Clustering grouped data via hierarchical mixture models

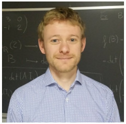
a finite dimensional perspective

Raffaele Argiento

Università degli Studi di Bergamo

Montréal, June 26th – July 1st 2022

Alessandro Colombi

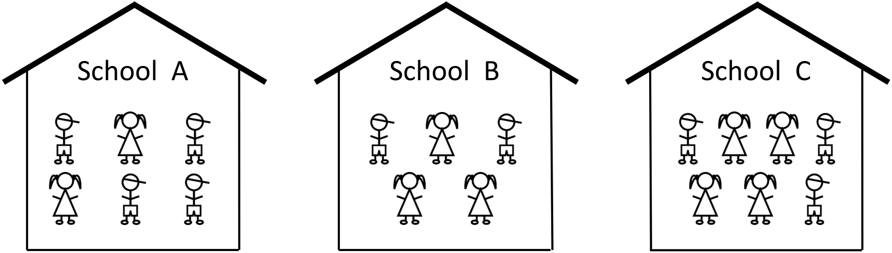


Federico Camerlenghi

Lucia Paci

Grouped data

- ✓ Observations are organized in different groups
- ✓ Partially exchangeable data
- ✓ Sharing of information across groups to learn distinctive features of the units



Dealing with grouped data

✓ Main approaches to model **grouped data** in Bayesian nonparametrics:

1. Hierarchical processes (Teh et al., 2006; Camerlenghi et al., 2019; Argiento et al., 2020)
2. Nested process (Rodríguez et al., 2008) and the recent extensions have been proposed by Denti et al. (2021) and Beraha et al. (2021)
3. Clayton-Levy copulas – Betatrice Franzolin talk

Our proposal

☛ A **Bayesian hierarchical mixture model** where the group-specific mixing distribution belongs to the class of **finite Dirichlet Processes** (Argiento and De Iorio, 2022), i.e. *Gibbs priors with negative parameter* (De Blasi et al., 2013)

☛ We assign the joint law of the mixing distributions assuming a **shared support** (D'Angelo et al., 2022) and allowing clustering within and between groups

Mixture model

Let y_{ji} be the observed variable for group $j = 1, \ldots, d$, and individual $i = 1, \ldots, n_j$

We assume that the data in each group j come from a finite mixture with **random number, M, of components** (i.e. mixture of finite mixture) that is

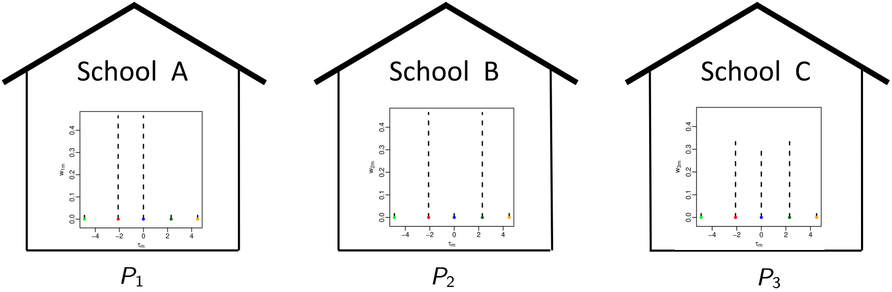
$$y_{j1}, \ldots, y_{jn_j} \mid w_{jm}, \tau_m, M \sim \sum_{l=1}^{M} w_{jm} f(y_{ji} \mid \tau_m),$$

where

- $f(y_{ji} \mid \tau_m)$ are the kernels of the mixture
- $\{\tau_m, m = 1, \ldots M\} \subset \Theta$ are **kernel parameters shared across groups**.
- $\{w_{jm}, j = 1, \ldots, M\}$ are the group-specific **mixing weights**
- M is the shared number of components

