
Bayésien NP

Ibrahim SEYDI

Table des matières

1	Bayésien non paramétrique	2
1.1	Processus de Dirichlet	2
1.2	Modèles de mélanges à processus de Dirichlet (DPMM)	8
1.3	Échantillonnage de Gibbs avec lois conjuguées pour les DPMM	10
	Bibliographie	11

1.1 Processus de Dirichlet

1.1.1 Loi de Dirichlet

La **loi de Dirichlet** est une loi de probabilité définie sur le simplexe de dimension $K - 1$:

$$\Delta^{K-1} = \left\{ \mathbf{p} = (p_1, \dots, p_K) \in \mathbb{R}^K \left| \forall k, p_k \geq 0, \sum_{k=1}^K p_k = 1 \right. \right\}$$

C'est une généralisation multidimensionnelle de la loi Beta. On note $\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})$ si \mathbf{p} suit une loi de Dirichlet de paramètre $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K) \in \mathbb{R}_+^K$, avec $\sum_{k=1}^K \alpha_k > 0$, dont la densité est donnée par :

$$f(p_1, \dots, p_K \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K p_k^{\alpha_k - 1}, \quad \text{pour } \mathbf{p} \in \Delta^{K-1}$$

où la constante de normalisation $B(\boldsymbol{\alpha})$ est :

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}$$

et $\Gamma(\cdot)$ est la fonction Gamma.

L'espérance de chaque composante est :

$$\mathbb{E}[p_k] = \frac{\alpha_k}{\sum_{j=1}^K \alpha_j}$$

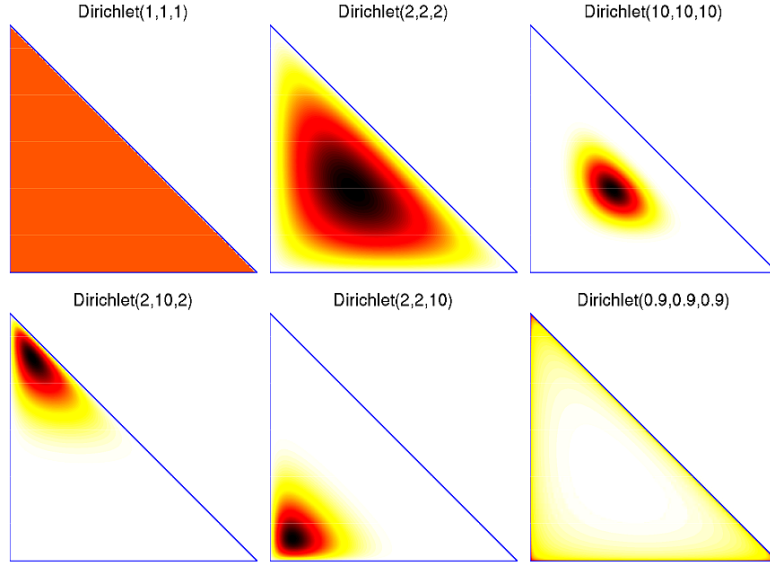


FIGURE 1.1 – Exemples de lois de Dirichlet sur $\mathbf{p} = (p_1, p_2, p_3)$ représentées en 2D ($p_3 = 1 - p_1 - p_2$) [tirée de (Ghahramani, 2005)]

Conjugaison de la loi de Dirichlet

Soient $\mathbf{p} \sim \text{Dir}(\alpha_1, \dots, \alpha_K)$ et $\mathbf{c} = (c_1, \dots, c_K) \sim \mathcal{M}(n, \mathbf{p})$ (c_k est le nombre d'observations de la catégorie k parmi n tirages). Alors la loi a posteriori est donnée par :

$$\mathbf{p} \mid \mathbf{c} \sim \text{Dir}(\alpha_1 + c_1, \dots, \alpha_K + c_K)$$

La loi de Dirichlet est conjuguée à la multinomiale.

