Il processo di Pitman-Yor e modelli mistura di processo di Pitman-Yor

Sara Capozio, Luca De Simone, Anna Petranzan

July 2024

1 Introduzione

Il processo di Pitman-Yor è una generalizzazione del processo di Dirichlet applicabile nel momento in cui si vogliono modellare fenomeni che seguono distribuzioni asimmetriche con code pesanti. Il processo di Dirichlet risulta inadeguato nel modellare tali tipologie di distribuzioni in quanto il numero medio di cluster distinti K_n generati dal processo cresce logaritmicamente all'aumentare della numerosità campionaria:

$$E[K_n] \approx a \log(n).$$

Dalla seguente espressione emerge che da un certo n in poi il numero medio di cluster si stabilizza e questo determina, al crescere di n, cluster molto simili in termini di numerosità.
Da un punto di vista empirico tale rappresentazione risulta essere poco realistica in quanto se volessimo modellare fenomeni con distribuzione fortemente asimmetrica come ad esempio la misura di frequenza delle parole o il numero di follower per utente su Twitter, K_n dovrebbe seguire una distribuzione a code pesanti come una distribuzione power law.

Definizione 1. Una distribuzione di probabilità P segue una distribuzione power law se la sua funzione di densità presenta la seguente forma:

$$p(x) = c \cdot x^{-a}.$$

Il processo che permette di generare un numero di clusters distinti distribuiti tramite una distribuzione power law è il processo di Pitman-Yor, $PY(d, c, G_0)$, che a differenza del processo di Dirichlet è caratterizzato da un parametro aggiuntivo ovvero d.

Definizione 2. Sia G_0 la misura base di probabilità su $\mathbb{X}$ (supporto delle osservazioni), $d \in [0, 1)$ coefficiente di penalizzazione e $c > -d$ parametro di concentrazione, allora la misura di probabilità aleatoria

$$\tilde{P} = \sum_{j=1}^{\infty} p_j \cdot \delta_{X_j}$$

dove:

$$(X_j)_{j\geq 1} \overset{\text{iid}}{\sim} G_0$$
$$(p_j)_{j\geq 1} \sim GEM(c, d)$$

è detto processo di Pitman-Yor.

Anche nel seguente caso i pesi di un processo di Pitman-Yor possono essere costruiti tramite stick breaking le cui componenti sono definite nel seguente modo:

$$(W_j)_{j\geq 1} \overset{\text{ind}}{\sim} Beta(1-d; c+j\cdot d)$$

$$p_1 = W_1$$

$$p_j = W_j \cdot \prod_{r=1}^{j-1}(1-W_r).$$

2 Casi particolari processo Pitman-Yor

La seguente sezione si pone l'obiettivo di illustrare due casi particolari del processo di Pitman-Yor:

1. Caso d=0: fissando il parametro di penalizzazione a zero il processo di Pitman-Yor coincide con il processo di Dirichlet. Siano $E_i \overset{\text{iid}}{\sim} Exp(1)$ con $i = 1, 2, \ldots,$ i tempi di inter-arrivo di un processo di Poisson omogeneo con tasso di arrivo unitario e $\Gamma_k = E_1 + ... + E_k$ i tempi di attesa del k-esimo evento. Sia γ_α un subordinatore gamma, ovvero un processo stocastico gamma ad incrementi stazionari e indipendenti, tale per cui $\gamma_\alpha = \sum_{j=1}^{\infty} J_k$, dove $J_1 > J_2 > .... > 0$ è la sequenza dei salti ordinati nell'intervallo di tempo $[0, \alpha]$. Sia $\nu(x)$, la misura di Lévy associata ad un processo gamma che misura la frequenza dei salti di ampiezza x:

$$\nu(x) = \alpha \int_x^{\infty} xe^{-x}\, dx$$

Sfruttando le quantità appena esplicitate, è possibile definire un metodo alternativo allo stick breaking per la determinazione dei pesi di un processo di Dirichlet:

$$\tilde{P} = \sum_{k=1}^{\infty} \frac{\nu(\Gamma_k)^{-1}}{\sum_{k=1}^{\infty}\nu(\Gamma_k)^{-1}} \cdot \delta_{X_k} = \sum_{k=1}^{\infty} \frac{J_k}{\sum_{k=1}^{\infty} J_K} \cdot \delta_{X_k} \overset{\text{d}}{=} V_1 \cdot \delta_{X_1} + \sum_{k=2}^{\infty} \left[\left(\prod_{j=1}^{k-1}(1-V_j) \right) V_k \right] \cdot \delta_{X_k}$$

Nel momento in cui gli atomi vengono campionati dalla misura base, alla prima variabile estratta verrà assegnato peso maggiore, in quanto il tempo di attesa necessario per osservare X_1 è ridotto con una frequenza ad essa associata pari all'incremento maggiore, ovvero J_1. Dal Teorema centrale del limite è noto che la somma di N variabili indipendenti e identicamente distribuite, per $N \to \infty$ si distribuisce come una distribuzione Normale. Per questo motivo ci si attende che il processo di Dirichlet tenderà a distribuirsi in modo simmetrico. Al contrario, nel momento in cui si è interessati a modellare dei fenomeni asimmetrici, entra in gioco il processo di Pitman-Yor.

2. Caso c=0: fissando il parametro di concentrazione a zero il processo ottenuto è un processo di Pitman-Yor $PY(d)$. Siano $E_i \overset{\text{iid}}{\sim} Exp(1)$, i tempi di inter-arrivo di un processo di Poisson omogeneo con tasso di arrivo unitario e $\Gamma_k = E_1 + ... + E_k$ i tempi di attesa del k-esimo evento. Sia γ_α un subordinatore stabile. La caratteristica fondamentale del seguente processo consiste nel preservare la proprietà di asimmetria del fenomeno. Infatti, è noto che la somma di N variabili aleatorie indipendenti e identicamente distribuite seguano, a loro volta, una legge stabile con gli stessi parametri di stabilità, dove per $\alpha < 2$, la legge stabile coincide con una power law. La misura di Lévy associata ad un processo che segue una legge stabile è la seguente:

$$\nu(x) = c \cdot x^{\alpha}.$$

Sfruttando le quantità appena esplicitate, è possibile definire un metodo alternativo al metodo stick breaking per il campionamento dei pesi di un processo di $PY(d)$:

$$\tilde{P} = \sum_{k=1}^{\infty} \frac{\Gamma_k^{-1/\alpha}}{\sum_{k=1}^{\infty} \Gamma_k^{-1/\alpha}} \cdot \delta_{X_k} = \sum_{k=1}^{\infty} \frac{J_k}{\sum_{k=1}^{\infty} J_k} \cdot \delta_{X_k} \overset{\mathrm{d}}{=} V_1 \cdot \delta_{X_1} + \sum_{k=2}^{\infty} \left[\left(\prod_{j=1}^{k-1} (1 - V_j) \right) V_k \right] \cdot \delta_{X_k}.$$

3 Analisi simulativa

Tramite l'utilizzo del software R sono state svolte varie simulazioni al fine di comprendere i ruoli dei parametri c e d ed analizzare le differenze tra il processo di Dirichlet e quello di Pitman-Yor.