The vector of mixing distributions

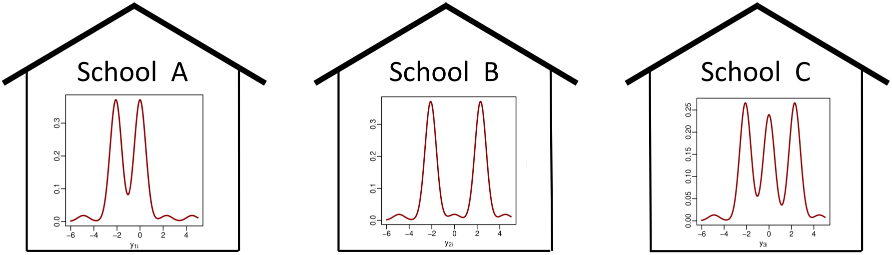
- ✓ Let $P_j := \sum_{m=1}^{M} w_{jm}\delta_{\tau_m}$, we consider hierarchical mixture model where P_j is a group specific **mixing distribution**
- ✓ $(P_1, \ldots, P_d)$ is a **vector** of species sampling processes (see Antonio Lijoi keynote lecture)



The vector of mixing distributions

- ✓ Let $P_j := \sum_{m=1}^{M} w_{jm}\delta_{\tau_m}$, we consider hierarchical mixture model where P_j is a group specific **mixing distribution**
- ✓ $(P_1, \ldots, P_d)$ is a **vector** of species sampling processes (see Antonio Lijoi keynote lecture)
- ✓ The sampling model in group $j = 1, \ldots, d$ is

$$y_j | P_j \sim \int_{\Theta} f(y_j|\theta) P_j(\mathrm{d}\theta)$$



$y_1|P_1$ $y_2|P_2$ $y_3|P_3$

Prior via normalization

We assign a prior distribution on the mixing weights by **normalization**

$$\boxed{S_{j1}, \ldots, S_{jM}, \mid M, \gamma_j \overset{iid}{\sim} \text{Gamma}(\gamma_j, 1) \quad T_j = \textstyle\sum_{m=1}^{M} S_{jm}}$$

so that

$$w_{j1} = \frac{S_{j1}}{T_j}, \ldots, w_{jM} = \frac{S_{jM}}{T_j} \sim \text{Dirichlet}_M(\gamma_j, \ldots, \gamma_j)$$

Prior via normalization

We assign a prior distribution on the mixing weights by **normalization**

$$\boxed{S_{j1}, \dots, S_{jM}, \mid M, \gamma_j \stackrel{iid}{\sim} \text{Gamma}(\gamma_j, 1) \quad T_j = \sum_{m=1}^{M} S_{jm}}$$

so that

$$w_{j1} = \frac{S_{j1}}{T_j}, \dots, w_{jM} = \frac{S_{jM}}{T_j} \sim \text{Dirichlet}_M(\gamma_j, \dots, \gamma_j)$$

For one group, i.e. $d = 1$, Argiento and De Iorio (2022)

- ✓ obtain an analytical expression of the clustering eppf
- ✓ design an a priori/a posterior Chinese Restaurant type process yielding to an "easy" marginal posterior sampler
- ✓ characterize the posterior distribution of P_1 and design a conditional sampler

The latent θ_{ji}'s

Summary of the model

$$
\begin{aligned}
&y_{ji} \mid \theta_{ji} \overset{\text{ind}}{\sim} f(y_{ji} \mid \theta_{ji}) \qquad i = 1, \ldots, n_j \\
&\theta_{j1}, \ldots, \theta_{jn_j} \mid P_j \overset{\text{iid}}{\sim} P_j \quad j = 1, \ldots, d \\
&P_1, \ldots, P_d \mid \gamma_j, \Lambda \sim \textit{Vector-Fin-Dirichlet}(\gamma_j, \Lambda, p_0)
\end{aligned}
$$

- ✓ $f(y \mid \theta)$: is the density of a $N(\mu, \sigma^2)$
- ✓ M is assumed to be a 1−shifted Poisson distribution with parameter Λ
- ✓ S_j: are independent Gamma(γ_j, 1)
- ✓ $p_0(\tau)$: is the conjugate Norm-Inv-Gamma(μ_0, κ_0, ν_0,σ_0^2)
- ✓ Additional level of hierarchy can be added by assuming a prior on Λ and the γ_j's

