BML Lecture 4

Bayesian nonparametrics

Julyan ARBEL

What comes to your mind when you hear about

Bayesian nonparametrics ?

Infinite-dim spaces
Dirichlet process
Random process
Clustering

Nonparametric means : ① infinite dimensional

or dimension that grows with n

Bayes : $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(\theta) \cdot \mathbb{P}(X \mid \theta)$

De Finetti theorem

infinite exchangeable data : $p(X_1, \dots, X_k) = p(X_{\sigma(1)}, \dots, X_{\sigma(k)})$

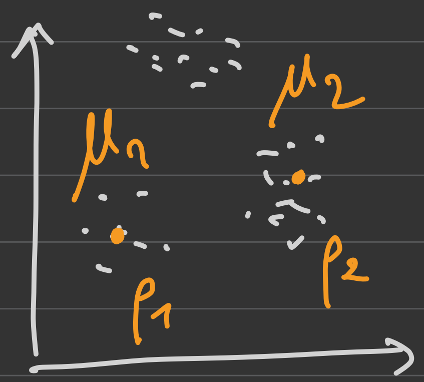
$\forall k \geq 1 \,,\ \forall \sigma \in S^k$

$$\Leftrightarrow \quad p(X_1, \dots, X_k) = \int \prod_{i=1}^{k} p(X_i \mid \theta)\, \underline{\underline{P}}(d\theta)$$

(4)

conditional iid.

Mixture Models, generative model



$n = 1 \dots N$ $\qquad$ $K = \cancel{2}$ clusters

$$\begin{cases} z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2, \dots, \rho_K) & \rho_1 + \rho_2 = 1 \\ & >0 \quad >0 \\ x_n \mid z_n \overset{iid}{\sim} N(\mu_{z_n}, \Sigma) & \end{cases}$$

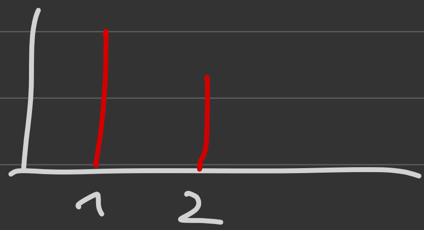
Parameters $(\mu_1, \mu_2, \dots, \mu_K) \; \underline{(\rho_1, \rho_2, \dots, \rho_K)}$

$$\mu_k \sim N(\mu_0, \Sigma_0) \qquad \rho_1 \sim \text{Beta}(a_1, a_2)$$
$$\rho_2 = 1 - \rho_1 .$$

Now, $(\rho_1, \dots, \rho_K) \sim \text{Dirichlet}(a_1, \dots, a_K)$

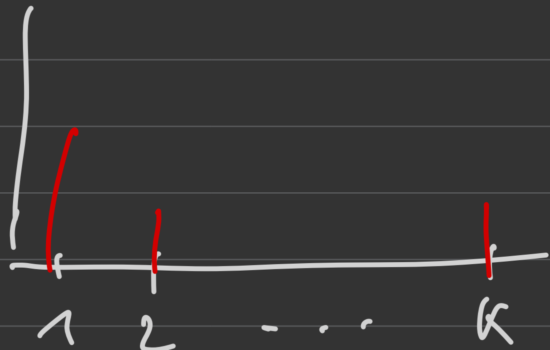
N obs. K clusters. What to do if $K > N$?

Beta



(p_1, p_2)

