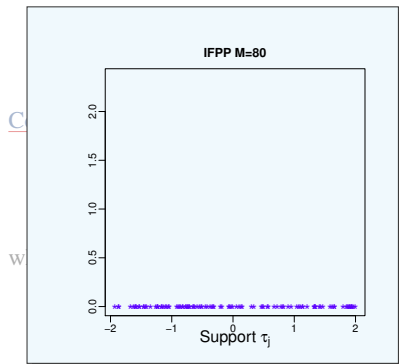
✓ $\{S_1, \dots, S_M\}$ is an *independent finite point process* with $q_0 = 0$ with Janossy density

$$\mathbb{J}_m(ds_1, \dots, ds_m) = m! q_m \prod_{j=1}^m h(s_j) ds_j. \quad m = 1, 2, \dots$$

where h is a density on \mathbb{R}^+ .

- ✓ the support $\{\tau_j\}$ is an iid sequence from P_0 ;
- ✓ $\{S_j\}$ and $\{\tau_j\}$ are independent.

Normalized independent finite point processes (Norm-IFPP)



$P(h, \{q_m\}, P_0)$, on $\Theta \subset \mathbb{R}^s$

a finite point process

$$\nu_j \delta_{\tau_j}(\cdot) \stackrel{d}{=} \sum_{j \in \mathcal{J}} \frac{S_j}{T} \delta_{\tau_j}(\cdot), \quad (1)$$

$$\forall j \in \mathcal{J} \quad S_j < \infty.$$

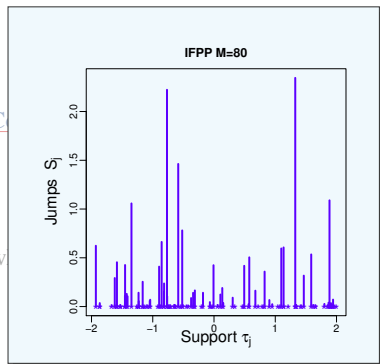
the point process with $q_0 = 0$ with Janossy density

$$j_m(a s_1, \dots, a s_m) = m! q_m \prod_{j=1}^m h(s_j). \quad m = 1, 2, \dots$$

where h is a density on \mathbb{R}^+ .

- ✓ the support $\{\tau_j\}$ is an iid sequence from P_0 ;
- ✓ $\{S_j\}$ and $\{\tau_j\}$ are independent.

Normalized independent finite point processes (Norm-IFPP)



$P(h, \{q_m\}, P_0)$, on $\Theta \subset \mathbb{R}^s$

a finite point process

$$\nu_j \delta_{\tau_j}(\cdot) \stackrel{d}{=} \sum_{j \in \mathcal{J}} \frac{S_j}{T} \delta_{\tau_j}(\cdot), \quad (1)$$

$$\sum_{j \in \mathcal{J}} S_j < \infty.$$

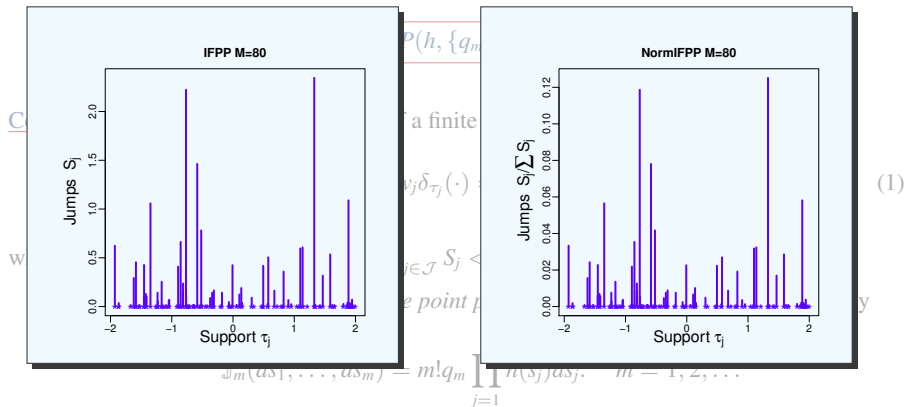
the point process with $q_0 = 0$ with Janossy density

$$j_m(as_1, \dots, as_m) = m! q_m \prod_{j=1}^m h(s_j) ds_j. \quad m = 1, 2, \dots$$

where h is a density on \mathbb{R}^+ .

- ✓ the support $\{\tau_j\}$ is an iid sequence from P_0 ;
- ✓ $\{S_j\}$ and $\{\tau_j\}$ are independent.