Group specific clustering

- ✓ For each group j consider the vector latent parameters:

$$\theta_{j1}, \ldots, \theta_{jn_j} | P_j \stackrel{iid}{\sim} P_j$$

- ✓ As customary in mixture model framework we denote by

$$\mathcal{T}_j = \{\theta^\star_{j1}, \ldots \theta^\star_{jK_j}\}$$

the set of the K_j unique values among the θ's

Group specific clustering

✓ For each group j consider the vector latent parameters:

$$\theta_{j1}, \ldots, \theta_{jn_j} | P_j \overset{iid}{\sim} P_j$$

✓ As customary in mixture model framework we denote by

$$\mathcal{T}_j = \{\theta^\star_{j1}, \ldots \theta^\star_{jK_j}\}$$

the set of the K_j unique values among the θ's

✓ Let's now

$$\mathcal{T} = \cup_{j=1}^{d} \mathcal{T}_j$$

Group specific clustering

✓ For each group j consider the vector latent parameters:

$$\theta_{j1}, \ldots, \theta_{jn_j} | P_j \overset{iid}{\sim} P_j$$

✓ As customary in mixture model framework we denote by

$$\mathcal{T}_j = \{\theta^\star_{j1}, \ldots \theta^\star_{jK_j}\}$$

the set of the K_j unique values among the θ's

✓ Let's now

$$\mathcal{T} = \cup_{j=1}^{d} \mathcal{T}_j = \{\theta_1^{\star\star}, \ldots, \theta_K^{\star\star}\}$$

the set of unique values among the $\theta^\star$'s

✓ we refer to K as number of global **clusters**

Clustering within and between groups

✓ We define a **local clustering** for group j by letting

$$\rho_j = \{A_{j1}, \dots, A_{jK}\}$$

where for each $i = 1, \dots, n_j$

$$(j, i) \in A_{jk} \quad \text{iff} \quad \theta_{ji} = \theta_k^{\star\star} \quad k = 1, \dots, K$$

✓ this is a pseudo-clustering, indeed A_{jk} can be an empty set and then

$$\# A_{jk} := n_{jk} \geq 0$$

Clustering within and between groups

- ✓ We define a **local clustering** for group j by letting

$$\rho_j = \{A_{j1}, \ldots, A_{jK}\}$$

where for each $i = 1, \ldots, n_j$

$$(j,i) \in A_{jk} \quad \text{iff} \quad \theta_{ji} = \theta_k^{\star\star} \quad k = 1, \ldots, K$$

- ✓ this is a pseudo-clustering, indeed A_{jk} can be an empty set and then

$$\#A_{jk} := n_{jk} \geq 0$$

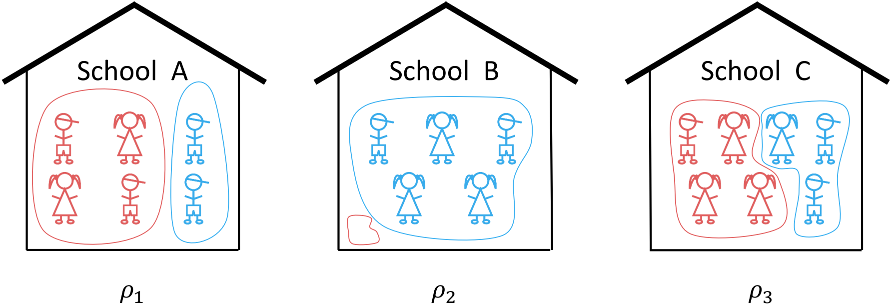
- ✓ The **global clustering** is given by merging the local clusterings