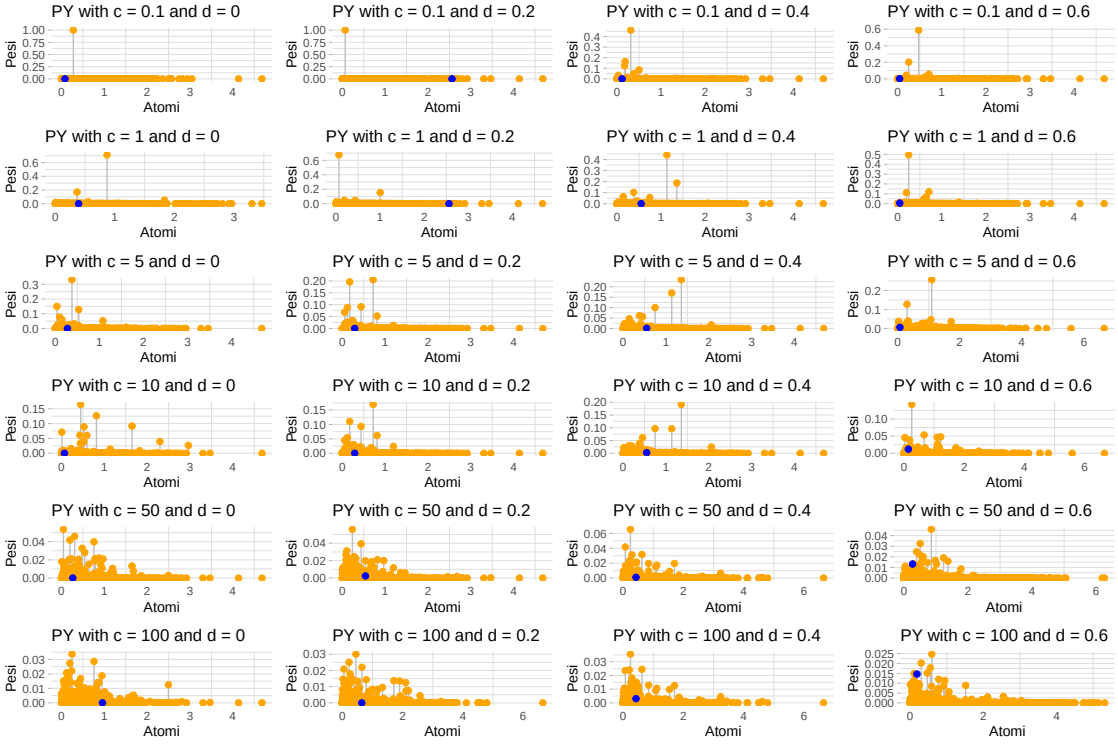

Figure 1: Scatterplot dei pesi simulati

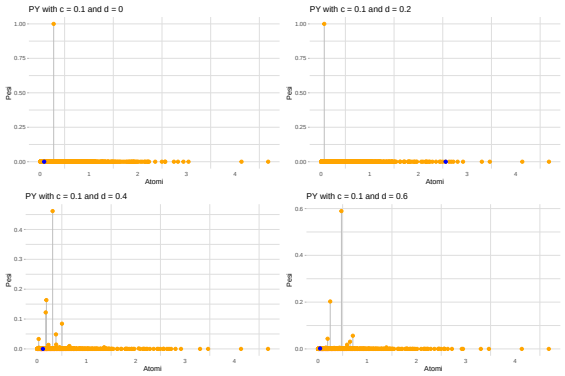

Figure 2: $c = 0.1$

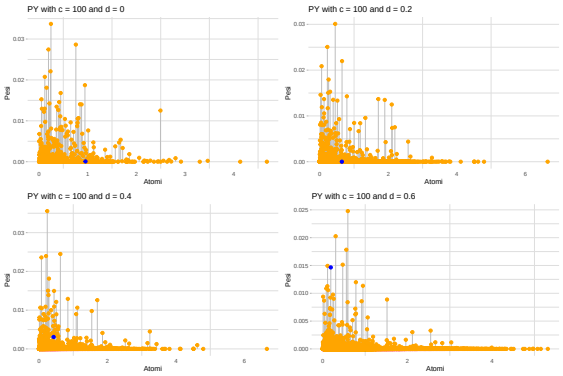

Figure 3: $c = 100$

Figure 4: Scatterplot dei pesi simulati fissato c

Dai seguenti grafici risulta possibile studiare il ruolo dei parametri c e d all'interno del processo.
Il parametro di concentrazione c definisce la forza della prior guess G_0, ovvero stabilisce quanto si è convinti che le realizzazioni del processo si distribuiscano secondo la misura base G_0. Per questo studio si è supposto a priori di utilizzare una distribuzione $Gamma(\alpha = 1, \beta = 2)$, in quanto più attinente a modellare fenomeni asimmetrici con code più o meno pesanti.
Il parametro di penalizzazione d permette di modellare le code della distribuzione a priori. Infatti, dalla simulazione emerge che il crescere di d penalizza la massa di probabilità assegnata agli atomi più frequenti distribuendola su un maggior numero di realizzazioni x_j rendendo le code della distribuzione più pesanti.

Dall'analisi (Figura 4) è possibile osservare che i parametri c e d lavorano nella stessa direzione, ovvero per valori bassi di questi, la massa di probabilità delle realizzazioni di $\tilde{P}$ è concentrata su pochi atomi. Al contrario, il loro aumentare, permette ad un numero maggiore di atomi di avere una massa di probabilità significativamente positiva.