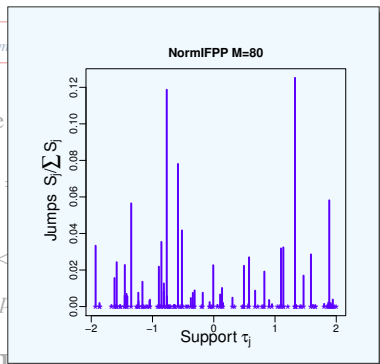
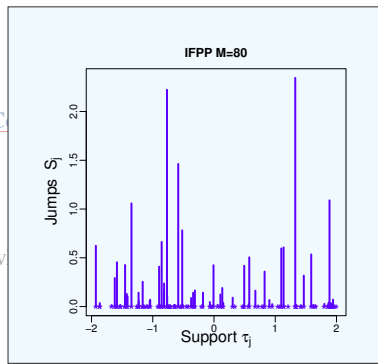
Normalized independent finite point processes (Norm-IFPP)



where h is a density on \mathbb{R}^+ .

- ✓ the support $\{\tau_j\}$ is an iid sequence from P_0 ;
- ✓ $\{S_j\}$ and $\{\tau_j\}$ are independent.

Normalized independent finite point processes (Norm-IFPP)



A simple example **Finite (Poisson) Dirichlet process**:

Number of jumps $\mathcal{P}_1(\Lambda)$ $m = 1, 2, \dots$, $q_m = \frac{e^{-\Lambda} \Lambda^{m-1}}{(m-1)!}$; Distribution $\text{Gamma}(\gamma, 1)$ $h(s) = \frac{1}{\gamma} s^{\gamma-1} e^{-s}$;

Given $M \left(\frac{S_1}{T}, \dots, \frac{S_M}{T} \right) \sim \text{Dirichlet}_M(\gamma, \dots, \gamma)$. **FDP**

The **variables** $\theta_1, \dots, \theta_n | P \stackrel{iid}{\sim} P$ where $P \sim \text{NormIFPP}$ induce a **random partition** ρ of data indexes $\{1, \dots, n\}$.

☛ Since P is a.s. discrete we observe ties with positive probability:

- ✓ $\theta_1^*, \dots, \theta_{K_n}^*$: unique values in $\theta_1, \dots, \theta_n$
- ✓ $\rho = \{C_1, \dots, C_{K_n}\}$: $i \in C_j \Leftrightarrow \theta_i = \theta_j^*, \#C_j = n_j$

The variables $\theta_1, \dots, \theta_n | P \stackrel{iid}{\sim} P$ where $P \sim \text{NormIFPP}$ induce a **random partition** ρ of data indexes $\{1, \dots, n\}$.

Since P is a.s. discrete we observe ties with positive probability:

- ✓ $\theta_1^*, \dots, \theta_{K_n}^*$: unique values in $\theta_1, \dots, \theta_n$
- ✓ $\rho = \{C_1, \dots, C_{K_n}\}$: $i \in C_j \Leftrightarrow \theta_i = \theta_j^*$, $\#C_j = n_j$

Prior of ρ : exchangeable partition probability function (Pitman 1996)

$$\mathbb{P}(\rho = \{C_1, \dots, C_{K_n}\}) = \text{eppf}(\#C_1, \dots, \#C_{K_n}) := \sum_{j_1, \dots, j_{K_n}} \mathbb{E} \prod_{i=1}^{K_n} w_{j_i}^{(\#C_i)}$$

Theorem 1 – *Eppf-characterization*

Let (n_1, \dots, n_k) be a vector of positive integers such that $\sum_{i=1}^k n_i = n$. Then, the eppf associated with a Norm-IFPP($h, \{q_n\}, P_0$) is

$$\text{eppf}(n_1, \dots, n_k) = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} \Psi(u, k) \prod_{i=1}^k \kappa(n_i, u) du$$

where

$$\Psi(u, k) := \left\{ \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} \psi(u)^m q_{m+k} \right\},$$

moreover, $\psi(u)$ is the Laplace transform of the density $h(s)$, i.e.

$$\psi(u) := \int_0^{\infty} e^{-us} h(s) ds, \quad \text{and} \quad \kappa(n_i, u) := \int_0^{\infty} u^{n_i} e^{-us} h(s) ds = (-1)^{n_i} \frac{d}{du^{n_i}} \psi(u).$$

Why it is important to have an expression of the eppf?