$$\rho = \{A_1, \ldots, A_K\}$$

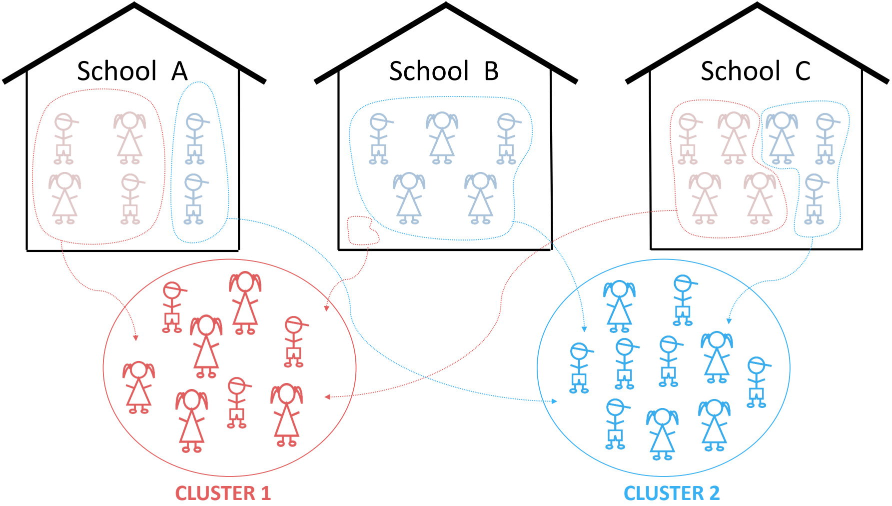
where $A_k = \cup_{j=1}^{d} A_{jk}$ and clearly

$$\#A_k = n_k = \sum_j n_{jk} > 0$$

Clustering within and between groups



Clustering within and between groups



Characterization of the local clustering

Theorem - *partial eppf*

The **joint distribution** of the local clusterings $\rho_1, \ldots, \rho_d$ and K, induced by our hierarchical mixture model on the data indices, is characterized via the following *partial* exchangeable partition probability function

$$\boxed{\pi\left(\rho_1, \ldots, \rho_d, K\right) = V(n_1, \ldots, n_d, K) \prod_{j=1}^{d} \prod_{k=1}^{M^{(na)}} \frac{\Gamma(\gamma_j + n_{kj})}{\Gamma(\gamma_j)}}$$

where

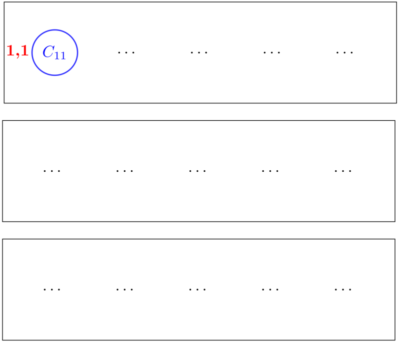
$$V(n_1, \ldots, n_d, K) = \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{d} \frac{1}{\Gamma(n_j)} \frac{u_j^{n_j - 1}}{(u_j+1)^{nj+\gamma_j K}} \left[\Lambda \prod_{j=1}^{d} \psi_{\gamma_j}(u_j) + K\right] \Lambda^{K-1} \exp\left[-\Lambda\left(\prod_{j=1}^{d} \psi_{\gamma_j}(u_j) - 1\right)\right] du_1 \ldots du_d$$

and

$\psi_{\gamma_j}(u_j) = \frac{1}{(u_j+1)^{\gamma_j}}$ is the Laplace transform of a gamma distribution

A Restaurant Franchise process

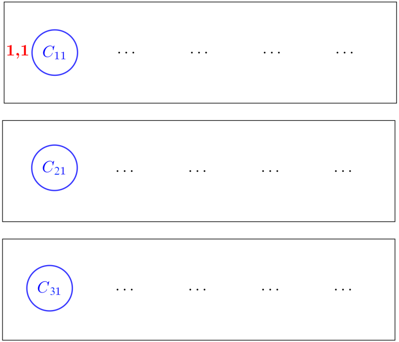
- ✓ A restaurant franchise with a *shared menu across the restaurants*
- ✓ The first customer of the first restaurant seats at table one and choose a dish (colour).



A Restaurant Franchise process

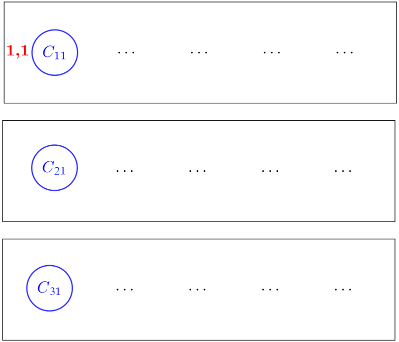
✓ A restaurant franchise with a *shared menu across the restaurants*

