Processus Stick-Breaking et extensions pour le traitement bayésien de processus ponctuels

Florencia Chimard, Jean Vaillant

Processus Stick-Breaking et extensions pour le traitement bayésien de processus ponctuels

Florencia CHIMARD et Jean VAILLANT

Université des Antilles et de la Guyane, GRIMAAG, EA3590
BP 592 Campus de Fouillole, 97157 Pointe-à-Pitre CEDEX

Mots-clés: Processus Ponctuel, Stick-Breaking, Processus de Dirichlet, Modèle bayésien hiérarchique

Résumé: Les processus ponctuels sont souvent utilisés comme modèles de répartitions spatiales ou spatio-temporelles d'occurrences. Pour tenir compte de la variabilité de l'environnement, des mélanges de processus ponctuels sont proposés ainsi que l'approche bayésienne hiérarchique pour mener à bien leur inférence statistique. Les lois a priori Stick-Breaking et ses extensions permettent une approche non paramétrique.

Abstract: Point processes are often used as tools for describing spatial or spatio-temporal point patterns. In order to take into account the environmental variability, mixtures of point processes are proposed along with the appropriate bayesian statistical inference. Stick-Breaking Priors and their extensions allow a nonparametric approach.

1 Introduction

Un processus ponctuel (PP) est un mécanisme stochastique qui modélise des localisations de points dans un espace donné. Une répartition de points dans l'espace ou dans le temps peut-être considérée comme la réalisation d'un processus ponctuel caractérisée par un ensemble de coordonnées. Dans la pratique, on ne dispose souvent que d'une réalisation du processus. Les processus ponctuels temporel, spatial et spatio-temporel sont utilisés depuis un certain nombre d'années dans des domaines telles que la biologie, l'épidémiologie, la sismologie, la neurologie, etc... afin de modéliser des séries d'occurrences d'événements (Kallenberg (1976), Daley et Vere-Jones (1988), Karr (1991), van Lieshout (2000)). Un processus ponctuel N(.) est caractérisé sous certaines conditions (Daley et Veres-Jones, 2000) de façon unique par le processus d'intensité conditionnelle ($\lambda(.)$) qui lui est associé. En général, $\lambda(.)$ ne dépend pas seulement du spatial ou/et du temps mais aussi de ce qui s'est passé précédemment. Souvent, dans les répartitions d'occurrences, il existe des variables latentes traduisant les mécanismes sous-jacents au phénomène observé. Dans une telle situation, nous pouvons faire appel aux modèles de mélanges par processus Stick-Breaking (SBP) (Ishwaran et James (2001)) pour caractériser ces variables non observées. Dans nos travaux, nous faisons intervenir plus exactement le processus de Dirichlet (DP) et ses extensions: SBP, Stick-Breaking à noyaux (KSBP).

Le DP a été introduit par Ferguson (1973,1974) et une nouvelle construction plus simple que celle de Ferguson a été introduite par Sethuraman et Tiwari en 1982. Ce processus découle de la loi de Dirichlet qui est utilisée de manière intensive dans de nombreux domaines comme la biologie (Kottas et al. (2007)), la météorologie ou l'astronomie (Ishwaran et James (2002)) par exemple. Ces champs d'applications ont été étendus au domaine de l'informatique en utilisant avec succès des modèles hiérarchiques de Dirichlet et classification par estimation de mélanges de lois (An et al. (2008), Teh et al.(2004)). Le principal avantage est de pouvoir proposer dans le cadre de l'inférence bayésienne des lois a priori non paramétriques (Müller et Quintana (2004)). Ces lois a priori peuvent être définies à partir de paramètres appelés par convention hyperparamètres. On voit ainsi que nous pouvons inclure plusieurs niveaux d'a priori. Le modèle bayésien est alors dit hiérarchique. Le but de notre exposé est de passer en revue quelques extensions du processus de Dirichlet en discutant des motivations et des applications possibles.

2 Processus de Dirichlet et mélanges

Les processus de Dirichlet définissent une mesure de probabilité sur l'espace des mesures de probabilité. Ils permettent donc de définir, dans le cadre de l'estimation bayésienne, un a priori sur une distribution de probabilité inconnue. On pourra se référer aux articles de Ferguson (1973,1974), Sethuraman (1994), Teh (2004) pour de plus amples détails. Une propriété importante est que les réalisations d'un processus de Dirichlet sont discrètes, avec une probabilité 1. Si $\mathrm{G} \sim DP(\alpha, \mathrm{G}_0)$, alors la représentation *Stick-Breaking* de G introduite par Sethuraman (1994) est simplement de la forme suivante:

$$\left\{ \begin{array}{lll} V_k & \overset{i.i.d}{\sim} & Beta(1,\alpha) \qquad\qquad \theta_k \overset{i.i.d}{\sim} \mathrm{G}_0 \\ p_k & = & V_k \prod_{j=1}^{k-1}(1-V_j) \qquad \mathrm{G}(.) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(.) \end{array} \right. \tag{1}$$

où $\delta_{\theta_k}(.)$ est la mesure de Dirac en θ_k. Il est important de noter que la suite $p = (p_k)_{k\in\mathbb{N}^*}$ construite dans l'équation (1) satisfait que les p_k sont des variables aléatoires indépendantes de θ_k tel que $0 \leq p_k \leq 1$ et que $\sum_{k=1}^{\infty} p_k = 1$. Le processus de Dirichlet peut donc être vu comme un mélange infini dénombrable de mesures de Dirac.
Une motivation fondamentale pour l'utilisation du processus de Dirichlet est que la loi a posteriori est également un processus de Dirichlet.