- Computation The eppf fully characterize the predictive structure of P , i.e. it provide us with a Chinese Restaurant representation of the clustering ρ .

Why it is important to have an expression of the eppf?

- Computation The eppf fully characterize the predictive structure of P , i.e. it provide us with a Chinese Restaurant representation of the clustering ρ .
- Interpretation It allows us to compute the prior distribution on the number of clusters, i.e for $k = 1, \dots, n$

$$\mathbb{P}(K_n = k) = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} \Psi(u, k) B_{n,k}(\kappa(\cdot, u))$$

where $B_{n,k}(\kappa(\cdot, u))$ is the *partial Bell polynomial*

Why it is important to have an expression of the eppf?

- Computation The eppf fully characterize the predictive structure of P , i.e. it provide us with a Chinese Restaurant representation of the clustering ρ .
- Interpretation It allows us to compute the prior distribution on the number of clusters, i.e for $k = 1, \dots, n$

$$\mathbb{P}(K_n = k) = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} \Psi(u, k) B_{n,k}(\kappa(\cdot, u))$$

where $B_{n,k}(\kappa(\cdot, u))$ is the *partial Bell polynomial*

- Difficulties The analytical expression of the eppf involves:
 - ① an integral respect to u ;
 - ② an infinite sum $\Psi(u, k)$;
 - ③ the Laplace transform of $h(s)$.

Idea To avoid the analytical computation of the integral respect to u we augment the state space of the process by a latent variable U_n – *disintegration trick*.

The joint law of the partition ρ and U_n is

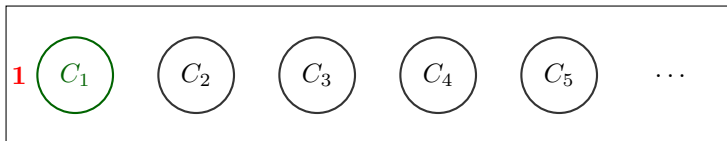
$$eppf(n_1, \dots, n_k, du) = \frac{u^{n-1}}{\Gamma(n)} \Psi(u, k) \prod_{i=1}^k \kappa(n_i, u) du$$

while the marginal law of U_n is

$$f_{U_n}(u; n) = (-1)^n \frac{u^{n-1}}{\Gamma(n)} \frac{d}{du^n} \mathbb{E}(\psi(u)^M)$$

To draw a partition ρ from a Norm-IPPF

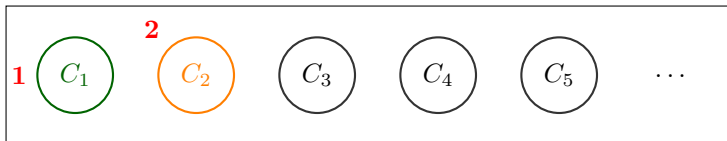
- ✓ The first customer sits at table 1, and $U_1 = u$ is drawn;



To draw a partition ρ from a Norm-IPPF

- ✓ The first customer sits at table 1, and $U_1 = u$ is drawn;
- ✓ Given that k tables are occupied by n customer, and $U_n = u$, customer $n + 1$ sits:
 - A new table $k + 1$ with probability proportional to

$$\frac{\text{epf}(n_1, \dots, n_k, 1; u)}{\text{epf}(n_1, \dots, n_k; u)} = \frac{\Psi(u, k+1)}{\Psi(u, k)} \kappa(1, u)$$



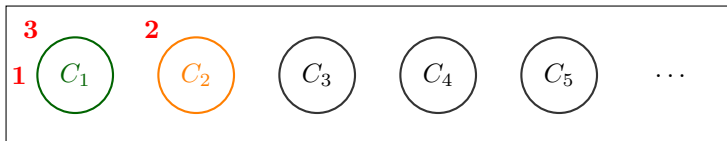
To draw a partition ρ from a Norm-IPPF

- ✓ The first customer sits at table 1, and $U_1 = u$ is drawn;
- ✓ Given that k tables are occupied by n customer, and $U_n = u$, customer $n + 1$ sits:
 - A *new* table $k + 1$ with probability proportional to

$$\frac{\text{epf}(n_1, \dots, n_k, 1; u)}{\text{epf}(n_1, \dots, n_k; u)} = \frac{\Psi(u, k+1)}{\Psi(u, k)} \kappa(1, u)$$