✓ The first customer of the first restaurant seats at table one and choose a dish (colour). Each restaurant lays a table serving the same dish
$M^{(na)} = 1$



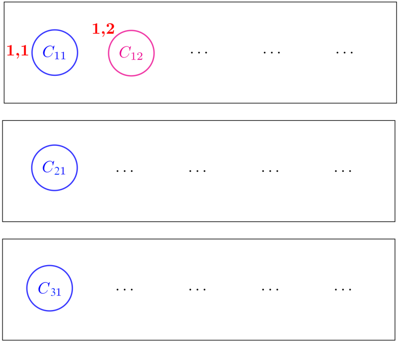
A Restaurant Franchise process

- ✓ A restaurant franchise with a *shared menu across the restaurants*
- ✓ The first customer of the first restaurant seats at table one and choose a dish (colour). Each restaurant lays a table serving the same dish $M^{(na)} = 1$
- ✓ then the $i-$th customer of the $j-$th restaurant



A Restaurant Franchise process

- ✓ A restaurant franchise with a *shared menu across the restaurants*
- ✓ The first customer of the first restaurant seats at table one and choose a dish (colour). Each restaurant lays a table serving the same dish $M^{(na)} = 1$
- ✓ then the $i-$th customer of the $j-$th restaurant



1. Pr(Sits to a table serving a new dish)

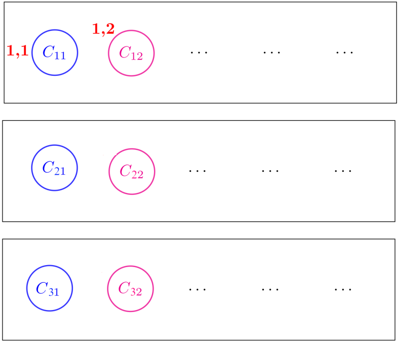
$$\boxed{= \frac{V(n_1,\dots,n_j+1,\dots,n_d,K+1)}{V(n_1,\dots,n_j,\dots,n_d,K)}\Lambda\gamma_j}$$

A Restaurant Franchise process

✓ A restaurant franchise with a *shared menu across the restaurants*

✓ The first customer of the first restaurant seats at table one and choose a dish (colour). Each restaurant lays a table serving the same dish $M^{(na)} = 1$

✓ then the $i-$th customer of the $j-$th restaurant



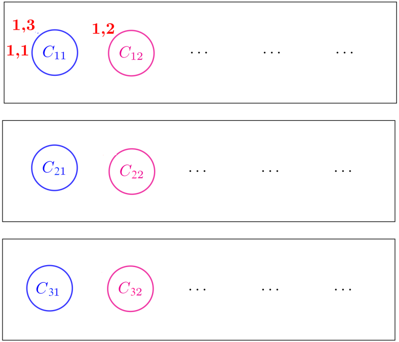
1. Pr(Sits to a table serving a new dish)

$$\boxed{= \frac{V(n_1,\dots,n_j+1,\dots,n_d,K+1)}{V(n_1,\dots,n_j,\dots,n_d,K)}\Lambda\gamma_j}$$

In this case a new table serving the new dish is laid in every restaurant $M^{(na)} = M^{(na)} + 1$

A Restaurant Franchise process

- ✓ A restaurant franchise with a *shared menu across the restaurants*
- ✓ The first customer of the first restaurant seats at table one and choose a dish (colour). Each restaurant lays a table serving the same dish $M^{(na)} = 1$
- ✓ then the $i-$th customer of the $j-$th restaurant



1. Pr(Sits to a table serving a new dish)

$$\boxed{= \frac{V(n_1,\dots,n_j+1,\dots,n_d,K+1)}{V(n_1,\dots,n_j,\dots,n_d,K)} \Lambda \gamma_j}$$

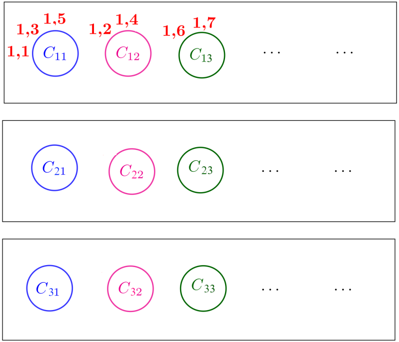
   In this case a new table serving the new dish is laid in every restaurant $M^{(na)} = M^{(na)} + 1$