Lien Dirichlet/Gamma

On peut générer une variable de la loi de Dirichlet en normalisant des variables de lois Gamma indépendantes. Soit K variables indépendantes $\lambda_k \sim \text{Gamma}(\alpha_k, 1)$, pour $k = 1, \dots, K$. Alors :

$$\left(\frac{\lambda_1}{\sum_k \lambda_k}, \dots, \frac{\lambda_K}{\sum_k \lambda_k} \right) \sim \text{Dir}(\alpha_1, \dots, \alpha_K)$$

Quelsues propriétés intéressantes

Soient $\mathbf{p} \sim \text{Dir}(\alpha_1, \dots, \alpha_K)$ et $\boldsymbol{\theta} \sim \text{Dir}(\alpha_1 b_1, \dots, \alpha_1 b_N)$ avec $\sum_i b_i = 1$, alors :

- $(p_1 + p_2, p_3, \dots, p_K) \sim \text{Dir}(\alpha_1 + \alpha_2, \alpha_3, \dots, \alpha_K)$
- $(\theta_1 p_1, \dots, \theta_N p_1, p_2, \dots, p_K) \sim \text{Dir}(\alpha_1 b_1, \dots, \alpha_1 b_N, \alpha_2, \dots, \alpha_K)$
- $\left(\frac{p_2}{\sum_{k=2}^K p_k}, \dots, \frac{p_K}{\sum_{k=2}^K p_k} \right) \sim \text{Dir}(\alpha_2, \dots, \alpha_K)$

1.1.2 Processus de Dirichlet

Le processus de Dirichlet est un objet fondamental en statistique bayésienne non paramétrique. Il permet de définir une distribution de probabilité sur l'espace des mesures de probabilité sur un espace mesurable donné.

Definition 1 (Ferguson (1973)) Soit $(\Omega, \mathcal{B}(\Omega))$ un espace mesurable. Soit G_0 une mesure de probabilité sur $(\Omega, \mathcal{B}(\Omega))$ et $\alpha > 0$. On dit que G suit un **processus de Dirichlet** de base (α, G_0) , et on écrit : $G \sim \text{DP}(\alpha, G_0)$, si pour toute partition finie (A_1, \dots, A_K) de Ω , on a :

$$(G(A_1), \dots, G(A_K)) \sim \text{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_K))$$

Proposition 1 Pour tout $A \in \mathcal{B}(\Omega)$, on a :

$$\mathbb{E}[G(A)] = G_0(A), \quad \text{Var}[G(A)] = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}$$

α est appelé **paramètre de concentration** : plus α est grand, plus G est concentrée autour de G_0 .

Propriété de conjugaison du DP

Soit $G \sim \text{DP}(G_0, \alpha)$ et soit $\theta \sim G$. Alors la loi a posteriori de G sachant θ est également un processus de Dirichlet :

$$\mathbb{P}(G \mid \theta) = \frac{\mathbb{P}(\theta \mid G) \mathbb{P}(G)}{\mathbb{P}(\theta)} = \text{DP} \left(\alpha + 1, \frac{\alpha}{\alpha + 1} G_0 + \frac{1}{\alpha + 1} \delta_{\theta_1} \right)$$

Soit un échantillon $(\theta_1, \dots, \theta_n)$ tel que : $\theta_i \mid G \sim G$ indépendants pour $i = 1, \dots, n$. Alors :

$$\mathbb{P}(G \mid \theta_1, \dots, \theta_n) = \text{DP} \left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i} \right)$$

Cette propriété (prouvée par Ferguson (1973)) est analogue à la conjugaison de la loi de Dirichlet à la loi multinomiale. Par ailleurs, on peut voir ici que le paramètre de concentration α peut être considéré comme exprimant la confiance dans la mesure de base G_0 .

Représentations du DP

Il existe plusieurs manières équivalentes de comprendre et représenter le processus de Dirichlet.

I - Représentation par urn de Polya (Blackwell & MacQueen, 1973)

Blackwell et MacQueen (1973) proposent une vision alternative et très intuitive du processus de Dirichlet. Leur approche repose non pas sur la définition formelle via des distributions de Dirichlet, mais sur l'étude des **lois prédictives** successives d'une suite de variables aléatoires. On considère une suite d'observations $\theta_1, \theta_2, \dots$ tirées d'une loi inconnue G , et on suppose que cette loi G elle-même est tirée selon un processus de Dirichlet :

$$\theta_1, \theta_2, \dots \sim G \quad \text{et} \quad G \sim \text{DP}(\alpha, G_0).$$

Le premier tirage θ_1 est effectué selon la **loi prédictive a priori**, c'est-à-dire la mesure de base G_0 : $\theta_1 \sim G_0$. Après avoir observé θ_1 , on met à jour notre connaissance de G (qui est encore un processus de Dirichlet, cf (rajouter conjugaison de DP)) :

$$G \mid \theta_1 \sim \text{DP}(\alpha + 1, \frac{\alpha G_0 + \delta_{\theta_1}}{\alpha + 1})$$