Nel processo di Pitman-Yor gli atomi campionati presentano una maggior numerosità e una densità inferiore rispetto a quelli generati da un processo di Dirichlet. Infatti, analizzando più nel dettaglio la costruzione stick breaking dei due processi è possibile osservare che nel processo di Dirichlet, le variabili aleatorie $W_j \overset{\text{iid}}{\sim} Beta(1, c)$ sono indipendenti e identicamente distribuite con valore atteso pari a

$$E[W_j] = \frac{1}{c+1}.$$

Mentre, nel processo di Pitman-Yor, le variabili $W_j \overset{\text{ind}}{\sim} Beta(1-c, c+jd)$ sono solamente indipendenti con valore atteso

$$E[W_j] = \frac{1-d}{1-d+c+jd},$$

che decresce all'aumentare di j.

Le formule evidenziano che mediamente i pesi generati dal processo di Dirichlet sono maggiori rispetto a quelli di Pitman-Yor, la cui differenza aumenta al crescere di j. Questo comporta che, per esaurire la probabilità totale, è necessario generare un numero maggiore di pesi e conseguentemente un maggior numero di atomi. Tale considerazione determina, per valori elevati di d, un problema computazionale dettato dall'incapacità dell'algoritmo di esaurire totalmente la probabilità residua $1 - \sum_{j=1}^{N} p_j$ e quindi la generazione di una misura di probabilità. Per arginare la problematica, una strategia che si può adottare è quella di stabilire a priori un troncamento fissando una certa quantità $\epsilon > 0$ che si è disposti ad accettare come residua. Se $1 - \sum_{j=1}^{N} p_j > \epsilon$ allora si prosegue la generazione di nuove realizzazioni. Tale tecnica risulta però computazionalmente onerosa e per questa ragione, nell'analisi, si sono considerati diversi valori di troncamento per le diverse combinazioni

di c e d. Si noti che utilizzare diversi ϵ potrebbe far emergere delle difficoltà nel confronto tra grafici, in quanto una maggior concentrazione dei pesi potrebbe essere imputabile ad una maggior numerosità dovuta ad un diverso valore del parametro.
In Figura 1 e 4 i residui ottenuti $1 - \sum_{j=1}^{N} p_j$ sono identificati dai pallini blu.

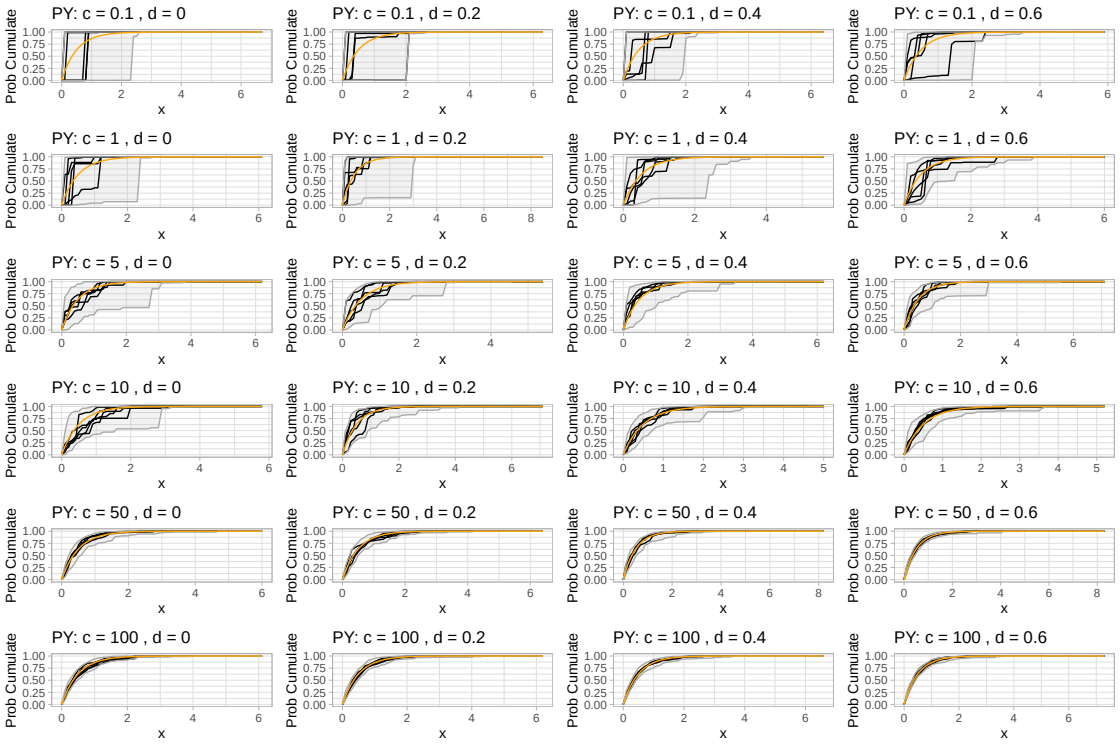

Figure 5: ECDF simulate

I grafici dell'ECDF (Figura 5) permettono di osservare ulteriormente l'effetto dei parametri. All'aumentare di c e d, la variabilità del processo diminuisce, infatti le realizzazioni di $\tilde{P}$ si concentrano attorno alla misura base G_0.

4 Proprietà

Sia $\tilde{P} \sim PY(c, d, G_0)$ definito sullo spazio $\mathbb{X}$ e $A \subseteq \mathbb{X}$, un sottoinsieme misurabile di $\mathbb{X}$, allora:

$$E[\tilde{P}(A)] = G_0(A),$$

$$\text{Var}[\tilde{P}(A)] = G_0(A) \cdot (1 - G_0(A)) \cdot \frac{1-d}{c+1}.$$

Il valore atteso del processo di Pitman-Yor, come nel caso del processo di Dirichlet, coincide con la misura base valutata su un sottoinsieme misurabile di $\mathbb{X}$. Al contrario, la variabilità, nel caso del processo di Pitman-Yor, è controllata da due parametri: il parametro di concentrazione c e il parametro di penalizzazione d. Come osservato precedentemente, è possibile affermare che la variabilità del processo tende a diminuire al crescere di c e/o per valori di d elevati prossimi a uno.

