Introduction aux partitions aléatoires

Adrien de Larrard,
encadré par Jim Pitman et Jean Bertoin

Résumé

Une partition aléatoire est une variable aléatoire dans l'ensemble des partitions d'un ensemble fini ou dénombrable. En probabilités, la classe des partitions aléatoires la plus étudiée est celle des paritions échangeables, invariantes sous l'action des permutations. Dans ce mémoire, nous allons définir les principales notions utiles à l'étude des partitions aléatoires échangeables de $\mathbb{N}$ ainsi que le théorème de Kingman. Enuite, nous allons donner différentes constructions de la partition de Poisson Dirichlet (α, θ) : une construction basée sur les subordinateurs et une autre, en utilisant le Chinese Restaurant Process. Enfin, nous allons voir de quelles façons les partitions aléatoires s'appliquent en statistiques et quelles sont les généralisations du Chinese Restaurant Process.

1 Définitions et premières constructions

1.1 Définitions

Commençons par quelques définitions (toutes ces notions sont traitées en détail dans le cours de Saint Flour de Jim Pitman [5]) :

Partitions d'un ensemble fini ou dénombrable Une partition

$$\pi = \{A_1, ..., A_k\}$$

de $\{1, ..., n\} := [n]$ est une suite de blocs $A_i \subset [n]$ deux à deux disjoints dont la réunion est $[n]$. On note $\mathcal{P}_{[n]}$ l'ensemble des partitions de $[n]$ et $\mathcal{P}_{[n]}^k$

l'ensemble des partitions de $[n]$ en k éléments. Pour chaque partition π de $[n]$ nous définissons une relation d'équivalence sur les entiers dans $\{1, .., n\}$ par la relation suivante : $i \sim^{\pi} j$ si et seulement si i et j sont dans le même ensemble A_k de la partition.

Plus généralement si B est une partie (finie ou dénombrable) de $\mathbb{N}$, alors une partition $\{A_1, A_2...\}$ de B est une suite de blocs $A_i \subset B$ toujours deux à deux disjoints dont la réunion est B. On notera $\mathcal{P}_B$ l'ensemble des partitions de B.

Partition aléatoire Une partition aléatoire Π_n de $[n]$ est une variable aléatoire sur un espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathcal{P}_{[n]}$. Dans toute la suite on supposera cet espace de probabilités fixé. On dit qu'une partition aléatoire est échangeable s'il existe une fonction symétrique p telle que pour toute partition particulière $\{A_1, ..., A_k\} \in \mathcal{P}_{[n]}$,

$$\mathbb{P}(\Pi_n = \{A_1, ..., A_k\}) = p_n(|A_i|, ..., |A_k|), \tag{1}$$

où pour tout j, $|A_j|$ est le cardinal de A_j et p vérifie la relation : pour toute permutation $\sigma \in \mathcal{S}_k$ et tous $n_1, ..., n_k$ tel que $\sum_{i=1}^{k} n_i = n$

$$p_n(n_1, n_2, ..., n_k) = p_n(n_{\sigma(1)}, ..., n_{\sigma(k)}). \tag{2}$$

La fonction p_n est appelée fonction de probabilité de la partition échangeable (EPPF) de la partition aléatoire Π_n. Une partition échangeable de $\mathbb{N}$ est une séquence de partitions échangeables $(\Pi_n)_{n\in\mathbb{N}}$ qui satisfait la propriété de consistance suivante : pour tous $m \leq n$,

$$\Pi_{n|m} =^d \Pi_m$$

Ainsi si la séquence $(\Pi_n)_{n\in\mathbb{N}}$ satisfait la propriété de consistance il existe une fonction p (cette fonction est appelée EPPF) telle que pour tout $\{n_1, ..., n_k\}$, tels que $n_1 + ... + n_k = n$,

$$p(n_1, ..., n_k) = p_n(n_1, ..., n_k).$$

Si la relation de consistance est vérifiée alors p doit satisfaire la règle d'addition suivante :

$$p(n_1, ..., n_k) = \sum_{i=1}^{k} p(..., n_i + 1, ...) + p(n_1, ..., n_k, 1)$$

Maitenant que nous avons defini les EPPF, voici une construction basée sur les polynômes de Bells