Plutôt que de recalculer cette distribution, on s'intéresse à la **loi prédictive** de θ_2 :

$$\theta_2 \mid \theta_1, \alpha, G_0 \sim \frac{\alpha}{\alpha + 1} G_0 + \frac{1}{\alpha + 1} \delta_{\theta_1}$$

- Avec probabilité $\frac{\alpha}{\alpha+1}$, on tire une nouvelle valeur selon G_0 (innovation)
- Avec probabilité $\frac{1}{\alpha+1}$, on tire à partir de θ_1 (renforcement, propriété du *rich get richer*)

Par récurrence, on obtient la **loi prédictive** suivante :

$$\theta_{n+1} \mid (\theta_1, \dots, \theta_n), \alpha, G_0 \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i}$$

Cette loi est une combinaison (convexe) de la mesure de base G_0 , qui introduit de nouvelles valeurs et de l'empirisme des observations passées, favorisant la répétition de valeurs déjà vues.

Ainsi, plus une valeur θ_i a été observée fréquemment, plus la probabilité qu'elle soit tirée à nouveau est élevée. Inversement, la probabilité d'observer une nouvelle valeur diminue avec n , selon $\frac{\alpha}{\alpha+n}$. Ce mécanisme est à l'origine du phénomène de **clustering**, caractérisant le DP. De plus, la décroissance de cette probabilité vers 0 lorsque $n \rightarrow \infty$ illustre que toutes les valeurs finiront par être répétées, soulignant ainsi le caractère discret du DP, et ceci en dépit de la continuité ou non de la mesure de base.

II - Représentation par processus du restaurant chinois (CRP)

Le processus du restaurant chinois (Chinese Restaurant Process, CRP) est une construction probabiliste qui permet de représenter la loi sur les partitions formées par un DP. Il repose sur l'approche suivante :

"Un restaurant possède une infinité de tables. Les clients (observations) arrivent un par un et choisissent une table selon une certaine règle de proba. Une fois tous les clients assis, les tables forment des clusters, c'est-à-dire une partition des observations."

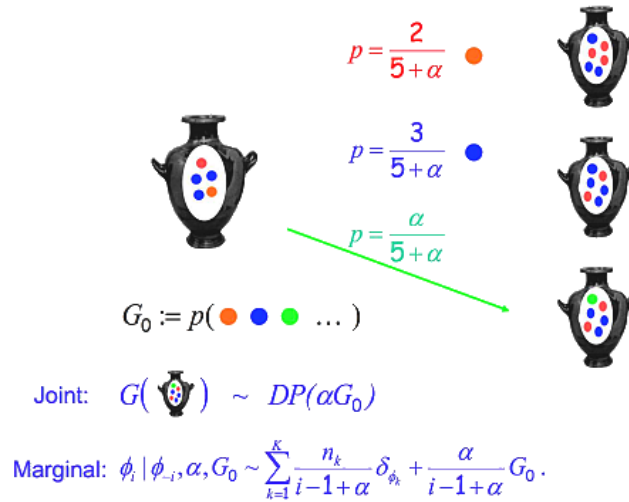


FIGURE 1.2 – Enter Caption [tirée de (MacEachern, 2016)]

Soit $\alpha > 0$ un paramètre de concentration du processus. À chaque étape n , le client n choisit selon la règle suivante :

- une table k déjà occupée avec proba :

$$\mathbb{P}(\text{client } n \text{ rejoint la table } k) = \frac{n_k}{\alpha + n - 1}$$

où n_k est le nombre de clients déjà assis à la table k ;

- une **nouvelle table** avec proba :

$$\mathbb{P}(\text{client } n \text{ crée une nouvelle table}) = \frac{\alpha}{\alpha + n - 1}$$

Ce mécanisme reflète la propriété *rich get richer* : plus une table est populaire, plus elle a de chances d'attirer de nouveaux clients.

Le DP $G \sim DP(\alpha, G_0)$ est une loi sur des lois, c'est un objet discret. En échantillonnant n valeurs à partir de G , on observe souvent des valeurs répétées, formant ainsi une partition des observations. Le CRP décrit cette partition, sans expliciter les valeurs. Il est donc équivalent au DP du point de vue de la structure de clustering. Le CRP a la particularité de former une loi **échangeable** sur les partitions :

$$\mathbb{P}(\theta_1, \dots, \theta_n) = \mathbb{P}(\theta_{\sigma(1)}, \dots, \theta_{\sigma(n)})$$

pour toute permutation σ de $\{1, \dots, n\}$. Il s'agit d'une conséquence du théorème de De Finetti.