2. Pr(Sits to an existing table)

$$\boxed{\frac{V(n_1,\dots,n_j+1,\dots,n_d,K)}{V(n_1,\dots,n_j,\dots,n_d,K)} (n_j + \gamma_j)}$$

A Restaurant Franchise process

- ✓ A restaurant franchise with a *shared menu across the restaurants*
- ✓ The first customer of the first restaurant seats at table one and choose a dish (colour). Each restaurant lays a table serving the same dish $M^{(na)} = 1$
- ✓ then the $i-$th customer of the $j-$th restaurant



1. Pr(Sits to a table serving a new dish)

$$= \frac{V(n_1,\ldots,n_j+1,\ldots,n_d,K+1)}{V(n_1,\ldots,n_j,\ldots,n_d,K)}\Lambda\gamma_j$$

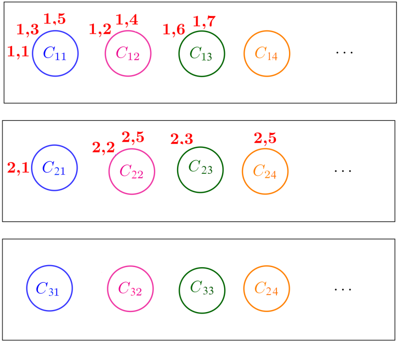
   In this case a new table serving the new dish is laid in every restaurant $M^{(na)} = M^{(na)} + 1$

2. Pr(Sits to an existing table)

$$\frac{V(n_1,\ldots,n_j+1,\ldots,n_d,K)}{V(n_1,\ldots,n_j,\ldots,n_d,K)}(n_j + \gamma_j)$$

A Restaurant Franchise process

- ✓ A restaurant franchise with a *shared menu across the restaurants*
- ✓ The first customer of the first restaurant seats at table one and choose a dish (colour). Each restaurant lays a table serving the same dish $M^{(na)} = 1$
- ✓ then the $i-$th customer of the $j-$th restaurant



1. Pr(Sits to a table serving a new dish)

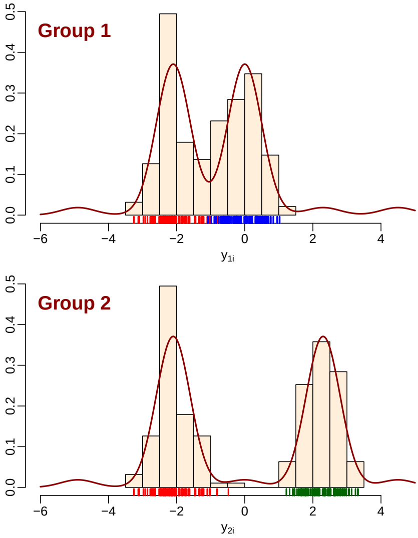
$$\boxed{= \frac{V(n_1,\ldots,n_j+1,\ldots,n_d,K+1)}{V(n_1,\ldots,n_j,\ldots,n_d,K)}\Lambda\gamma_j}$$

   In this case a new table serving the new dish is laid in every restaurant $M^{(na)} = M^{(na)} + 1$

2. Pr(Sits to an existing table)

$$\boxed{\frac{V(n_1,\ldots,n_j+1,\ldots,n_d,K)}{V(n_1,\ldots,n_j,\ldots,n_d,K)}(n_j + \gamma_j)}$$

Allocated and non allocated components



- ✓ hierarchical mixture with $M = 5$ **shared components**
- ✓ I draw $n_j = 200$ observation for each group
- ✓ $K_1 = 2$ **clusters** in the first group and $K_2 = 2$ **clusters** in the second group
- ✓ For a total of $M^{(a)} = 3$ **allocated components**. Note that $M^{(a)} = K$
- ✓ and a total of $M^{(na)} = M - M^{(a)} = 2$ non-allocated (empty) components