1.2 Partitions de Gibbs et polynôme de Bells

On dit de V que c'est une espèce de structure combinatoire si pour tout ensemble fini F_n de cardinal n il existe une construction d'un ensemble $V(F_n)$ de V-structures de F_n telle que le nombre de V structures d'un ensemble à n éléments est v_n. Par exemple V(F_n) peut être $F_n \times F_n$, $F_n^{F_n}$ ou les permutations de F_n dans F_n correspondant respectivement à $v_n = n^2$, n^n ou $n!$. Si W est une autre espèce de structure combinatoire, telle que le nombre de W-structures sur un ensemble à j éléments est w_j, alors $V \circ W(F_n)$ est la structure composée sur F_n et son cardinal est donné par la relation :

$$|(V \circ W)(F_n)| = B_n(v_\bullet, w_\bullet) = \sum_{k=1}^{n} v_k B_{n,k}(w_\bullet), \tag{3}$$

où B_n est le n-ème polynôme de Bells (4).

Etant donné deux suites de réels positifs $v_\bullet = \{v_1, ...\}$ et $w_\bullet = \{w_1, w_2, ...\}$, si $B_{n,k}$ est le (n,k)eme polynôme de Bells defini de la façon suivante :

$$B_{n,k}(w_\bullet) = \sum_{A_1,\dots,A_k \in \mathcal{P}_{[n]}^k} \prod_{i=1}^{k} w_{|A_i|},$$

et si

$$B_n(v_\bullet, w_\bullet) = \sum_{k=1}^{n} v_k B_{n,k}(w_\bullet), \tag{4}$$

alors nous pouvons définir une EPPF de la façon suivante : pour tous $(n_1, n_2, ..., n_k)$ tels que $n_1 + ... + n_k = n$

$$p_n(n_1, n_2, ..., n_k, v_\bullet, w_\bullet) = \frac{v_k \Pi_{i=1}^k w_{n_i}}{B_n(v_\bullet, w_\bullet)}.$$

Les partitions définies avec l'EPPF précedente sont appelées partitions de Gibbs ; elles apparaissent dans des problèmes de combinatoire. A priori elles sont définies uniquement pour un n fixé et la règle de consistance n'est pas vérifiée pour toutes suites $v_\bullet$ et $w_\bullet$ de réels positifs. Pitman et Gnedin en 2006 ont prouve que pour avoir une partition aléatoire infinie sur $\mathbb{N}$, les poids $w_\bullet$ doivent être de la forme $w_{n_j} = [1-\alpha]_{n_j}$ pour un $\alpha \in (-\infty, 1]$. Dans ce cas la, l'EPPF prend la forme suivante :

$$p(n_1, ..., n_k) = V_{(n,k)} \Pi_{i=1}^{k} [1-\alpha]_{n_i},$$

où $V_{n,k}$ est une constante et $[1-\alpha]_k := (1-\alpha)(1-\alpha+1)...(1-\alpha+k-1)$.

2 Théorèmes de Kingman

Dans cette partie nous allons étudier les constructions de partitions aléatoires basées sur le théorème de Kingman

2.1 Constructions de partitions par échantillonnage aléatoire

Si $(x_1, x_2, ..., x_n)$ est une suite de n nombres, alors nous pouvons construire la partition $\Pi(x_1, ..., x_n)$ par la relation d'équivalence :

$$i \sim j \Leftrightarrow x_i = x_j.$$

De la même manière, nous pouvons définir une partition aléatoire avec une suite de variables aléatoire échangeables. On dit d'une suite de variable aléatoire $(Z_1, Z_2, ...)$ qu'elle est échangeable si pour toute permutation σ de $\mathbb{N}$,

$$(Z_1, Z_2, ...) =^d (Z_{\sigma(1)}, Z_{\sigma(2)}, ...).$$

Alors pour toute suite de variables aléatoires échangeables, on peut définir une partition aléatoire avec toujours la relation d'équivalence :