Il est possible d'écrire la loi du CRP. Soit $\pi_{[n]}$ une partition de l'ensemble $\{1, \dots, n\}$ en K clusters $\{c_1, \dots, c_K\}$. La probabilité d'obtenir cette partition sous un $CRP(\alpha, n)$ est :

$$\mathbb{P}(\pi_{[n]}) = \frac{\alpha^K \prod_{k=1}^K (|c_k| - 1)!}{\prod_{j=1}^n (\alpha + j - 1)}$$

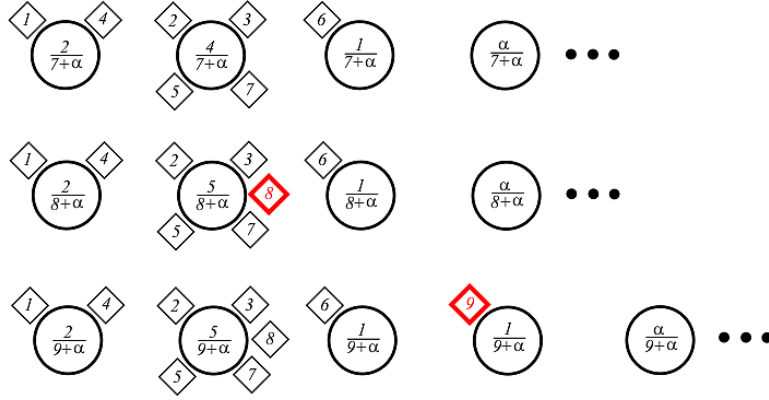


FIGURE 1.3 – Enter Caption [tirée de (Sudderth, 2006)]

Cette expression montre que la probabilité dépend uniquement des tailles des clusters (et non de l'ordre d'arrivée), ce qui confirme l'échangeabilité.

Une petite remarque qui vaut son détour est lorsque $n \rightarrow \infty$, le nombre de clusters K_n suit asymptotiquement :

$$K_n \sim \alpha \log n$$

Ce comportement est intéressant parce qu'il montre que la complexité du modèle augmente avec la taille des données, ce qui renvoie au fait que le CRP (et donc le DP) est bien un outil non paramétrique.

III - Représentation par Stick Breacking (SB)

Le processus de Stick Breacking introduit par Sethuraman (1994), fournit une construction pratique et intuitive du DP, en mettant en évidence sa structure **componants-wise**. Le processus de Dirichlett $G \sim \text{DP}(\alpha, G_0)$ peut s'écrire comme :

$$G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot)$$

où :

- $\theta_k \sim G_0$ iid sont les positions des composantes
- $\pi_k \in [0, 1]$ sont les poids associés à chaque composante, avec $\sum_k \pi_k = 1$ presque sûrement
- δ_{θ_k} est la mesure de Dirac en θ_k

Les poids π_k sont générés à partir d'une suite de variables indépendantes $V_k \sim \text{Beta}(1, \alpha)$ par :

$$\pi_1 = V_1, \quad \pi_k = V_k \prod_{j=1}^{k-1} (1 - V_j), \quad \text{pour } k \geq 2$$

Ce mécanisme est ce qu'on appelle stick-breaking parce que conceptuellement on commence avec un bâton de longueur 1, et à chaque étape, on casse une partie V_k du bâton, cette dernière est

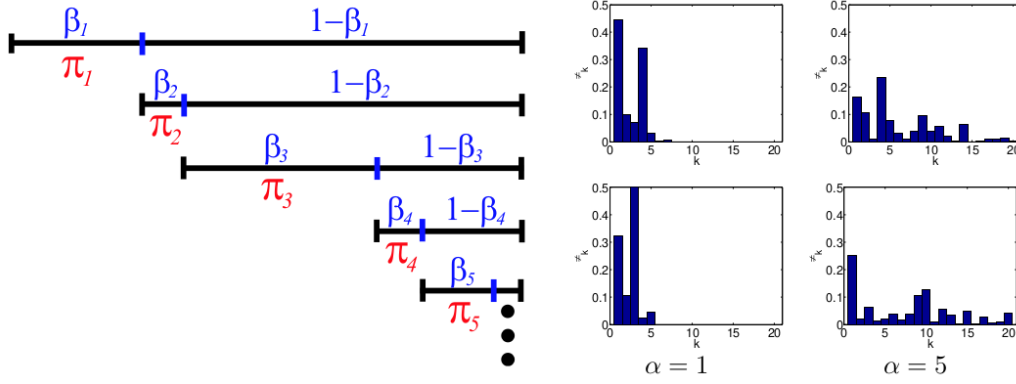


FIGURE 1.4 – Enter Caption [tirée de (Sudderth, 2006)]

assignée à π_k , puis on répète le processus avec le reste du bâton.