Posterior Characterization 1/2

Theorem 2 – *Posterior law*

Given the sequence of partitions $\rho_1, \ldots, \rho_d$, the number of clusters K and the unique values $\theta_1^{\star\star}, \ldots, \theta_K^{\star\star}$ for each j there exists an auxiliary random variable U_j such that conditionally to $U_j = u_j$ the law of P_j coincides with the normalization of the following:

$$\sum_{k=1}^{K} S_{jk}^{(a)} \delta_{\theta_k^{\star\star}}(\cdot) + \sum_{m=1}^{M^{(na)}} S_{jm}^{(na)} \delta_{\tau_m}(\cdot) \qquad \tau_m \overset{\text{i.i.d.}}{\sim} p_0$$

1. **Allocated** jumps: for each $k = 1, \ldots, K$

$$\boxed{S_{jk}^{(a)} \sim \text{gamma}(n_{jk} + \gamma_j, u_j + 1)}$$

2. **Non-allocated** jumps: for each $m = 1, \ldots, M^{(na)}$

$$\boxed{S_{jm}^{(na)} \sim \text{gamma}(\gamma_j, u_j + 1)}$$

Posterior Characterization 2/2

Theorem 2 – *Posterior law*

3. Number of **non-allocated components**

$$\boxed{M^{(na)} \sim Q_1 \mathscr{P}_0(\Lambda \prod_{j=1}^d \psi_{\gamma_j}(u_j)) + (1 - Q_1)\mathscr{P}_1(\Lambda \prod_{j=1}^d \psi_{\gamma_j}(u_j))}$$

   where $\mathscr{P}_i$ is the $i-$shifted Poisson distribution, and Q_1 is a simple algebraic expression depending on $\psi(u_j)$, Λ and $M^{(a)}$.

4. **Latent variable**: For each $j = 1, \dots, d$

$$\boxed{[U_j | \boldsymbol{S}_j^{(a)}, \boldsymbol{S}_j^{(na)}] \sim \mathrm{Gamma}(n_j, \sum_k S_{jk}^{(a)} + \sum_m S_{jm}^{(na)})}$$

- ✓ This result is the finite dimensional counterpart of the posterior characterization for the Normalized Completely Random Measures (Camerlenghi et al., 2019)
- ✓ **quasi-conjugacy**: conditionally to the latent variables U_j the posterior weights can still be obtained by normalization of independent gamma r.v.'s with updated parameters

Conditional blocked Gibbs sampler: sketch

Augment the state space introducing the rv's U_j

Parameters: $U_j,\ \theta_j,\ P_j \quad j = 1, \ldots, d$

Conditional blocked Gibbs sampler: sketch

Augment the state space introducing the rv's U_j

Parameters: $U_j,\ \boldsymbol{\theta}_j,\ P_j \quad j = 1, \dots, d$

Sequentially update the parameter as follows:

1. sample $U_j | rest$ from a Gamma$(n_j, \sum_m S_{jm})$

2. Sample $\boldsymbol{\theta}_j | rest$, for each $i = 1, \dots, n_j$ consider the discrete distribution

$$Pr(\theta_{ji} = \tau_m | rest) \propto S_{jm} f(y_{ji} | \tau_m), \quad m = 1, \dots, M$$

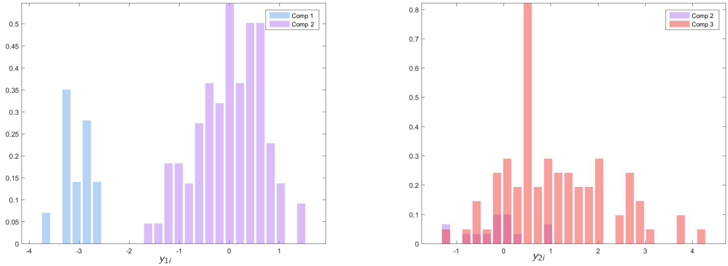
3. Update the r.p.m. $P_j | rest$ using the posterior characterization provided by Theorem 2

A hint to the simulation study

- ✓ $n = 200$ observations from two groups, both with two components, one shared component. y_{new}