- at an *occupied* table $j = 1, \dots, k$ with probability proportional to

$$\frac{\text{epf}(n_1, \dots, n_j+1, \dots, n_k, 1; u)}{\text{epf}(n_1, \dots, n_j, \dots, n_k; u)} = \frac{\kappa(n_j+1, u)}{\kappa(n_j, u)}, \quad j = 1, \dots, k$$



To draw a partition ρ from a Norm-IPPF

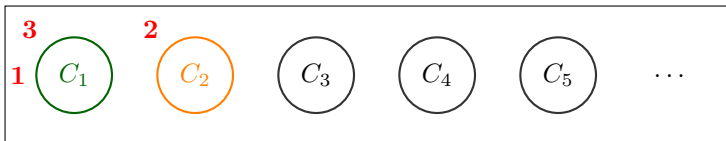
- ✓ The first customer sits at table 1, and $U_1 = u$ is drawn;
- ✓ Given that k tables are occupied by n customer, and $U_n = u$, customer $n + 1$ sits:
 - A *new* table $k + 1$ with probability proportional to

$$\frac{eppf(n_1, \dots, n_k, 1; u)}{eppf(n_1, \dots, n_k; u)} = \frac{\Psi(u, k+1)}{\Psi(u, k)} \kappa(1, u)$$

- at an *occupied* table $j = 1, \dots, k$ with probability proportional to

$$\frac{eppf(n_1, \dots, n_j+1, \dots, n_k, 1; u)}{eppf(n_1, \dots, n_j, \dots, n_k; u)} = \frac{\kappa(n_j+1, u)}{\kappa(n_j, u)}, \quad j = 1, \dots, k$$

- we draw $U_n \sim f_{U_n}(u | n_1, \dots, n_k) \propto eppf(n_1, \dots, n_k; u)$



I just recall that to compute the eppf we need to evaluate the infinite sum

$$\Psi(u, k) := \left\{ \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} \psi(u)^m q_{m+k} \right\}$$

✓ We have a closed form expression for three cases (*conjugacy*):

- If M is assumed Shifted Poisson on $\{1, 2, \dots\}$, then

$$\Psi(u, k) = \Lambda^{k-1} (\Lambda \psi(u) + k) \exp\{\Lambda(\psi(u) - 1)\}$$

- If M is assumed Negative Binomial with parameters $0 \leq p \leq 1$ and $r > 0$

$$\Psi(u, k) = \frac{\Gamma(r+k-1)}{\Gamma(r)} p^{k-1} (1-p)^r \frac{p\psi(u)(r-1)+k}{(1-p\psi(u))^{k+r}}$$

- If M is assumed fixed, i.e. $M = \tilde{M} \geq 1$ with probability 1,

$$\Psi(u, k) = \begin{cases} \frac{\tilde{M}!}{(\tilde{M}-k)!} \psi(u)^{\tilde{M}-k} & \text{if } k \leq \tilde{M} \\ 0 & \text{if } k > \tilde{M} \end{cases}$$

Let S_m the unnormalized weights, conditionally to M , $S_m \stackrel{iid}{\sim} h(s)$

- $S_j \sim \text{Gamma}(\gamma, 1)$ – Finite Dirichlet Process (FDP):

$$\psi(u) = \frac{1}{(u+1)^\gamma}, \quad \kappa(u, n_j) = \frac{1}{(u+1)^{n_j+\gamma}} \frac{\Gamma(\gamma+n_j)}{\Gamma(\gamma)}$$

- $S_j \sim \text{Unif}(0, 1)$:

$$\psi(u) = \frac{1-e^u}{u}, \quad \text{and} \quad \kappa(n_j, u) = \frac{\gamma(n_j+1, u)}{u^{n_j+1}}$$

Let S_m the unnormalized weights, conditionally to M , $S_m \stackrel{iid}{\sim} h(s)$

- $S_j \sim \text{Gamma}(\gamma, 1)$ – Finite Dirichlet Process (FDP):

$$\psi(u) = \frac{1}{(u+1)^\gamma}, \quad \kappa(u, n_j) = \frac{1}{(u+1)^{n_j+\gamma}} \frac{\Gamma(\gamma+n_j)}{\Gamma(\gamma)}$$

- $S_j \sim \text{Unif}(0, 1)$:

$$\psi(u) = \frac{1-e^u}{u}, \quad \text{and} \quad \kappa(n_j, u) = \frac{\gamma(n_j+1, u)}{u^{n_j+1}}$$

- Levy Processes approach – Fix $\psi(u) = e^{-\int_0^\infty (e^{ux} - 1)\omega(z)dx}$, where $\omega(z)$ is called *Levy intensity*, and compute $h(s)$ such that

$$h(s) = \int_0^s \omega(z)h(s-z) \frac{z}{s} dz$$

- ⇒ This latter construction is the finite dimensional version of a Normalized Completely Random Measure (Lijoi et al. 2007)

Let P be a finite Dirichlet process, i.e. a Norm-IFPP such that

$$M \sim q_m, \text{ and } S_j \stackrel{iid}{\sim} \text{gamma}(\gamma, 1).$$

• We will use the notation $P \sim FDP(\gamma, \Lambda, P_0)$.