$$i \sim j \text{ ssi } Z_i = Z_j \quad .$$

Urne de Polya Un exemple classique est ceui de l'urne de Polya : prenons un vecteur $u = (u_1, ..., u_r)$ de k entiers ($k \geq 1$) positifs, et notons $|u| = u_1 + ... + u_k$. Chaque entier u_i correspond à une balle de couleur i ; il y a donc en tout k couleurs différentes. Toutes ces balles se trouvent dans une même urne et sont indistinguables au toucher. L'urne suit la dynamique suivante : A chaque instant n, une personne pioche une balle dans cette urne et remet la boule en plus d'une autre boule de même couleur dans l'urne. Pour chaque instant n, $Z_n := i$ si la n-eme boule tirée est de couleur i. Nous avons ainsi défini une suite $(Z_1, Z_2, ...)$ de variables échangeables et la partition associée est trivialement la partition engendrée par les différentes couleurs (deux entiers i et j sont dans la même classe si $X_i = X_j$ soit si les deux boules X_i et X_j sont de la même couleur).

2.2 Théorème de de Finetti

Ce théorème est très important en théorie des partitions aléatoires et est utilisé pour prouver le théorème de Kingman.

Theorem 1 (de Finetti) *La distribution de toute suite infinie de variables échangeables* $(Z_1, Z_2, ...)$ *peut s'ecrire sous la forme suivante :*

$$\mathbb{P}[Z_1, Z_2, .., Z_n] = \int \prod_{i=1}^{n} \mathbb{P}[Z_i|B] d\mathbb{P}(B),$$

où B est un élément aléatoire qui rend les variables aléatoires $\{Z_1, ..., Z_n\}$ *conditionellement indépendantes.*

La principale preuve de ce théorème est donnée dans ([1]) et fait intervenir des martingales rétrogrades.

2.3 Construction de Kingman

Si $x = (x_1, x_2, ...)$ est une partition de masse, c'est à dire

$$\forall i \in \mathbb{N},\ x_i \geq 0 \ \text{ et } \ \sum_{i=1}^{n} x_i \leq 1, \tag{5}$$

nous pouvons associer une partition aléatoire appelée partition de la boite de peinture (en anglais paint box partition) de la façon suivante :

On prend une suite de variables aléatoires iid $X_1, X_2, ...$ à valeurs dans $\mathbb{N}$ avec pour loi

$$\mathbb{P}(X_1 = i) = x_i \text{ et } \mathbb{P}(X_i = 0) = 1 - \sum_i x_i. \tag{6}$$

Alors deux entiers i et j appartiennent au même bloc de la partition aléatoire π si et seulement si $X_i = X_j > 0$. Si F est une loi de probabilité sur $\mathbb{R}_+$ dont les atomes sont $(x_i)_{i\in\mathbb{N}}$, alors on dit de la partition Π_∞ qu'elle est générée par l'échantillonnage de F.

Kingman [3] a prouvé que l'on pouvait construire toutes les partitions $(\Pi_n)_{n\in\mathbb{N}}$ de cette façon :

Theorem 2 (Kingman)
1. *Soit $\Pi_\infty = (\Pi_n)_{n\in\mathbb{N}}$ une partition aléatoire échangeable de $\mathbb{N}$, alors il existe une suite $(P_i^\downarrow)_{i\in\mathbb{N}}$ telle que la partition Π_∞ pouvait être construite en échantillonnant les $P_i^\downarrow$ (comme en [6]).*
2. *De plus si pour tout $n \geq 0$, $(N_{n,i})_{i\geq 0}$ est le réordonnement décroissant de la taille des blocs de Π_n, alors $\frac{N_{n,i}}{n}$ a presque sûrement une limite $P_i^\downarrow$ lorsque $n \to \infty$. De plus, la distribution conditionelle de Π_∞ sachant les $(P_n^\downarrow)_{n\in\mathbb{N}}$ est donnée en [7]*