5 Distribuzione Predittiva

Sia $X_i|\tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$ con $\tilde{P} \sim PY(c, d, G_0)$. La distribuzione predittiva di X_{n+1} date le osservazioni $X_1, \ldots, X_n$:

$$P(X_{n+1} \in A \mid X_1, \ldots, X_n) = \frac{c + k \cdot d}{n + c} \cdot G_0(A) + \sum_{j=1}^{k} \frac{n_j - d}{n + c} \cdot \delta_{X_j^*}(A),$$

per ogni $A \subseteq \mathbb{X}$, sottoinsieme misurabile di $\mathbb{X}$, dove $\{X_1^*, \ldots, X_k^*\}$, con $k \leq n$, denotano i valori unici nel set $\{X_1, \ldots, X_n\}$ con frequenze $n_1, \ldots, n_k$.

Da questa espressione segue che la distribuzione congiunta di $\{X_1, \ldots, X_n\}$ può essere definibile a partire dalla generalizzazione dello schema dell'urna di Polya, tale per cui:

$$X_1 \text{ è campionato da } G_0, \quad \textit{per} \quad i = 2, \ldots, n$$

$$X_i \mid X_1, \ldots, X_{i-1} \sim \begin{cases} \text{nuovo valore campionato da } G_0, & \text{con probabilità } \dfrac{c + k \cdot d}{i - 1 + c}, \\ \text{valore già osservato } X_j^*, & \text{con probabilità } \dfrac{n_j - d}{i - 1 + c}. \end{cases}$$

In un processo di Pitman-Yor, la probabilità di osservare una nuova osservazione dipende non solo dal parametro di concentrazione c, ma anche dal parametro di penalizzazione d e dal numero di osservazioni distinte generate precedentemente. Infatti, il coefficiente di penalizzazione d influisce sia sulla probabilità di generare nuovi cluster e sia sulla crescita di quelli già esistenti. Tanto più d e c assumono valori elevati tanto più la probabilità di generare nuovi cluster distinti aumenta riducendo, allo stesso tempo, la probabilità di popolare i cluster già esistenti. Dunque, a differenza del processo di Dirichlet, in cui i cluster generati tendono ad avere in media una numerosità simile, nel processo di Pitman-Yor si avranno molti cluster poco densi e pochi molto popolati.

6 Modelli mistura di Pitman-Yor

Il processo di Pitman-Yor risulta utile applicarlo quando si suppone che la distribuzione delle osservazioni $X_1, \ldots, X_n$ sia discreta. Se le osservazioni seguono una distribuzione continua, non può essere applicato direttamente per problemi di stima di densità, per questo motivo vengono introdotti i cosiddetti modelli mistura.

Una mistura di processo di Pitman-Yor su $\mathbb{X}$ è una PDF aleatoria $\tilde{f}$ definita come:

$$\tilde{f}(x) = \int_{\Theta} K(\theta, x) d\tilde{P}(\theta)$$

dove $\tilde{P} \sim PY(c, d, G_0)$ con G_0 misura di probabilità su Θ. Questo modello può essere riscritto alternativamente tramite la sua forma gerarchica:

$$\begin{aligned} X_i|\theta_i &\overset{\text{ind}}{\sim} K(X_i, \theta_i); \\ \theta_i|\tilde{P} &\overset{\text{iid}}{\sim} \tilde{P}; \\ \tilde{P} &\sim PY(c, d, G_0). \end{aligned}$$

L'obiettivo è stimare la densità della distribuzione delle osservazioni $\boldsymbol{X} = (X_1, \ldots, X_n)$ tramite

$$\hat{f} = E[\tilde{f}(x)|\boldsymbol{X}].$$

Tale quantità risulta stimabile mediante metodo Monte Carlo. Siano $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ un insieme di M realizzazioni dalla distribuzione $\boldsymbol{\theta}|\boldsymbol{X}$ dove $\boldsymbol{\theta}^{(m)} = (\theta_1^{(m)}, \ldots, \theta_n^{(m)})$ per ogni $m = 1, \ldots, M$, allora:

$$\hat{f}(x) \approx \frac{1}{M} \sum_{m=1}^{M} E\left[\tilde{f}(x) \mid \boldsymbol{X}, \boldsymbol{\theta}^{(m)}\right] \tag{1}$$

dove, condizionatamente alle variabili ausiliari $\boldsymbol{\theta}$:

$$E[\tilde{f}(x)|\boldsymbol{X}, \boldsymbol{\theta}] = \frac{c + d \cdot k}{c + n} \int_{\Theta} K(X, \theta) g_0(\theta)\, d\theta + \sum_{j=1}^{k} \frac{(n_j - d)}{c + n} K(X; \theta_j^*) \tag{2}$$

Per campionare dalla distribuzione condizionata di $\boldsymbol{\theta}|\boldsymbol{X}$ si utilizza l'algoritmo Gibbs Sampling dove le distribuzioni full conditional per θ_i condizionatamente a $\boldsymbol{\theta}_{(-i)} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ e $\boldsymbol{X} = (X_1, \ldots, X_n)$, sono definite nel seguente modo:

$$\begin{aligned} P(\theta_i \in d\theta|\boldsymbol{X}, \boldsymbol{\theta}_{(-i)}) \propto \frac{c + dk_{(-i)}}{c + n - 1} \int_{\Theta} K(X_i, \theta) g_0(\theta)\, d\theta \cdot \frac{K(X_i, \theta) \cdot g_0(\theta)}{\int_{\Theta} K(X_i, \theta) \cdot g_0(\theta)\, d\theta} \\ + \sum_{j=1}^{k_{(-i)}} \frac{(n_{j(-i)} - d)}{c + n - 1} K(X_i; \theta_{j(-i)}^*) \delta_{\theta_{j(-i)}^*}(\theta). \end{aligned} \tag{3}$$

dove g_0 è la funzione densità associata alla misura di probabilità G_0 e $k_{(-i)}$ il numero di valori distinti in $\boldsymbol{\theta}_{(-i)}$ con frequenze $(n_{1(-i)}, \ldots, n_{k_{(-i)}(-i)})$.

7 Simulazione modello Location Scale Pitman-Yor

Sia dato il modello Location Scale Pitman-Yor con kernel Gaussiano definito nel seguente modo:

$$\begin{aligned} X_i \mid \theta_i, \sigma^2 &\overset{\text{ind}}{\sim} N(\theta_i, \sigma^2) \qquad i = 1, \ldots, n \\ \theta_i \mid \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} \qquad i = 1, \ldots, n \\ \tilde{P} &\sim PY(c, d, G_0) \\ G_0 &\sim N(0, 1) \\ \sigma^2 &\sim Inv - Gamma(a, b) \end{aligned}$$

con $c > -d, d \in [0, 1], a > 0$ e $b > 0$.