Le processus de Dirichlet dans un mélange peut intervenir comme loi mélangeante, nous parlerons indifféremment de DPM (Dirichlet Process Mixture) ou MDP (Mixture of Dirichlet Processes), ou encore comme loi mélangée et nous parlerons alors de HDP

(Hierarchical Dirichlet Process). Une des plus importantes applications du processus de Dirichlet est son utilisation comme loi a priori dans les composantes d'un modèle de mélanges non paramétrique. Muller et Quintana (2004) discutent de l'analyse bayésienne non paramétrique et comparent différentes variantes ou généralisations du processus de Dirichlet. En particulier, supposons que les observations y_i sont définies comme suit:

$$\begin{array}{rcl} y_i \mid \theta_i & \sim & k(.\mid \theta_i) \\ \theta_i \mid \mathrm{G} & \sim & \mathrm{G} \end{array} \tag{2}$$

où $k(.\mid \theta_i)$ dénote la loi des observations y_i conditionnellement à θ_i. En choisissant comme loi *a priori* pour G, le processus de Dirichlet, on peut reformuler le problème d'estimation de densité selon le modèle hiérarchique (MH) suivant et connu sous le nom de MDP.

$$\begin{array}{rcl} y_i \mid \theta_i & \sim & k(.|\theta_i) \\ \theta_i \mid \mathrm{G} & \sim & \mathrm{G} \\ \mathrm{G} & \sim & DP(\alpha, \mathrm{G}_0) \end{array} \tag{3}$$

Cette écriture peut-être reformulée comme en figure 1 avec $\mathrm{G} \sim DP(\alpha, \mathrm{G}_0)$.

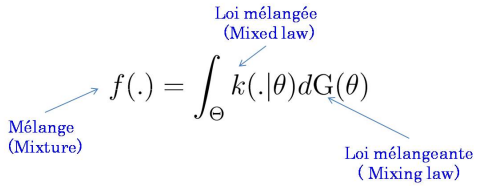

Figure 1. Relation de mélange par rapport à θ.
Chimard et al. (2010)

3 Processus Stick-Breaking et extensions

Ishwaran et James (2001) ont introduit une classe de processus appelée Stick-Breaking Priors construite de la façon suivante:

$$\begin{cases} V_k \overset{ind}{\sim} Beta(a_k, b_k) & \theta_k \overset{i.i.d}{\sim} \mathrm{G}_0 \\ p_k = V_k \prod_{j=1}^{k-1}(1 - V_j) & \mathrm{G}(.) = \mathrm{G}_\infty(1, \alpha, .) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(.) \end{cases} \tag{4}$$

Les processus appartenant à cette classe diffèrent dans la façon d'obtenir les coefficients V_k. On peut citer trois exemples de processus:

1. le processus de Dirichlet vu à la partie précédente: $V_k \sim Beta(1, \alpha)$.

2. le processus de Pitman-Yor encore appelé processus de Poisson-Dirichlet à deux paramètres qui a été développée récemment par Pitman et Yor en 1997: $V_k \sim Beta(a_k, b_k)$ avec,

$$\begin{cases} a_k = 1 - a, \ 0 \leq a < 1 \\ b_k = b + ka, \ b > -a \end{cases}$$

3. le processus de Beta à deux paramètres: $V_i \sim Beta(a, b)$.

Dunson et Park (2008) ont généralisé cette classe de processus stick-breaking en introduisant les processus stick-breaking à noyaux (KSBP):

$$\begin{aligned} \mathrm{G}_x &= \sum_{k=1}^{\infty} \pi_k(x; V_k, \Gamma_k)\mathrm{G}_k^* \\ \pi_k(x; V_k, \Gamma_k) &= \pi_k(x; V_k, \Gamma_k) \prod_{l<k} (1 - \pi_l(x; V_l, \Gamma_l)) \\ \pi_k(x; V_k, \Gamma_k) &= V_k \times K(x, \Gamma_k) \\ V_k &\overset{ind.}{\sim} Beta(a_k, b_k) \\ \Gamma_k &\overset{i.i.d}{\sim} H \\ \mathrm{G}_k^* &\sim Q \end{aligned} \tag{5}$$

où $K \to [0, 1]$ est une fonction de noyau bornée qui est initialement supposée connue, et x un point quelconque de l'espace étudié. Ce modèle est similaire au (4), mais maintenant le SBP est augmenté en employant une fonction de noyau qui quantifie l'a priori associé. Il est employé pour la segmentation d'images en imposant la condition que les pixels voisins dans l'espace sont plus probablement associés dans la même classe (An et al. (2008)).