Lorsque $\sum_i P_i = 1$ presque sûrement, alors on dit que la partition Π a des *bonnes fréquences*. Dans ces cas la, $\tilde{P}_j$ dénote la taille du $j$eme atome decouvert dans le processus d'échantillonnage aléatoire. La séquence $(\tilde{P}_j)_{j\geq 0}$ est appelée *sized-biased permutation* de P_i. C'est à dire que $\tilde{P}_j = P_{\pi_j}$ où, pour toute suite d'entier $(i_j, 1 \leq j \leq k)$ d'entiers positifs distincts, la probabilité de l'évenement ($\pi_j = i_j$ pour tout $1 \leq j \leq k$) sachant $(P_1, P_2, ...)$ est

$$P_{i_1} \frac{P_{i_2}}{1 - P_{i_1}} \cdots \frac{P_{i_k}}{1 - P_{i_1} - \ldots - P_{i_{k-1}}}.$$

Dans ce cas, la loi de Π_n est déterminée par la loi de la suite des fréquences $P_i^\downarrow$ à travers la loi de la permutation conditionnée à la taille par la relation $\tilde{P}_j$. Plus precisemment, l'EPPF p de [1] est donnée par :

$$p(n_1, n_2, ..., n_k) = \mathbb{E}[\prod_{i=1}^{k} \tilde{P}_i^{n_i - 1} \prod_{i=1}^{k-1} (1 - \sum_{j=1}^{i} \tilde{P})_j]. \tag{7}$$

3 Partitions de Poisson-Kingman

Dans cette section, nous allons voir comment construire des partitions aléatoires avec des subordinateurs. L'ensemble de cette partie est tirée de [4] et [5]. Lorsque $P_i^{\downarrow} = J_i^{\downarrow}/T$ sont déterminés par les points réordonnés d'un processus de Poisson, avec densité de Lévy ρ, alors la partition associée est appelée PD(ρ). Nous notons aussi PD($\rho|t$) la partition aléatoire construite comme précédemment en conditionnant par $T = t$. Pour une loi de probabilité γ sur $]0, \infty[$, nous definissons aussi :

$$PD(\rho, \gamma) := \int_0^\infty PK(\rho|t)\gamma(dt). \tag{8}$$

Soit

$$J_1^{\downarrow} \geq J_2^{\downarrow} \geq ... \geq 0$$

les sauts d'un subordinateurs réordonnés par ordre décroissant. Si l'on a $T = \sum_i J_i^{\downarrow} < \infty$ alors nous pouvons définir une suite décroissante de fréquences aléatoires $P_i^{\downarrow} = \dfrac{J_i^{\downarrow}}{T}$. Nous allons voir quelle est la partition de Kingman associée à cette suite de fréquences $(P_i^{\downarrow})_{i \in \mathbb{N}}$.

Nous avons besoin de faire les hypothèses suivantes :

1. Pour tout intervalle I, nous supposons que $N_I = \sum_i \mathbb{1}(J_i^{\downarrow})$ est une variable de Poisson avec pour moyenne $\Lambda(I)$ pour une mesure de Lévy Λ sur $(0, \infty)$ et que les variables $(N_{I_1}, ..., N_{I_k})$ sont indépendantes pour toute collection d'intervalles disjoints $I_1, ..., I_k$.
2. Nous supposons que la mesure Λ definie précédemment est absolument continue par rapport à la mesure de Lebesgue et a pour densité ρ
3. Nous supposons enfin que la loi de T est aussi absolument continue par rapport à la mesure de Lebesgue avec pour densité : $f(t)dt = \mathbb{P}(T \in dt)$.

Le théorème suivant, qui utilise la formule de Palm et les propriétés d'un subordinateur donne l'EPPF d'une partition de Poisson-Kingman :

Theorem 3 *Pour toute partition $\{A_1, ..., A_k\}$ de $[n]$, pour tout $x_i > 0$, $1 \leq i \leq k$, $t > 0$, nous avons :*

$$\mathbb{P}[\Pi_n = \{A_1, ..., A_k\},\ \tilde{J}_i \in dx_i,\ 1 \leq i \leq k,\ T \in dt]$$

$$= \prod_{i=1}^{k} \rho(x_i) dx_i f(t - \sum_{i=1}^{k} x_i) dt \prod_{i=1}^{k} (\frac{x_i}{t})^{|A_i|}.$$