Nella seguente sezione si riportano i risultati relativi all'applicazione di tale modello ad un dataset simulato di $n = 1000$ osservazioni indipendenti e identicamente distribuite generate dalla seguente mistura di distribuzioni Normali:

$$0.3 \cdot N(y, 0, 1) + 0.6 \cdot N(y, 3, 1) + 0.1 \cdot N(y, 7, 1).$$

La densità media a posteriori $E[\hat{f}(x)|\boldsymbol{X}]$ si è stimata utilizzando un algoritmo di tipo marginale articolato su due step:

STEP 1: simulazione del vettore $\boldsymbol{\theta}$ da $P(\theta_i \mid \boldsymbol{X}, \boldsymbol{\theta}_{-i})$, Eq.(3), full conditional stimata tramite l'algoritmo Gibbs Sampling utilizzando un numero di iterazioni pari a 100 e un burn in pari al 10% delle iterazioni totali.

STEP 2: stima di $E[\hat{f}(x)|\boldsymbol{X}]$ via Monte Carlo combinando le equazioni (1) e (2).

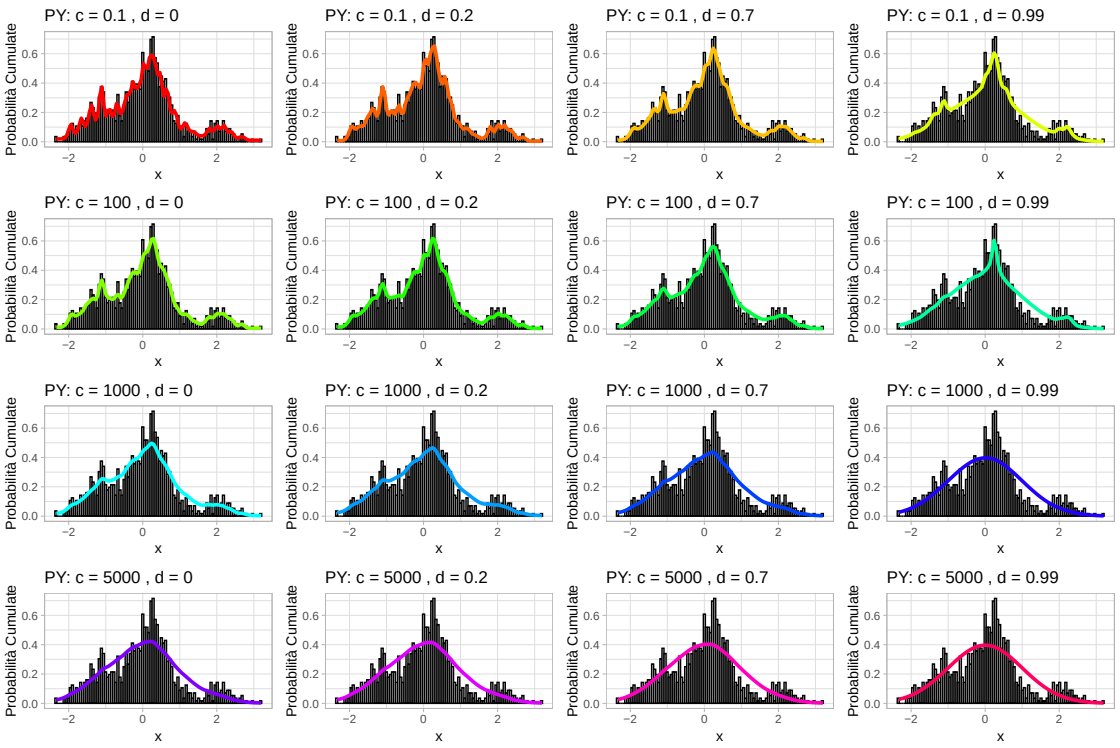

Figure 6: Modelli mistura su dati simulati al variare dei parametri

L'obiettivo dell'analisi consiste nello studiare come varia la densità media stimata a posteriori al variare dei parametri c e d. Dai grafici in Figura 6 emerge che per valori bassi di c e d, la funzione di densità stimata si adatta perfettamente ai dati cogliendo picchi e fluttuazioni della mistura stessa. Questo deriva dal fatto che per valori ridotti dei parametri, l'impatto della prior guess è pressoché ininfluente come evidenziato dall'Equazione (2). Infatti, in tale scenario la densità a posteriori viene campionata soprattutto dalla componente empirica:

$$\sum_{j=1}^{k} \frac{(n_j - d)}{c+n} K(X; \theta_j^*)$$

ottenendo così una buona approssimazione alla mistura.

Al contrario al crescere di c e d, la funzione media stimata a posteriori risulta essere sempre più smooth. Quando la componente relativa alla prior guess è più preponderante rispetto a quella empirica, la densità media a posteriori viene campionata dalla componente a priori

$$\frac{c + d \cdot k}{c+n} \int_{\Theta} K(X, \theta) g_0(\theta)\, d\theta$$

ottenendo così una densità a posteriori approssimabile alla misura base.

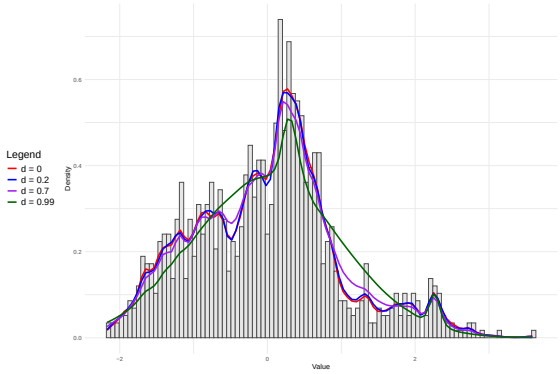

Figure 7: $c = 100$

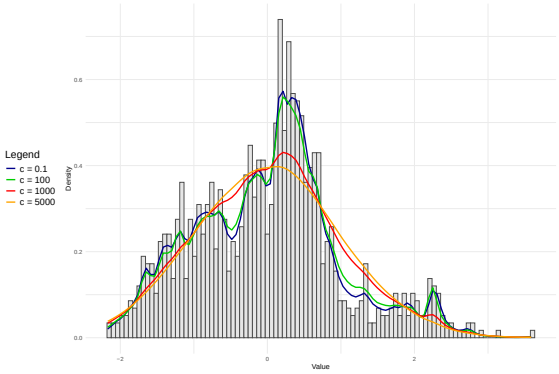

Figure 8: $d = 0.7$

Figure 9: Modelli mistura su dati simulati mediante algoritmo marginale

In conclusione, la Figura 9 mostra ancor più nel dettaglio il ruolo dei parametri parametri c e d nel modellare il livello di smoothness della funzione di densità stimata.