Let P be a finite Dirichlet process, i.e. a Norm-IFPP such that

$$M \sim q_m, \text{ and } S_j \stackrel{iid}{\sim} \text{gamma}(\gamma, 1).$$

• We will use the notation $P \sim FDP(\gamma, \Lambda, P_0)$.

✓ The eppf of a $FDP(\gamma, \Lambda, P_0)$ is given by [see also Miller Harrison (2016)]

$$\text{eppf}(n_1, \dots, n_k) = V(n, k) \prod_{j=1}^k \frac{\Gamma(\gamma + n_j)}{\Gamma(\gamma)},$$

where $V(n, k) = \int_0^\infty \tilde{f}(u) du.$, and \tilde{f} is a function that depends on the prior on q_M .

Let P be a finite Dirichlet process, i.e. a Norm-IFPP such that

$$M \sim q_m, \text{ and } S_j \stackrel{iid}{\sim} \text{gamma}(\gamma, 1).$$

• We will use the notation $P \sim FDP(\gamma, \Lambda, P_0)$.

✓ The eppf of a $FDP(\gamma, \Lambda, P_0)$ is given by [see also Miller Harrison (2016)]

$$\text{eppf}(n_1, \dots, n_k) = V(n, k) \prod_{j=1}^k \frac{\Gamma(\gamma + n_j)}{\Gamma(\gamma)},$$

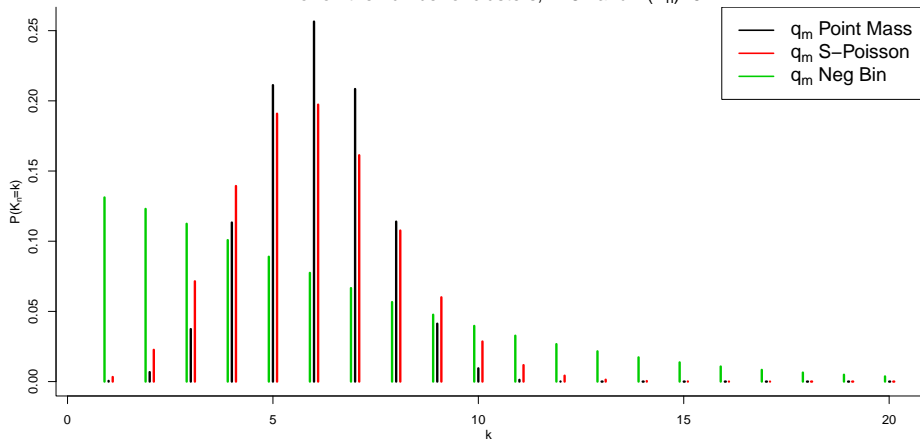
where $V(n, k) = \int_0^{\infty} \tilde{f}(u) du.$, and \tilde{f} is a function that depends on the prior on q_M .

• We will consider M as a Shifted Poisson or a Negative Binomial.

✓ Let \mathcal{C} denote the generalized Stirling numbers of second kind, then

$$P(K_n = k) = V(n, k) \mathcal{C}(n, k, \gamma)$$

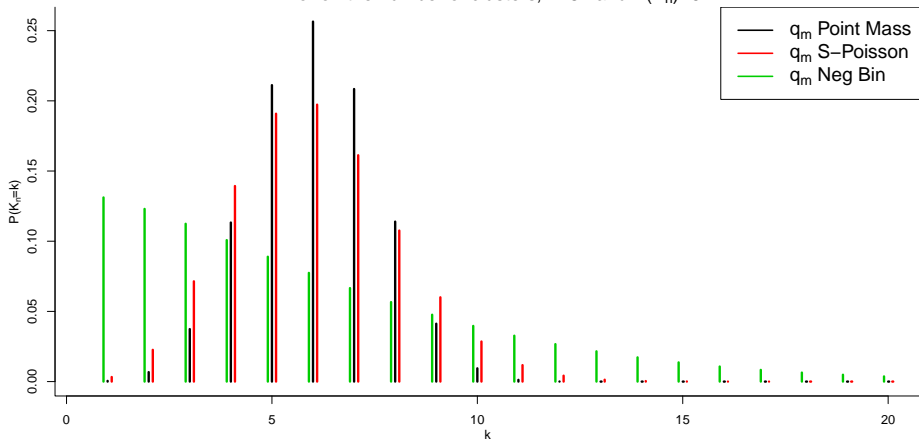
Prior on the number of clusters, $n=82$ and $E(K_n)=6$