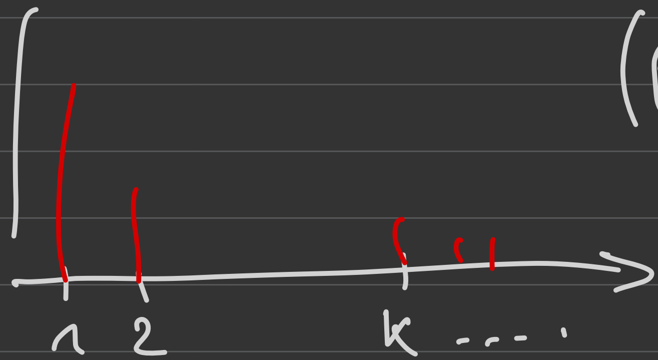
Dirichlet



$(p_1, \dots, p_K)$

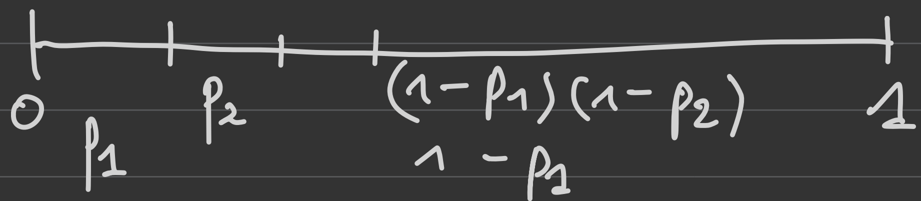
Stick-breaking

$K = \infty$



$(p_1, p_2, \dots) \in \text{Simplex}(\infty)$

$\sum p_i = 1, \quad p_i \geq 0$



$(p_1, p_2, \dots) \sim \text{GEM}.$

Chapt 1 Introduction

Chapter 2 Dirichlet process

1. Definition

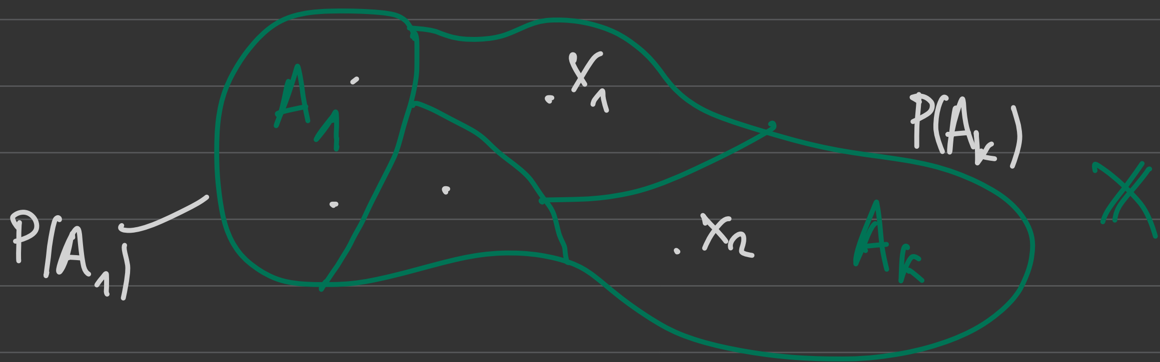
Ferguson, 1973

Space $\mathbb{X}$, P is a Dirichlet process (DP) on $\mathbb{X}$

if $\exists \Big| \begin{array}{l} \alpha > 0 \text{ concentration} \\ P_0 \text{ (fixed) proba measure : base measure} \end{array}$

$\forall k$, $\forall$ partition $(A_1, \dots, A_k)$ of $\mathbb{X}$ measurable

$$\big(P(A_1), \dots, P(A_k)\big) \sim \mathrm{Dir}\big(\alpha P_0(A_1), \dots, \alpha P_0(A_k)\big).$$



$P \sim DP(\alpha, P_0)$

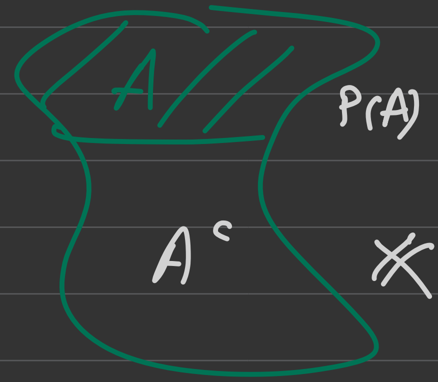
↳ is a proba distri. on $\mathbb{X}$

$X_1, \dots, X_n \mid P \sim P$.

