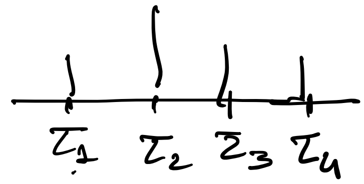
$$Y_1, Y_2, \ldots, Y_n \mid \theta_1, \theta_2, \ldots, \theta_n \overset{ind}{\sim} f(y_i \mid \theta_i)$$

$$\underbrace{\theta_1, \theta_2, \ldots, \theta_n}_{\underline{\theta}} \mid P \overset{iid}{\sim} P$$

$$P \sim Dir(\alpha P_0)$$

$$P = \sum_{J=1}^{\infty} w_J \, \delta_{Z_J}(\cdot)$$



Sicuramente ci sono ripetizioni in $\underline{\theta}$

$\theta_1^*, \theta_2^*, \ldots, \theta_k^*$ & Definiscono un clustering

$i \in C_J$ sse $\theta_i = \theta_J^*$

$$\rho = \{C_1, C_2, \ldots, C_k\}$$

So scrivere la prior di ρ si chiama

eppf

$$p(\rho) \cancel{\propto} \quad \prod_{J=1}^{k} \Gamma(|C_J|)$$

$$= (|C_J|-1)!$$

Questa prior è di tipo product partition ala funzione scelle numerosità si chiama coesione

Possiamo riscrivere il modello mistura come segue

$$Y_1, Y_2, \dots, Y_n \mid \rho, \theta_1^*, \dots, \theta_K^*$$

$$\sim \prod_{j=1}^{K} \Big\{ \prod_{i \in C_j} f(y_i \mid \theta_j^*) \Big)$$

$$\rho \sim \text{eppf}$$

$$\theta_1^*, \dots, \theta_K^* \mid K \overset{iid}{\sim} P_0$$

$\boxed{\text{PPMx}}$

$x_1, x_2, \dots, x_n$ $\qquad$ $x_i \in \mathbb{R}^q$ è un vettore di covariate per il dato i

$$x_j^* = (x_i)_{i \in C_j}$$

Cerchiamo di costruire una prior per ρ informata dalle x

$$p(\rho) \propto \prod_{j=1}^{k} \Gamma(|C_j|)\, \underbrace{g(x^*_j)}_{\text{similarità}}$$

Nel modello PPM l'inferenza si basa sur Pólya urn representation

$$p(\theta_{n+1} \mid \theta_1, \dots, \theta_n)$$

$$P(n+1 \in C_j \mid \rho, \theta^*_1, \dots \theta^*_k) = (*)$$

$$P(n+1 \in C^{new} \mid \rho, \theta^*_1, \dots \theta^*_k) = (**)$$

$$(*) = \frac{eppf(C_1, C_2, \dots, C_j \cup n+1, \dots, C_k)}{eppf(C_1, C_2, \dots, C_k)} \propto |C_j|$$

$$(**) = \frac{eppf(C_1, C_2, \dots, C_k, \{n+1\})}{eppf(C_1, C_2, \dots, C_k)} \propto \alpha$$

Nel PPMx

$$* \propto \frac{c(|e_j \cup \{n+1\}|)}{c(|e_j|)} \cdot \frac{g(x_j^* \cup x_{n+1})}{g(x_j^*)}$$

$$** \propto c(\{n+1\})\, g(x_{n+1})$$

---

da fare: plot nello spazio

p vs σ

Vedere se queste 2 features inducono un cluster