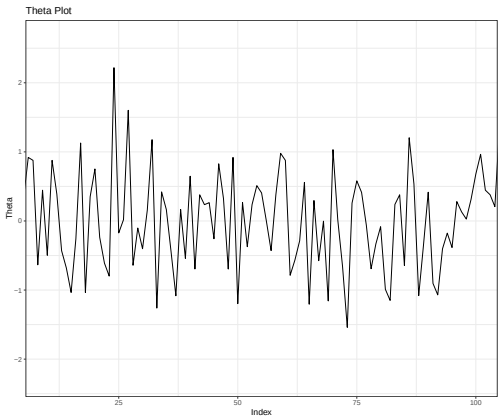

Figure 10: Trace plot parametri θ

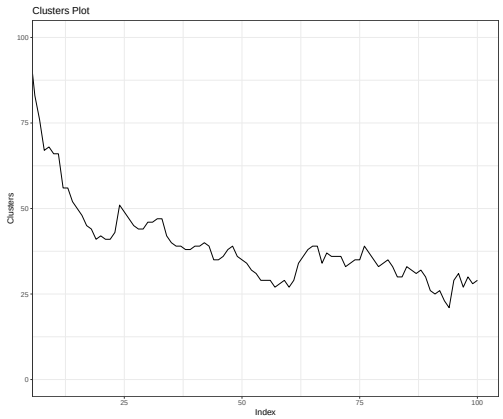

Figure 11: Trace plot numero di cluster

Figure 12: Trace plot

La Figura 12 permette di monitorare la convergenza dell'algoritmo di generazione. Da questa emergono alcuni problemi nella convergenza della stima del numero di cluster, in quanto il trace plot non tende a stabilizzarsi attorno ad un valore medio costante ma, al contrario, continua a decrescere. Per quanto riguarda il vettore dei parametri $\boldsymbol{\theta}$, le prestazioni dell'algoritmo risultano essere discrete. In conclusione, le stime di densità che si ottengono sono accettabili, ma non pienamente soddisfacenti. Una soluzione per ottenere delle stime più accurate potrebbe essere aumentare il numero di iterazioni e il periodo di burn in. Utilizzando l'algoritmo Gibbs Sampling manualmente implementato, incrementare il numero delle iterazioni comporta un tempo computazionale eccessivo. Per questo motivo si è deciso di utilizzare la funzione *PYdensity* del pacchetto *BNPmix* per stimare la densità della mistura, considerando un numero di iterazioni pari a 1000 e un burn in pari a 100. Il pacchetto *BNPmix* presenta tre metodi diversi di aggiornamento dei valori di θ nella catena MCMC. Nello specifico, il metodo utilizzato è il metodo MAR (Marginal Sampler).

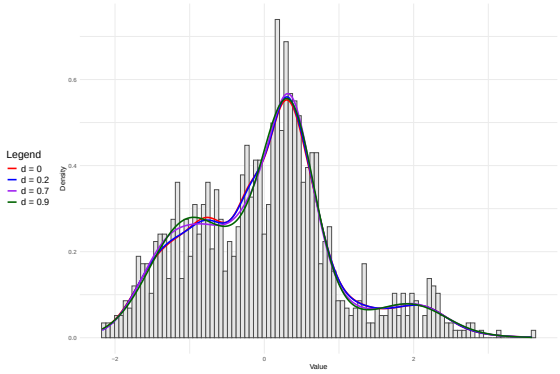

Figure 13: $c = 100$

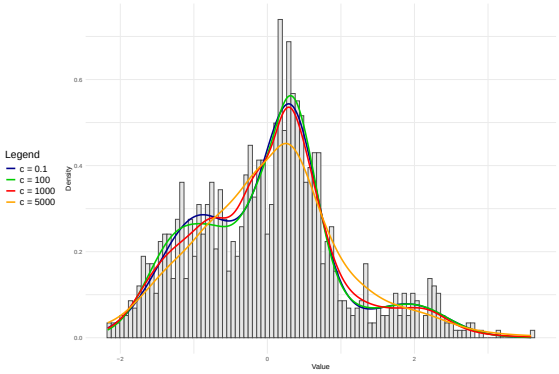

Figure 14: $d = 0.7$

Figure 15: Densità simulate con *BNPmix*

Confrontando i risultati in Figura 9 e quelli in Figura 15 si può osservare che le densità stimate dal pacchetto *BNPmix* risultano essere, in generale, molto più smooth. Inoltre, per ragioni computazionali è stato scelto come valore massimo per d il valore 0.9 e non 0.99 come in precedenza. In generale, l'andamento delle stime di densità mediante *BNPmix* al variare di d è molto più uniforme e risente meno del cambiamento del parametro, infatti le curve con $d = 0$, di colore rosso, e $d = 0.2$, di colore blu, in Figura 13, non sono praticamente distinguibili.

7.1 Stima del numero di clusters

Il pacchetto *BNPmix*, dopo aver stimato i parametri θ, utilizza tre diversi metodi basati sulla similarità per individuare il numero di cluster. I primi due metodi sfruttano un algoritmo hclust con linkage average o complete, mentre il terzo calcola la matrice di similarità ($\hat{P}$) per individuare la partizione ottima da cui estrarre il numero di cluster, così definita:

$$\hat{P} = (\hat{p}_{ij}),\ i, j = 1, \ldots, n$$

$$\hat{p}_{ij} = \frac{1}{R - R_0} \sum_{r=R_0+1}^{R} I_{\{\theta_i^r = \theta_j^r\}}$$

dove R identifica il numero di iterazioni e R_0 il numero di burn in.