$\in \mathbb{X}$

Moments

Let $A \subset \mathcal{X}$, $B \subset$. If $P \sim DP(\alpha, P_0)$, then



$$\mathbb{E}[P(A)] = \frac{\alpha P_0(A)}{\alpha} = P_0(A)$$

$$(P(A), P(A^c)) \sim \mathrm{Dir}(\underbrace{\alpha P_0(A)}_{a}, \alpha \underbrace{P_0(A^c)}_{1-P_0(A)})$$

$$P(A) \sim \mathrm{Beta}(a, b)$$

$$\mathbb{E}\,\mathrm{Beta}(a,b) = \frac{a}{a+b}$$

$$\mathbb{V}[P(A)] = \frac{P_0(A)(1-P_0(A))}{1+\alpha} \quad \text{concentration.}$$

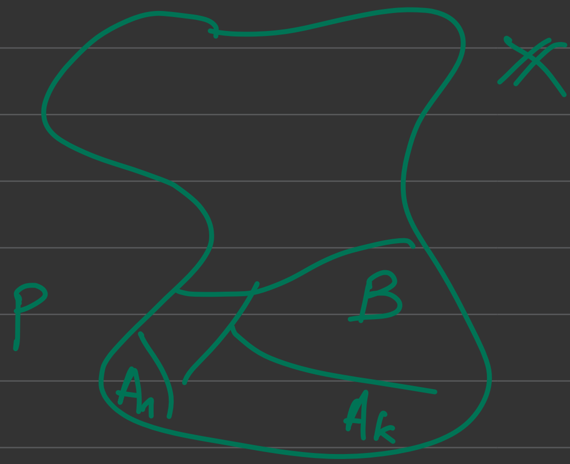
$$\mathbb{V} \;\_\_\_\_ = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\mathrm{Cov}(P(A), P(B)) = \overset{\text{EXO}}{\cdots} = \frac{P_0(A \cap B) - P_0(A)P_0(B)}{1+\alpha}$$

esp. $A \cap B = \emptyset \implies \mathrm{Cov} \leq 0$

$\mathbb{E}[P(A)] = P_0(A)$ can be written $\int P(A)\, dDP(P) = P_0(A)$

$$\begin{cases} P \sim DP(\alpha, P_0) \\ X \mid P \sim P \end{cases} \Longrightarrow X \sim P_0$$



<u>Self-similarity</u> of the DP

$$P_{|B}(A) = P(A \cap B)$$

$$P_B(A) = P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

<u>Theorem</u>: <u>if $P \sim DP(\alpha, P_0)$ on $\mathcal{X}$, then $P_B \sim DP(\alpha\, P_{0|B})$</u>.

$$\Big(\underbrace{P(A_1), \dots P(A_k)}, P(B^c)\Big) \sim \mathrm{Dir}_{k+1}$$

$$\left(\frac{P(A_1)}{P(B)}, \dots, \frac{P(A_k)}{P(B)}\right) \sim \mathrm{Dir}_k\Big(\underbrace{\alpha P_0(A_1)}_{\alpha P_{0|B}(A_1)}, \dots, \underbrace{\alpha P_0(A_k)}_{\alpha P_{0|B}(A_k)}\Big)$$

<u>Posterior conjugacy</u> : $X_1, \dots, X_n$ sample from DP :

$G = \alpha P_0$

$$\begin{cases} P \sim DP(\alpha, P_0) \\ X_1, \dots, X_n \mid P \sim P \end{cases} \quad (*) \qquad DP(\alpha, P_0) \longleftrightarrow DP(\alpha P_0)$$

$\alpha = G(\mathbb{X})$ $\qquad P_0 = \dfrac{G}{\alpha}$

<u>Th</u> (Ferguson) : the posterior in $(*)$ is

$$P \mid (X_1, \dots, X_n) \sim DP\Big(\underbrace{\alpha P_0 + \sum_{i=1}^{n} \delta_{X_i}} \Big)$$