Ainsi, l'EPPF d'une telle partition, appelée partition PK(ρ) est donnée par la formule :

$$p(n_1, ..., n_k) = \int_0^\infty ... \int_0^\infty \frac{f(v) dv \prod_{i=1}^k \rho(x_i) x_i^{n_i} dx_i}{(v + \sum_{i=1}^k x_i)^n}$$

3.1 Exemples de partitions de Poisson-Kingman

3.1.1 Le Gamma subordinateur

Le processus Gamma $(\Gamma_s, s \geq 0)$ est un subordinateur avec paramètre $\theta > 0$ et dont les marginales sont des lois Gamma :

$$\mathbb{P}(\Gamma_s \in dx)/dx = \frac{x^{s-1} e^{-x}}{\theta^s \Gamma(s)}$$

La mesure de Lévy du Gamma subordinateur est :

$$\rho(x) = \theta x^{-1} e^{-x},$$

et son exposant de Laplace est $\psi(u) = \theta \log(1+u)$. Avec un peu de calcul, nous pouvons prouver que l'EPPF correspondant à la partition de Poisson Kingman du subordinateur Gamma est :

$$p(n_1, n_2, ..., n_k) = \frac{\theta^k \Gamma(\theta)}{\Gamma(\theta + n)} \prod_{i=1}^{k} (n_i - 1)!.$$

Cette partition est plus couramment appelé partition $PD(0, \theta)$.

3.2 La subordinateur α stable

Un subordinateur α-stable $(T_s, s \geq 0)$, où $\alpha \in (0,1)$ est un processus de Lévy qui vérifie la relation :

$$T_s =^d s^{1/\alpha} T_1;$$

où T_1 est la loi stable(α) caracterisee par sa transformee de Laplace :

$$\mathbb{E}[e^{-\lambda T_1}] = e^{\lambda^\alpha}.$$

La distribution T_1 est absolument continue par rapport à la mesure de Lebesgue et a pour densité :

$$\rho_\alpha(x) = \frac{\alpha}{\Gamma(1-\alpha)} \frac{1}{x^{\alpha+1}}.$$

Dans le cas du subordinateur α stable, on peut prouver que si $U_1 = 1 - \tilde{P}_1$, et $U_i = 1 - ... - \tilde{P}_i$, alors les U_i sont indépendants et ont des lois Beta$(i\alpha, 1-\alpha)$.

3.3 La partition PD(α, θ)

La partition de Poisson-Dirichlet de paramètres (α, θ) est celle qui est construite avec l'échantillonnage aléatoire suivant :

$$(P_1^\downarrow, P_2^\downarrow, ...) = (B_1, \bar{B}_1 B_2, \bar{B}_1 \bar{B}_2 B_2, ...),$$

où les variables alétoires sont indépendantes et ont pour loi beta$(1-\alpha, \theta + i\alpha)$, et où $\bar{B}_i := 1 - B_i$.. L'EPPF associée à cette séquence de fréquences aléatoires est donnée en [9]. L'une des propriétés de cette EPPF est que pour tous $(i_1, ..., i_k)$ distincts,

$$(P_{i_1}^\downarrow, ..., P_{i_k}^\downarrow) =^d (\tilde{P}_{i_1}^\downarrow, ..., \tilde{P}_{i_k}^\downarrow);$$

les fréquences sont invariantes par 'size-biased permutation'.

4 Le Chinese restaurant Process

4.1 Définition

Dans cette partie, nous allons voir comment constuire la partition PD(α, θ) d'une autre manière, fondée sur la dynamique du Chinese Restaurant Process. Supposons que l'on ait un restaurant chinois avec un nombre infini dénombrable de tables et que les consommateurs (en nombre infinis) rentrent un par un dans ce restaurant. Supposons aussi que l'on ait deux paramètres α et θ qui vérifient l'une des conditions suivantes :

- Soit $\alpha = -\kappa < 0$ et $\theta = m\kappa$, pour $m \in \mathbb{N}$,
- soit $0 \leq \alpha \leq 1$ et $\theta > -\alpha$.