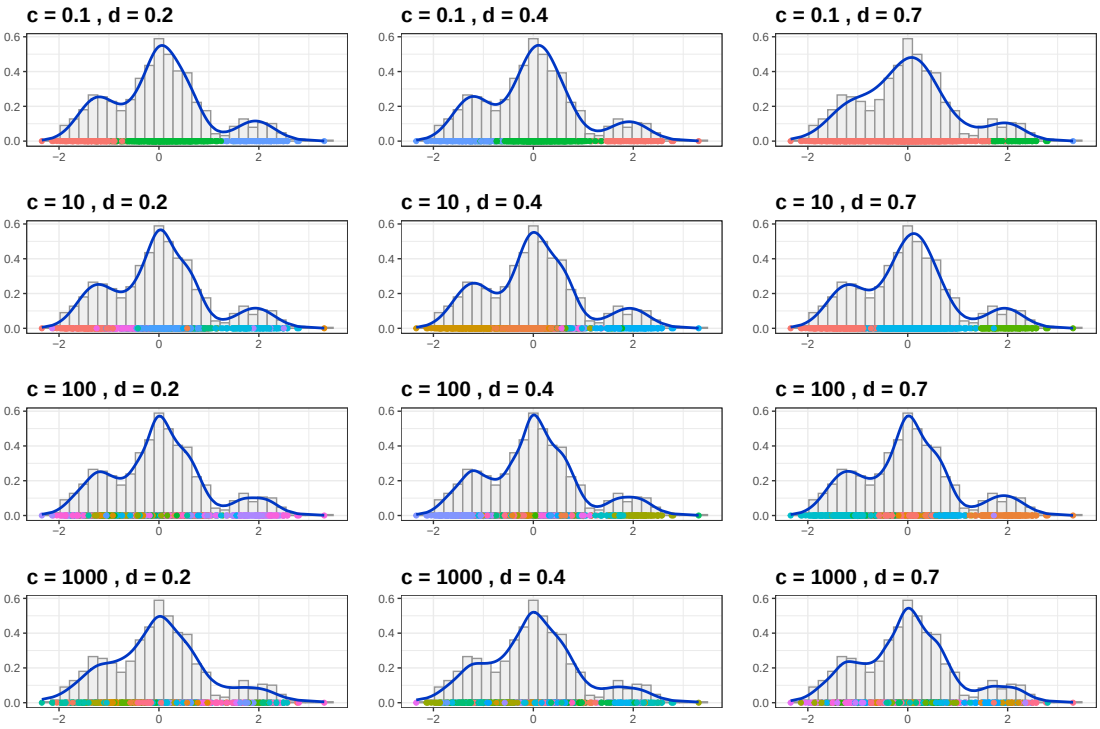

Figure 16: Numero di cluster al variare dei parametri c e d

Fissato n pari a 1000, per valori di c pari a 0.1 e 10, l'effetto della prior sulla stima della densità, e quindi sulla stima dei cluster, è pressoché nullo, indipendentemente dal valore di d. Il parametro d, infatti, agisce soprattutto sulla parte empirica

$$\sum_{j=1}^{k} \frac{(n_j - d)}{c + n} K(X; \theta_j^*)$$

andando a ridurre la probabilità di popolare cluster già esistenti di piccola numerosità. Questo si riflette sul grafico (Figura 16) in quanto all'aumentare di d il numero di cluster sembrerebbe diminuire. Al contrario quando c inizia ad assumere valori prossimi a n, il numero di cluster aumenta, effetto che tende ad amplificarsi quando d cresce.

7.1.1 Approfondimento previsione del numero di cluster

Nota la numerosità dei cluster nella popolazione di riferimento di numerosità n, risulta possibile stimare il numero di cluster a posteriori presenti in una nuova popolazione di numerosità m. Nel caso di un processo di Pitman-Yor, la forma dell'Exchangeable Partition Probability Function, EPPF, è la seguente:

$$\Pi_n(n_1, ..., n_k) = V_{n,k} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1},$$

dove i pesi non negativi $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ che rispettano la seguente relazione di ricorsività $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$, sono del tipo:

$$V_{n,k} = \frac{\prod_{i=1}^{k-1} (c + id)}{(c + 1)_{n-1}}$$

con $d \in [0,1)$ e $c > -d$.
La scrittura $(1-\sigma)_{n_j-1}$ è un simbolo di Pochhammer ed equivale al seguente rapporto: $\Gamma(n_j-\sigma)/\Gamma(1-\sigma)$.

Sulla base dell'EPPF, un'espressione esplicita per la distribuzione del numero di nuovi cluster distinti osservati in un nuovo campione addizionale, $K_m^{(n)}$, condizionatamente alle informazioni fornite da $X_1, \ldots, X_n$, è data da:

$$P(K_m^{(n)} = j | X_1, \ldots, X_n) = \frac{V_{n+m,k+j}}{V_{n,k}} \frac{\mathcal{C}(m, j; d, -n+kd)}{d^j}$$

dove $X_1, \ldots, X_n$ sono partizionate in $K_n = k$ clusters con frequenze rispettivamente $n_1, \ldots, n_k$ e $\mathcal{C}(m, j; d, -n+kd)$ è il coefficiente fattoriale generalizzato non centrale:

$$\mathcal{C}(m, j; d, -n+kd) = (j!)^{-1} \sum_{r=0}^{j} (-1)^r \binom{j}{r} (n - d(r+k))_m.$$

Tutto questo è usato per determinare lo stimatore non parametrico Bayesiano, considerando una funzione di perdita quadratica, che è pari a:

$$\hat{K}_m^{(n)} = E[K_m^{(n)} | K_n = k, N_n = \mathbf{n}]$$

con $\mathbf{n} = (n_1, \ldots, n_k)$. Nel caso del processo PY, con parametro (c, d) diventa:

$$P(K_m^{(n)} = j | K_n = k, N_n = \mathbf{n}) = \frac{(c/d+k)_j}{(c+n)_m} \mathcal{C}(m, j; d, -n+kd).$$

Lo stimatore è quindi pari a:

$$\hat{K}_m^{(n)} = \left(k + \frac{c}{d}\right) \left(\frac{(c+n+d)_m}{(c+n)_m} - 1 \right).$$

8 Applicazione a un dataset reale

In conclusione, il modello location scale Pitman-Yor è stato applicato a un dataset reale. Il dataset analizzato raccoglie dati sul calcestruzzo, in particolare circa la sua composizione e la resistenza alla compressione effettiva. Infatti, all'interno del composto, sono presenti diversi ingredienti fra cui cemento, acqua, aggregato fine (sabbia, per esempio) e aggregato grosso (frammenti di roccia). Il dataset consiste in 1030 osservazioni senza alcun valore mancante. Come variabile di interesse è stata considerata la quantità di aggregato grosso in kg per metro cubo di calcestruzzo.