Predictive distribution :

$$\mathbb{P}(X_{n+1} \mid X_1, \dots, X_n) = \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i} = P_n$$

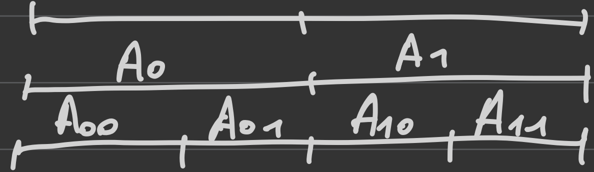
Updated parameter $G = \alpha P_0 + \sum_{i=1}^{n} \delta_{X_i}$ $\quad$ Dirac masses at X_i

— concentration $\alpha_n = G(\mathbb{X}) = \alpha + n$

— base measure $P_n = \dfrac{G}{\alpha + n}$

proof : tail-free property

$\mathcal{X} = [0,1]$



$$\boxed{\begin{array}{l} P \sim DP(\alpha, P_0) \text{ on } \mathcal{X} \\ \text{then } \{P(A_0), P(A_1)\} \perp\!\!\!\perp \{P(A_{00}), P(A_{01}), P(A_{10}), P(A_{11})\}. \end{array}}$$

Consider partition $(A_1, \dots, A_k)$ of $\mathcal{X}$.

$$N_j = \#(i : X_i \in A_j)$$

tail-free property $\Rightarrow$ $\big(P(A_1), \dots P(A_k)\big) \mid X_1 \dots X_n \overset{d}{=} \big(P(A_1), \dots P(A_k)\big) \mid N_1, \dots, N_k$

$$\begin{cases} \text{prior } \big(P(A_1), \dots P(A_k)\big) \sim Dir_k\big(\alpha P_0(A_1), \dots, \alpha P_0(A_k)\big) \\ \text{model } (N_1, \dots, N_k) \sim \text{Multinomial}\big(n, (P(A_1), \dots, P(A_k))\big) \end{cases}$$

posterior $\big(P(A_1), \dots, P(A_k)\big) \mid (N_1, \dots, N_k) \sim Dir_k\big(\alpha P_0(A_1) + N_1, \dots, \alpha P_0(A_k) + N_k\big)$

$$\Rightarrow \quad P \mid N_1, \dots, N_k = P \mid X_1, \dots, X_n \sim DP\Big(\alpha P_0 + \sum_{i=1}^{n} \delta_{X_i}\Big)$$

Predictive

Fundamentals of BNP inference

Ghosal van der Vaart

Chapter 4 on DP.

$$\mathbb{P}(X_{n+1} \mid X_1,\dots,X_n) = \underbrace{\frac{\alpha}{\alpha+n} P_0}_{\text{prior}} + \underbrace{\frac{n}{\alpha+n} \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}}_{\text{data}}$$

Marginalization of DP : Pólya-Urn

Blackwell - MacQueen.

$$\begin{cases} P \sim DP \\ X_1 \mid P \sim P \end{cases} \Longrightarrow X_1 \sim P_0$$

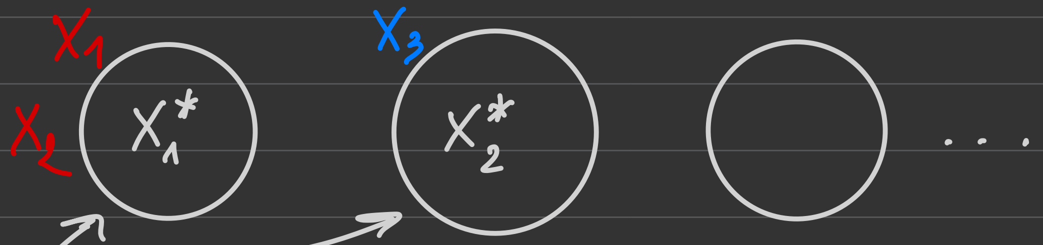
$$\begin{cases} P \sim DP \\ X_1, X_2 \mid P \sim P \end{cases} \Longrightarrow X_1 \sim P_0 , \quad X_2 \mid X_1 \sim \frac{\alpha}{\alpha+1} P_0 + \frac{1}{\alpha+1} \delta_{X_1}$$