✓ Note that, from the de Finetti Theorem

$$K_n \rightarrow M \text{ a.s. for } n \rightarrow \infty$$

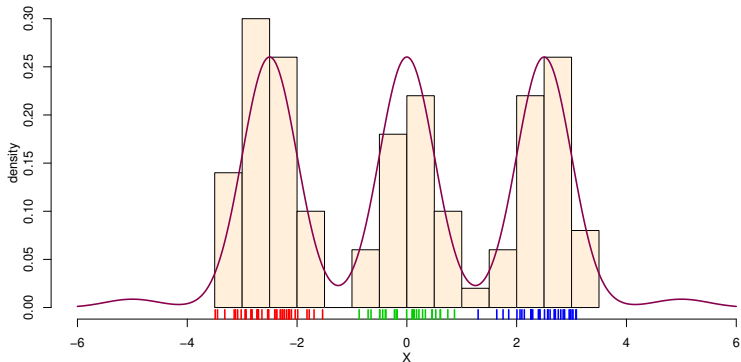
Prior on the number of clusters, $n=82$ and $E(K_n)=6$



Again the allocated and non-allocated components

In my illustrative example:

- ✓ The allocated components are $K_n = M^{(a)} = 3$
- ✓ The non-allocated components (empty) are $M^{(na)} := M - M^{(a)} = 2$.



Theorem 2 – Posterior law

Let $(\theta_1, \dots, \theta_n)$ be a sample from $P \sim \text{Norm} - \text{IFPP}(h, \{q_n\}, P_0)$, then there exist an auxiliary random variable U such that the conditional law of P , given θ^* and $U_n = u$ coincides with the normalization of the following:

$$\sum_{j \in \mathcal{J}^{(na)}} S_j^{(na)} \delta_{\tau_j}(\cdot) + \sum_{j \in \mathcal{J}^{(a)}} S_j^{(a)} \delta_{\theta_j^*}(\cdot) \quad \tau_j \stackrel{iid}{\sim} P_0$$

- ① **Non-allocated** jumps: the process $\{S^{(na)}\}$ is in IFPP with Janossy density given by

$$\mathbb{J}_m(ds_1, \dots, ds_m) = m! p_m^* \prod_{j=1}^m h^*(s_j) ds_j$$

$$h_u^*(s) \propto e^{-us} h(s) \quad \text{and} \quad q_m^* \propto \frac{(m+k)!}{m!} \psi(u)^m q_{m+k}, \quad m = 0, 1, 2, \dots$$

- ② **Allocated** jumps: for each $j \in \mathcal{J}^{(a)} = \{1, \dots, K_n\}$ the distribution of $S_j^{(a)}$ is proportional to $s^{n_j} e^{-us} h(s)$.

- ③ **Latent variable:** $[U_n | \mathcal{S}^{(a)}, \mathcal{S}^{(na)}] \sim \text{Gamma}(n, \sum_j S_j)$

- ✓ We let $\{f(\cdot, \theta), \theta \in \Theta\}$ be the family of Gaussian density.
- ✓ Then, the parameter $\theta = (\mu, \sigma^2)$ and P_0 is a *conjugate* prior for θ .

Mixture model

$$X_1, \dots, X_n | \theta_1, \dots, \theta_n \stackrel{\text{ind}}{\sim} f(x_i | \theta_i)$$

$$\theta_1, \dots, \theta_n | P \stackrel{\text{iid}}{\sim} P$$

$$P \sim FDP(\gamma, \Lambda, P_0)$$

$$(\gamma, \Lambda) \sim \text{gamma}(a_1, b_1) \times \text{gamma}(a_2, b_2)$$

- When Λ and γ are fixed, we choose them such that $\mathbb{E}(K_n)$ express our prior believes on the number of groups.
- **Result:** if we let $\gamma = \kappa/\Lambda$ then for $\Lambda \rightarrow \infty$ then P converges in law to the Dirichlet process $DP(\kappa, P_0)$.