Per stimare la densità del target è stato utilizzato sia l'algoritmo marginale implementato manualmente, i risultati sono riportati in Figura 19, che il pacchetto *BNPmix*, si vedano i risultati in Figura 20. Per la procedura marginale sono state considerate 100 iterazioni, mentre per la funzione *PYdensity* ne sono state prese in esame 1000. In entrambi i casi, il periodo di burn in è stato posto pari al 10% delle realizzazioni generate.

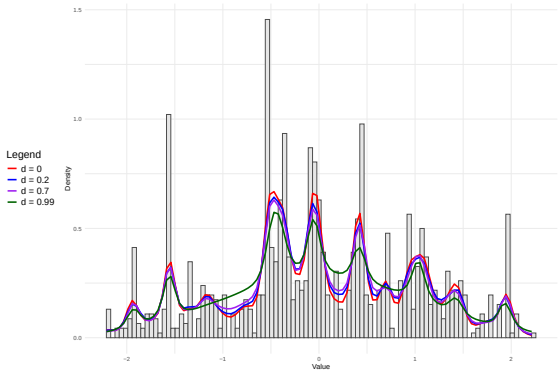

Figure 17: $c = 100$

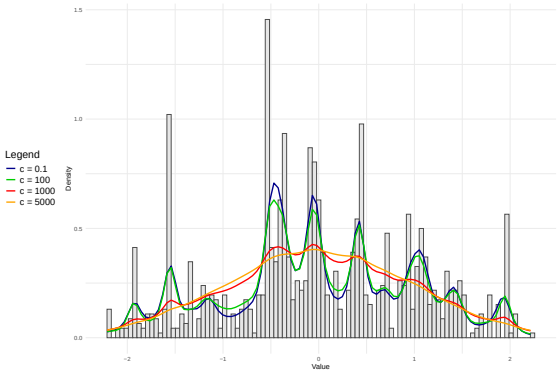

Figure 18: $d = 0.7$

Figure 19: Modelli mistura di PY stimati con l'algoritmo marginale al variare dei parametri

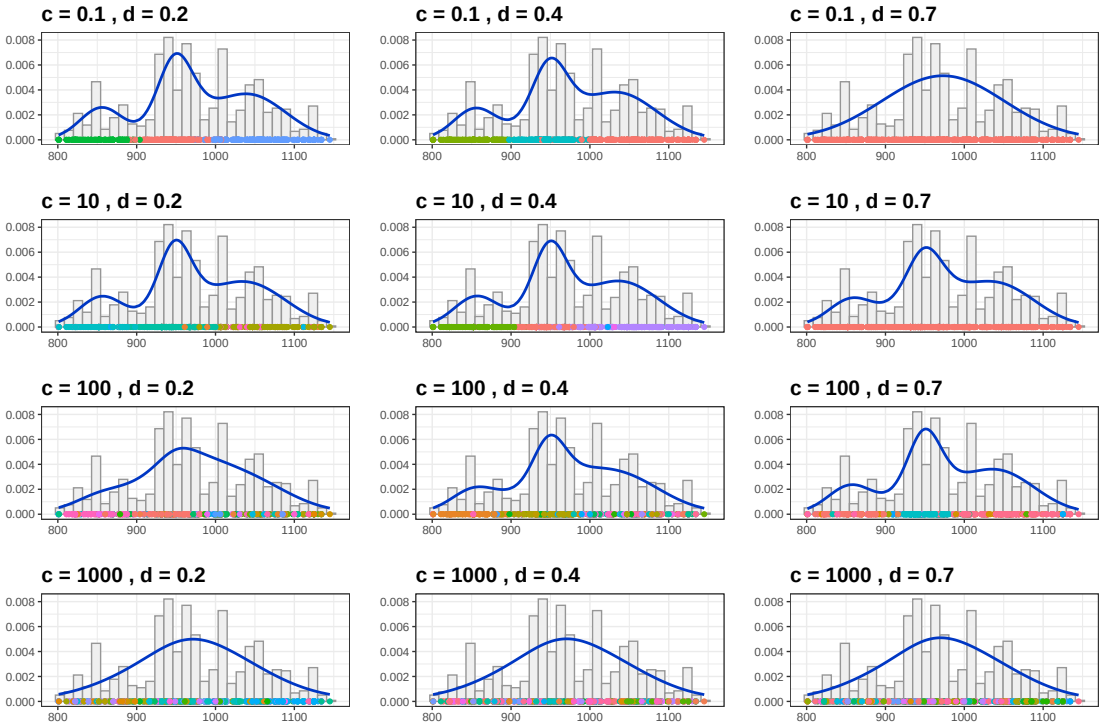

Figure 20: Numero di cluster al variare dei parametri c e d con il pacchetto BNPmix

Nuovamente, come emerge dalle figure, è possibile giungere alle medesime conclusioni fatte in precedenza circa la smoothness delle densità stimate al variare dei parametri c e d. Infatti, sembrerebbe che per valori bassi di concentrazione e penalizzazione i dati siano stati generati da tre misture che simboleggiano l'adozione di tre differenti tipologie di aggregato di calcestruzzo.

Bibliografia

Canale, A., Corradin, R. & Nipoti, B. (2022a). Importance conditional sampling for Pitman-Yor mixtures. *Statistics and Computing* **32(3)**, 40.

Canale, A., Corradin, R. & Nipoti, B. (2022b). Package 'BNPmix'. `https://cran.r-project.org/web/packages/BNPmix/BNPmix.pdf`.

Dassios, A. & Zhang, J. (2023). Exact simulation of Poisson-Dirichlet Distribution and Generalised Gamma process. *Methodology and Computing in Applied Probability* **25**.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I. & Ruggiero, M. (2015). Are Gibbs-type priors the natural generalization of the Dirichlet process? *IEEE Transactions Pattern Analysis and Machine Intelligence* **37(2)**, 212–229.

Ferguson, T. S. (1973). A Bayesian analysis of some Nonparametric problems. *The Annals of Statistics* **1(2)**, 209 – 230.

Ishwaran, H. & James, L. F. (2001). Gibbs Sampling methods for Stick-Breaking priors. *Journal of the American Statistical Association* **96(453)**, 161–173.

Orbanz, P. (2014). Lecture Notes on Bayesian Nonparametrics .

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25(2)**, 855 – 900.

Shanawad, V. (2021). Concrete compressive strength data. `https://www.kaggle.com/datasets/vinayakshanawad/cement-manufacturing-concrete-dataset`.