4 Statistique bayésienne pour les processus ponctuels spatio-temporels

L'analyse statistique d'un processus ponctuel peut-être basé sur différents types d'observations. Lorsqu'il y a un phénomène qui conduit à des occurrences d'événements localisés dans l'espace et dans le temps, on peut avoir les observations sous diverses formes (Vaillant (1991,1992)). En général, les données observées sont sous l'une des formes suivantes:

(i) une carte complète en temps continu dans $X \times [0, T]$ c'est-à-dire pour tout événement donné, on peut avoir sa localisation et sa date d'occurrence. Généralement, ce type d'observation est rare car coûteux.

(ii) k cartes exhaustives à des dates d'observations échelonnés dans le temps $t_1 < t_2 < \ldots < t_k$. En effet, dans la pratique, nous sommes rarement en temps continu, la

vraisemblance du processus ponctuel fournit par son processus intensité n'est pas disponible. Par contre, pour chaque carte spatiale, nous avons cette vraisemblance à partir des intensités cumulées par la période délimitée par deux dates d'observations.

(iii) des comptage d'événements dans des sous-espaces X_i de X à des dates d'observations $t_1 < t_2 < \ldots < t_k$ sans les localisations précises.

Kottas et Sansó (2007) ont proposé une méthode pour l'analyse d'une configuration spatiale de points, qui est supposée résulter d'un ensemble d'observations d'un processus de Poisson spatial non homogène. La méthode est basée sur la modélisation d'une fonction de densité, définie sur une région bornée, qui est directement liée à l'intensité du processus de Poisson. Ils ont développé un modèle de mélange de lois non paramétrique en utilisant une loi Bêta bivariée pour la loi mélangée et un a priori un processus Dirichlet pour la loi mélangeante. Green et Richardson (2002) ont considéré un modèle bayésien hiérarchique semi-paramétrique afin d'analyser des données spatiales de comptages épidémiologiques dans N zones géographiques. Le modèle de processus poissonien est considéré au plus bas niveau de la hiérarchie. Ils discutent de modèles de Markov cachés en cartographie.

Pour notre part, nous nous intéressons à des processus stick-breaking à noyaux pour la prise en compte des localisations locales dans les mélanges spatiaux de processus ponctuels. Des algorithmes de type MCMC pour le traitement approprié de données incomplètes sont présentés et discutés puis sont testés sur des données artificielles.

References

[1] An, Q., Wang, C., Shetervev, I., Wang, E., Dunson, D., and Carin, L. Hierarchical kernel stick-breaking process for multi-task image analysis.

[2] Chimard, F., Vaillant, J., and Daugrois, J.-H. Modélisation de répartitions d'occurrences spatio-temporelles et épidémiologie végétale. *soumis* (2010).

[3] Daley, D., and Vere-Jones, D. *An Introduction to the Theory of Point Processes.* New York, 1988.

[4] Dunson, D., and Park, J.-H. Kernel stick breaking processes. *Biometrica 95*, 2 (2008), 307–323.

[5] Ferguson, T. A bayesian analysis of some nonparametric problems. *The annals of statistics 1* (1973), 209–230.

[6] Ferguson, T. Prior distributions on spaces of probability measures. *The annals of statistics 2* (1974), 615–629.

[7] Green, P., and Richardson, S. Hidden markov models and disease mapping. *American Statistical Association 97*, 460 (2002).

[8] Ishwaran, H., and James, L. F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 453 (2001), 161–173.

[9] Ishwaran, H., and James, L. F. Approximate dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics 11* (2002), 508–532.

[10] Kallenberg, O. *Random Measure*. Berlin, 1976.

[11] Karr, A. F. *Point Processes and their statistical inference*. New York, 1991.

[12] Kottas, A., Duan, J., and Gelfand, A. Modeling disease incidence data with spatial and spatio-temporal dirichlet process mixtures. *Biometrical Journal 49* (2007), 1–14.

[13] Kottas, A., and Sanso, B. Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Statistical Planning and Inference 137* (Mars 2007), 3151–3163.

[14] Müller, P., and Quintana, F. Nonparametric bayesian data analysis. *Statistical Science 19*, 1 (2004), 95–110.

[15] Sethuraman, J. A constructive definition of dirichlet priors. *Statistica Sinica 4* (1994), 639–650.

[16] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical dirichlet processes. *Journal of the American Statistical Association 101* (2004).

[17] Vaillant, J. Negative binomal distributions of individuals and spatio-temporal cox processes. *Scan J. Statist 18* (1991), 235–248.

[18] Vaillant, J. Echantillonnage et étude statistique de populations en milieu héhétrogène. *Statistique Aplliquée 4* (1992), 15–26.