La dynamique des consommateurs est la suivante :

1. Le premier consommateur s'assoit toujours à la premiere table
2. Une fois que les n premiers consommateurs sont entrés et occupent k tables, le $n+1$-ème consommateur choisit une table inoccupée avec probabilité $\dfrac{\theta + k\alpha}{n+\alpha}$ et sinon choisit la $i$eme table occupée avec probabilité $\dfrac{n_i - \alpha}{n+\theta}$ où n_i est le nombre de consommateurs assis à la $i$eme table.

Le Chinese Restaurant Process que je viens de decrire est un processus à valeur partition : une fois que n consommateurs sont entrés, on a naturellement un partition de ces n consommateurs en fonction des tables que ces consommateurs occupent. On peut prouver que la partition Π_∞ créée par ce processus est échangeable et que l'EPPF associée est la suivante :

$$p_{\alpha,\theta}(n_1, ..., n_k) = \frac{[\theta+\alpha]_{k-1\uparrow\alpha}\Pi_{i=1}^{k}(1-\alpha)_{n_i-1\uparrow 1}}{[\theta+1]_{n-1\uparrow 1}}. \tag{9}$$

Ce qui correspond aux fréquences limites suivantes :

$$(\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, ...) = (W_1, \bar{W}_1 W_2, \bar{W}_1 \bar{W}_2 W_3, ...) \tag{10}$$

où les W_i sont indépendants et de loi Beta(1-$\alpha, \theta - i\alpha$) et où $\bar{W}_i := 1 - W_i$.

4.2 Applications en statistiques et le Indian Buffet Process

Le Chinese Restaurant Process a des applications en statistiques dans ce qu'on appelle les modèles de mélange (mixture models). Le principe est le suivant : on observe une suite d'objets ; chaque objet appartient à une classe particulière ; la loi des objets appartenant à une classe donnée depend de le classe en question. Par exemple pour les modèles de mélange Gaussien, les objets sont des variables aléatoires dans lesquels les paramètres de la gaussienne (moyenne et variance) dependent de la classe en question. Alors, la partition créée par la répartition des objets dans les différentes classes se comporte comme un Chinese restaurant process. On peut, par des méthodes de Monte Carlo, déterminer les lois de répartition des objets dans les classes.

Une des limites du modèle de mélange decrit précédemment est qu'il est n'est pas possible de donner plusieurs caracteristiques aux objets observés (dans le Chinese Restaurant Process, les consommateurs ne peuvent s'asseoir qu'à une seule table). Il existe une généralisation de ce processus appelé Indian Buffet Process ([2]). Dans ce processus, les consommateurs entrent dans un restaurant, et consomment un nombre Poissonnien de plats ; chaque consommateur prend à la fois des plats déjà pris et des plats nouveaux. De cette façon on construit une loi sur des classes d'équivalence de matrices aléatoires. En statistiques ([7]), le modèle est plus souple car chaque objet observé peut appartenir à plusieurs catégories (et ainsi Paul peut être à la fois brun, aux yeux bleus, et peser 70kg...). D'un point de vue probabiliste, il est difficile de généraliser la notion d'EPPF à ce problème, les calculs devenant rapidement peu lisibles et l'interprétation avec les processus de Lévy non triviale.

Références

[1] D. Aldous. Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII 1983*, pages 1–198. Springer, 1985.

[2] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process *In NIPS*. pages 475–482 *MIT Press*, 2005.

[3] J. F. C. Kingman. Random Partitions in Population Genetics. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 361(1704) :1–20, 1978.

[4] Jim Pitman. Poisson-kingman partitions *of Lecture Notes-Monograph Series*. pages 1–34, 2002.

[5] Jim Pitman. *Combinatorial Stochastic Processes : Ecole d'Eté de Probabilités de Saint-Flour XXXII - 2002 (Lecture Notes in Mathematics / Ecole d'Eté Probabilit.Saint-Flour)*. Springer, 1 edition, July 2006.

[6] Romain Thibaux and Michael I. Jordan. Hierarchical beta processes and the indian buffet process *In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, 2007.*