We **augment** the state space introducing the r.v. U_n

Parameter: $U_n, \theta, P, \Lambda, \gamma$

We **augment** the state space introducing the r.v. U_n

Parameter: $U_n, \theta, P, \Lambda, \gamma$

For \mathbf{g} in $1, \dots, G$:

1. sample $U_n | rest$ from a $\text{Gamma}(n, \sum_j S_j)$

2. Sample $\theta | rest$, for each $i = 1, \dots, n$ from the discrete distribution

$$\mathbb{P}(\theta_i = \tau_j) \propto S_j f(X_i | \tau_j), \quad j \in \mathcal{J} = \{1, \dots, M\}$$

3. Update the r.p.m. $P | rest$

3a. Update the r.p.m. P , given $\gamma, \Lambda, U, \theta$ we apply Theorem 2

3a.1 Sample $M^{(na)}$ from q_m^* that is the p.m.f.

$$\frac{(u+1)^\gamma k}{(u+1)^\gamma k + \Lambda} \mathcal{P}_1(\Lambda/(u+1)^\gamma) + \frac{\Lambda}{(u+1)^\gamma k + \Lambda} \mathcal{P}_0(\Lambda/(u+1)^\gamma),$$

where \mathcal{P}_i is the Shifted Poisson on $\{i, i+1, \dots\}$.

3a.2' Non-allocated jumps: sample

$$S_j^{(na)} \stackrel{iid}{\sim} \text{Gamma}(\gamma, u+1)$$

3a.2'' Allocated jumps: sample

$$S_j^{(a)} \stackrel{iid}{\sim} \text{Gamma}(n_i - \gamma, u+1)$$

3a.3' Non-allocated support points:

$$\tau_j \stackrel{iid}{\sim} P_0(d\tau_j)$$

3a.3'' Allocated support points: iid as

$$\tau_j = \theta_j^* \sim \prod_{i \in C_j} f(X_i | \theta_j^*) P_0(d\theta_j^*)$$

3b. Update γ , Λ , given U and θ

3b.1 Sample Λ from this mixture of gamma densities:

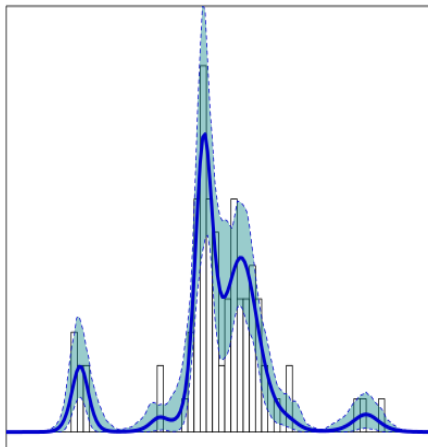
$$\frac{\psi(u)}{1+b_2} \text{Gamma}(k+a_2+1, 1-\psi(u)+b_2) \frac{1-\psi(u)+b_2}{1+b_2} \text{Gamma}(k+a_2, 1-\psi(u)+b_2)$$

where $\psi(u) = \frac{1}{(u+1)^\gamma}$ is the Laplace transform of a $\text{gamma}(\gamma, 1)$;

3b.2 Sample γ from the law

$$\mathcal{L}(\gamma) \propto (\Lambda\psi(u) + k) e^{\Lambda\psi(u)} \frac{1}{\psi(u)^k} \prod_{j=1}^k \frac{\Gamma(\gamma + n_j)}{\Gamma(\gamma)}$$

and we have to resort to an Adaptive Metropolis step to sample from this non standard full conditional.

Figure: $\Lambda = 10$, $\gamma = 0.21$ **Dataset:**

$n = 82$ galaxy velocities [$10^6 m/s$]

$$k(\cdot; \theta) = \mathcal{N}(\cdot; \mu, \sigma^2)$$

$$P_0(d\mu, d\sigma^2) = \mathcal{N}(d\mu, \sigma^2/k_0) \\ \times IG(d\sigma^2 | a, b)$$

$$(m_0, k_0, a, b) = (20.8, 0.01, 2, 1)$$

+ some robustness analysis

- ✓ We fix Λ and γ such that $\mathbb{E}(K_n) = 6$.
- ✓ Reversible Jump via `mixAK` R-package ([Komárek, 2009]; C++ linked to R). Our Gibbs is implemented in C++ code.
- ✓ 5000 burn-in, 10 thinning and final sample size of 5000.
- ✓ Integrated autocorrelation time [Kalli, Griffin and Walker, 2011]