Quelques propriétés sont intéressants à relever :

1. **Validité** : La somme des poids converge presque sûrement vers 1 :

$$\sum_{k=1}^{\infty} \pi_k = 1 \quad \text{p.s.}$$

2. **Espérance** :

$$\mathbb{E}[V_k] = \frac{1}{1 + \alpha} \quad \Rightarrow \quad \mathbb{E}[\pi_1] = \frac{1}{1 + \alpha}$$

3. **Comportement selon α** :

- Si $\alpha \ll 1$, la majorité de la masse est concentrée dans les premiers termes
- Si $\alpha \rightarrow \infty$, G devient proche de G_0 , avec des poids petits et plus uniformes.

Des variantes du SB existent en modifiant la loi des variables V_k :

- $V_k \sim \text{Beta}(1, \alpha) \Rightarrow \text{DP}$
- $V_k \sim \text{Beta}(a, b) \Rightarrow \text{Beta Two-parameter Process}$
- $V_k \sim \text{Beta}(1 - a, b + ka) \Rightarrow \text{Pitman-Yor Process (PYP)}$

Le SB est largement utilisée car il permet d'approximer un DP par une somme finie en s'arrêtant à K termes.

1.2 Modèles de mélanges à processus de Dirichlet (DPMM)

On considère le modèle suivant. Soient des données observées x_1, \dots, x_n induites par un paramètre latent θ_i , tiré d'une loi G une réalisation d'un DP :

$$\begin{aligned} x_i &\sim f(x_i \mid \theta_i) \\ \theta_i &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned}$$

où :

- $f(\cdot \mid \theta_i)$ est la loi conditionnelle (ou un noyau)
- G_0 est la mesure de base du DP, encodant l'a priori sur les paramètres θ
- α est paramètre de concentration, donnant le degré de dispersion de G autour de G_0

Une réalisation d'un processus de Dirichlet G est presque sûrement discret (**cf citer exemples dans les représentatitons**), ce qui pose problème pour modéliser une loi continue. Pour remédier à cela, on convolue G avec une densité $f(\cdot \mid \theta)$:

$$x_i \sim \int f(x \mid \theta) G(d\theta)$$

Ceci introduit une sorte de régularité entre observations. Ce modèle permet de générer des densités continues et est appelé modèle de mélanges à processus de Dirichlet (DPMM). On l'écrit :

$$\begin{aligned} x_i &\sim F(x_i \mid \theta_i) \\ F(\cdot \mid \theta_i) &= \int f(\cdot \mid \theta_i) G(d\theta_i) = \sum_{k=1}^{\infty} \pi_k f(\cdot \mid \theta_i) \\ \theta_i &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned}$$

Chaque observation x_i peut être associée à une classe latente c_i , qui indique à quel cluster elle appartient. Le modèle induit une loi sur les partitions (**cf CRP**) des données. Une réécriture équivalente de notre modèle, en utilisant la représentation par Stick-Breaking est la suivante :

$$\begin{aligned} x_i \mid \phi, c_i &\sim F(x_i \mid \phi_{c_i}), \quad F(\cdot \mid \phi_{c_i}) = \sum_{k=1}^{\infty} \pi_k f(\cdot \mid \phi_{c_k}) \\ c_i \mid \boldsymbol{\pi} &\sim \mathcal{M}(K, \pi_1, \dots, \pi_K) \\ \phi_c &\sim G_0 \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) = \beta_k \left(1 - \sum_{l=1}^{k-1} \pi_l \right) \\ \beta_k &\sim \text{Beta}(1, \alpha) \end{aligned}$$

Cette réécriture illustre de manière explicite le fait que le modèle DPM peut être vu comme un mélange infini et donne une marche à suivre pour l'inférence du modèle.