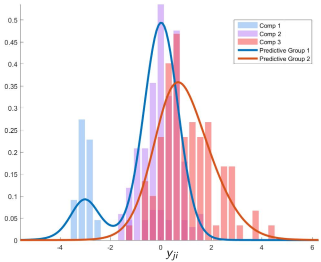
$$y_{1i} \stackrel{\text{i.i.d.}}{\sim} w_{11} N\left(\mu_1, \sigma_1^2\right) + w_{12} N\left(\mu_2, \sigma_2^2\right), i = 1, \ldots, n_1$$
$$y_{2i} \stackrel{\text{i.i.d.}}{\sim} w_{21} N\left(\mu_2, \sigma_2^2\right) + w_{22} N\left(\mu_3, \sigma_3^2\right), i = 1, \ldots, n_2,$$

- ✓ $(w_{11}, w_{12}) = (0.2, 0.8)$ and $(w_{21}, w_{22}) = (0.1, 0.9)$.
- ✓ $(\mu_1, \sigma_1^2) = (-3, 0.1)$, $(\mu_{12}, \sigma_{12}^2) = (\mu_{21}, \sigma_{21}^2) = (0, 0.5)$, and $(\mu_3, \sigma_3^2) = (1, 1.5)$.

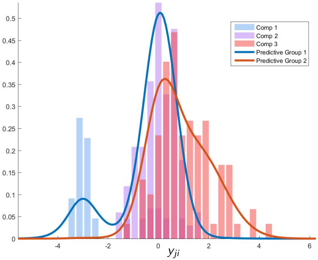


Results

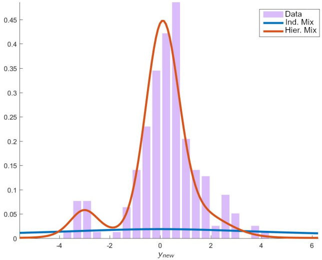
Independent mixtures



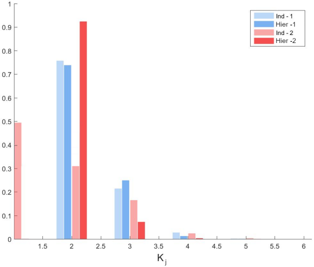
Hierarchical mixture



Predictive in a new group



Local clustering



Wrap-up

- ✓ A novel **Bayesian hierarchical model** for analyzing **grouped data** yielding to a clustering within and between groups
- ✓ **Analytical expression** of the partial exchangeable partition probability function (**eppf**), i.e., the law of the random partition induced by the model and a **posterior characterization** of the hierarchical process
- ✓ Restaurant Franchise process and set up both **marginal and conditional Gibbs sampler**
- ✓ Preliminary results are promising in terms of clustering identification and density estimation

Wrap-up

- ✓ A novel **Bayesian hierarchical model** for analyzing **grouped data** yielding to a clustering within and between groups
- ✓ **Analytical expression** of the partial exchangeable partition probability function (**eppf**), i.e., the law of the random partition induced by the model and a **posterior characterization** of the hierarchical process
- ✓ Restaurant Franchise process and set up both **marginal and conditional Gibbs sampler**
- ✓ Preliminary results are promising in terms of clustering identification and density estimation

Thanks!!!

Biased Bibliography

- ✓ Camerlenghi, F., Lijoi, A., Orbanz, P., & Prünster, I. (2019). Distribution theory for hierarchical processes. The Annals of Statistics, 47(1), 67-92.
- ✓ Raffaele Argiento, Andrea Cremaschi & Marina Vannucci (2020) Hierarchical Normalized Completely Random Measures to Cluster Grouped Data, Journal of the American Statistical Association, 318-333
- ✓ Argiento, R., & De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. Annals of Statistics (accepted)

References

Argiento, R., Cremaschi, A., and Vannucci, M. (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association*.

Argiento, R. and De Iorio, M. (2022). Is infinity that far? a Bayesian nonparametric perspective of finite mixture models. *Annals of Statistics*.

Beraha, M., Guglielmi, A., and Quintana, F. A. (2021). The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions. *Bayesian Analysis*, pages 1 – 33.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92.

D'Angelo, L., Canale, A., Yu, Z., and Guindani, M. (2022). Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. *Biometrics*.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013). Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37(2):212–229.

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2021). A common atoms model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association*, in press.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581.