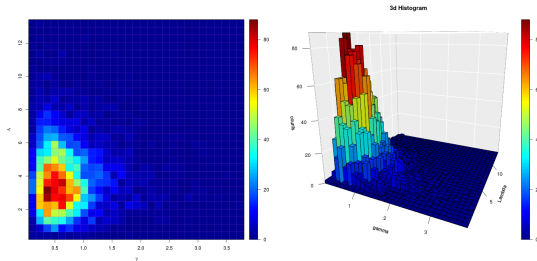
$$\hat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \hat{\rho}_l,$$

• A small value of τ implies good mixing and hence an efficient method.

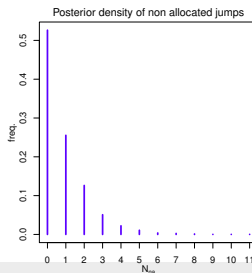
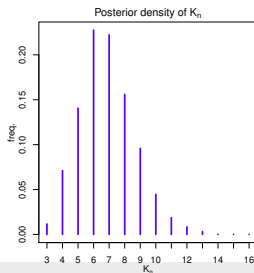
(Λ, γ)	Blocked Gibbs			Reversible Jump		
	time	$\mathbb{E}(M data)$	$\hat{\tau}$	time	$\mathbb{E}(M data)$	$\hat{\tau}$
(1000,0.0013)	15.13 min.	1003.47	1.53	22.69 min.	669.33	864.44
(100, 0.0136)	1.51 min.	103.19	1.51	2.12 min.	98.16	138.40
(10, 0.21)	12.50 sec.	13.18	1.33	12.03 sec.	10.31	3.45
(5,5)	9.60 sec.	9.34	1.26	9.25 sec.	7.10	6.29

✓ $\Lambda \sim \text{Gamma}(1, 0.01)$ and $\gamma \sim (2, 1)$.

✓ Performances: time 8.26 sec and $\tau = 3.89$, $\mathbb{E}(M^{(na)}|data) = 0.86$.



(Λ, γ)	LPLM
(1000, 0.0013)	-9.01
(100, 0.0136)	-8.97
(10, 0.21)	-8.96
(5, 5)	-8.26
random	-8.86



- ✓ **Finite mixture model**: We have proposed the new class of *finite independent normalized point processes* (Norm-IFPP) as the mixing measure.
- ✓ We have given an analytical expression of the *exchangeable partition probability function*, i.e. we characterized the law of the random partition induced by a Norm-IFPP on the data.
- ✓ We have characterized the **posterior distribution** of Norm-IFPP.
- ✓ We have designed a “conjugate” **blocked Gibbs** sampler for the Finite Dirichlet Mixture mixture model.
- ✓ Our Gibbs sampler outperforms the reversible jump in term of integrated autocorrelation time.

Thank you!!!

- ✓ **Finite mixture model**: We have proposed the new class of *finite independent normalized point processes* (Norm-IFPP) as the mixing measure.
- ✓ We have given an analytical expression of the *exchangeable partition probability function*, i.e. we characterized the law of the random partition induced by a Norm-IFPP on the data.
- ✓ We have characterized the **posterior distribution** of Norm-IFPP.
- ✓ We have designed a “conjugate” **blocked Gibbs** sampler for the Finite Dirichlet Mixture mixture model.
- ✓ Our Gibbs sampler outperforms the reversible jump in term of integrated autocorrelation time.

⇒ Argiento, De Iorio (2019) “Is infinity that far? A Bayesian nonparametric perspective of finite mixture models”, *arXiv:1904.09733*

- ⇒ Argiento, De Iorio (2019) “Is infinity that far? A Bayesian nonparametric perspective of finite mixture models”, *arXiv:1904.09733*
- ⇒ Fruhwirth-Schnatter, S., Celeux, G. and Robert, C. P. (2019) *Handbook of mixture analysis*.
- ⇒ Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019) From here to infinity: sparse finite versus dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification* **13**(1), 33–64..
- ⇒ Lijoi, A., Mena, R. H. and Prünster, I. (2007) Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 715–740.
- ⇒ Miller, J. W. and Harrison, M. T. (2018) Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, **113**, 340–356.
- ⇒ Nobile, A. (2004) On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics*, **32**, 2044–2073.
- ⇒ Richardson, S. and Green, P. J. (1997) On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 731–792.
- ⇒ Rousseau, J. and Mengersen, K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 689–710.