Extensions : Processus de Dirichlet Hiérarchiques (HDP)

Problème : Dans certaines situations, on a des groupes d'observations, et on veut modéliser des clusters partagés entre groupes, tout en permettant des variations locales.

Solution : Utiliser un processus de Dirichlet hiérarchique (HDP) :

$$\begin{aligned}x_{ji} &\sim F(x \mid \theta_{ji}) \\ \theta_{ji} &\sim G_j \\ G_j &\sim \text{DP}(\alpha, G_0) \\ G_0 &\sim \text{DP}(\gamma, H)\end{aligned}$$

Cela permet de partager les composantes θ_k entre différents groupes tout en modélisant des proportions spécifiques à chaque groupe j .

1.3 Échantillonnage de Gibbs avec lois conjuguées pour les DPMM

L'inférence exacte dans les DPMM est en général compliquée lorsque le nombre d'observations est modéré à grand. En pratique, l'estimation des quantités d'intérêt, telles que la distribution prédictive postérieure, est effectuée via des méthodes Monte Carlo, et plus particulièrement par échantillonnage de Gibbs lorsque des lois conjuguées sont utilisées. On se place dans le cadre d'un DPMM suivant :

$$\begin{aligned}x_i &\sim F(x_i \mid \theta_i) \\ F(\cdot \mid \theta_i) &= \sum_{k=1}^{\infty} \pi_k f(\cdot \mid \theta_k) \\ \theta_i &\sim G \\ G &\sim \text{DP}(\alpha, G_0)\end{aligned}$$

Ou réécrit de manière équivalente :

$$\begin{aligned}x_i \mid \phi, c_i &\sim F(x_i \mid \phi_{c_i}), \quad F(\cdot \mid \phi_{c_i}) = \sum_{k=1}^{\infty} \pi_k f(\cdot \mid \phi_{c_k}) \\ c_i \mid \boldsymbol{\pi} &\sim \mathcal{M}(K, \pi_1, \dots, \pi_K) \\ \phi_c &\sim G_0 \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) = \beta_k \left(1 - \sum_{l=1}^{k-1} \pi_l\right) \\ \beta_k &\sim \text{Beta}(1, \alpha)\end{aligned}$$

On pourrait chercher à :

1. Échantillonner le postérieur des paramètres latents (les θ_i ou les clusters c_i)
2. Estimer la densité prédictive $F(x_{n+1} \mid x_{1:n})$

Concernant ce second point, l'estimation se fait en pratique, en moyennant sur plusieurs échantillons MCMC. Soient $\theta = (\theta_1, \dots, \theta_n)$ les paramètres latents des données observées et $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_n^{(t)})$ un échantillon du postérieur obtenu à la t -ième itération d'une chaîne de Markov (pour $t = 1, \dots, T$). La densité prédictive théorique est donnée par :

$$F(x_{n+1} \mid x_{1:n}) = \int F(x_{n+1} \mid \theta_{n+1}) \mathbb{P}(\theta_{n+1} \mid x_{1:n}) d(\theta_{n+1})$$

Mais $\mathbb{P}(\theta \mid x_{1:n})$ est compliqué, donc on l'approxime avec des tirages postérieurs $\theta^{(t)}$. Grâce à la propriété ([citer plus haut](#)), pour un $\theta^{(t)}$, on a :

$$\theta_{n+1}^{(t)} \sim \frac{1}{n + \alpha} \sum_{i=1}^n \delta_{\theta_i^{(t)}} + \frac{\alpha}{n + \alpha} G_0$$

Bibliographie

- Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2).
- Ghahramani, Z. (2005). Non-parametric bayesian methods. UAI '05 Tutorial.
- Jara, A. (2017). Theory and computations for the Dirichlet process and related models : An overview. *International Journal of Approximate Reasoning*, 81 :128–146.
- MacEachern, S. N. (2016). Nonparametric Bayesian methods : a gentle introduction and overview. *Communications for Statistical Applications and Methods*, 23(6) :445–466.
- Ross, G. J. and Markwick, D. (2019). dirichletprocess : An R package for fitting complex bayesian nonparametric models.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4 :639–650.
- Sudderth, E. B. (2006). *Graphical Models for Visual Object Recognition and Tracking*. Ph.d. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.
- Xing, E. (2020a). Lecture 23 : Bayesian nonparametrics : Dirichlet processes.
- Xing, E. (2020b). Lecture 23 dirichlet processes. YouTube.