$\vdots$

$$X_{n+1} \mid X_1 \dots X_n \sim \frac{\alpha}{\alpha+n} P_0 + \frac{1}{\alpha+n} \sum_{i=1}^{n} \delta_{X_i}$$

α uncolored balls

Chinese Restaurant process : clustering distribution



rule : each new customer gets seated to tables with probability to

$$\begin{cases} \# \text{ of seated customers} \\ \alpha \text{ for new table.} \end{cases}$$

x_j^* unique observations (tables)

n_j : # customers on table x_j^*

$K = \#$ (populated) tables

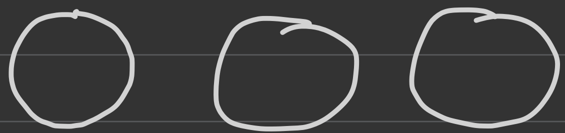
$$\mathbb{P}(X_{n+1} \mid X_1 \cdots X_n) = \frac{\alpha}{\alpha+n} P_0 + \frac{n}{\alpha+n} \frac{1}{n} \sum_{j=1}^{k} n_j \, \delta_{x_j^*}$$

$$1, \ \frac{1}{\alpha+1}$$

Theorem CRP $\longleftrightarrow$ $X_1,\dots,X_n \mid P \sim DP$

$$p(m_1,\dots,m_k) = \frac{\alpha^k}{(\alpha)_n}\prod_{j=1}^{k}(m_j-1)! = (*)$$

$$= \alpha(\alpha+1)\cdots(\alpha+n-1) \quad \text{ascending factorial}$$



$$(*) = \overset{X_1}{\frac{\cancel{\alpha}}{\cancel{\alpha}}}\;\overset{X_2}{\frac{1}{\alpha+1}}\;\overset{X_3}{\frac{2}{\alpha+2}}\cdots\overset{X_{n_1}}{\frac{m_1-1}{\alpha+n_1-1}}\;\overset{X_{n_1+1}}{\frac{\alpha}{\alpha+m_1}}\;\frac{1}{\alpha+m_1+1}\cdots\overset{X_{n_1+n_2}}{\frac{m_2-1}{\alpha+m_1+m_2-1}}\cdots$$

$$(*) = \frac{\alpha^K}{(\alpha)_n}\prod_{j=1}^{k}(m_j-1)!\,.$$

Ewens sampling formula $p(m_1,\dots,m_\ell)$

m_ℓ : # tables with ℓ customers. $\quad i=1,\dots,n$

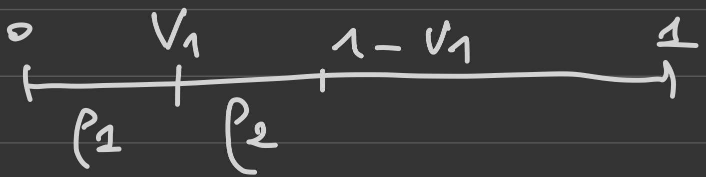
$$\sum_{\ell=1}^{n} m_\ell = K \qquad \sum_{\ell=1}^{n} \ell\, m_\ell = n$$

$$p(m_1, \dots, m_\ell) = \frac{n!}{\alpha_{(n)}} \alpha^k \frac{1}{\prod_{\ell=1}^{n} \ell^{m_\ell} m_\ell!}$$

<u>Stick-breaking</u> for DP:

Let $V_i \overset{iid}{\sim} \text{Beta}(1, \alpha)$, $\theta_i \overset{iid}{\sim} P_0$.

Let $\rho_1 = V_1$, $\rho_i = V_i \prod_{\ell=1}^{i-1} (1 - V_\ell)$



if $P = \sum_{i=1}^{\infty} \rho_i \delta_{\theta_i}$

then $P \sim DP(\alpha, P_0)$.