Méthodes bayésiennes non paramétriques pour le traitement du signal

Emmanuel DUFLOS[1,2], François CARON[3], Philippe VANHEEGHE[1,2]

[1]Ecole Centrale de Lille, LAGIS FRE CNRS 3303, Cité Scientifique, BP 48, 59651 Villeneuve d'Ascq Cedex, France

[2]INRIA Lille - Nord Europe - Equipe-Projet SEQUEL, 40, avenue Halley, 59650 Villeneuve d'Ascq - France

[3]INRIA Bordeaux - Sud-Ouest - Equipe-Projet ALEA, Institut de Mathématiques de Bordeaux
Université Bordeaux I, 351 cours de la libération - 33405 Talence Cedex

emmanuel.duflos@ec-lille.fr, francois.caron@inria.fr
philippe.vanheeghe@ec-lille.fr

Résumé – Les méthodes bayésiennes paramétriques font partie intégrante de la boîte à outil du chercheur en traitement statistique du signal. Les méthodes paramétriques font l'hypothèse que le modèle peut être caractérisé par un paramètre de dimension finie. Les méthodes bayésiennes non paramétriques étendent la gamme en considérant des modèles caractérisés par un paramètre de dimension infinie. Ceci évite ainsi de fixer la complexité ou l'ordre du modèle, le nombre de paramètres pouvant augmenter avec le nombre de données. Les propriétés de conjugaison et de marginalisation de plusieurs modèles les rendent particulièrement attractifs en inférence bayésienne. L'objet de cet article est de présenter plusieurs de ces modèles afin de populariser leur utilisation en traitement statistique du signal. Nous aborderons en particulier les Processus de Dirichlet (classification non supervisée/estimation de densité), processus Beta-Bernoulli (modèles à facteurs latents) et modèles de Markov caché infini.

Abstract – Parametric Bayesian methods are well known and widely used in statistical signal processing. These methods assume that a model can be characterized by a finite-dimensional parameter. Nonparametric Bayesian methods extend such a point of view by considering that the parameter belongs to an infinite-dimensional space. The model order or the complexity may therefore not be fixed in advance and the number of parameters can increase along with the number of available data. Conjugacy and marginalization properties of several models allow efficient Bayesian inference. The main objective of this paper is to present some of these models to facilitate their spread in the statistical signal processing community. We will present Dirichlet Processes (supervised/unsupervised classification, density estimation), Beta-Bernoulli Processes (latent factor models) and infinite Markov models.

1 Introduction

Les méthodes bayésiennes paramétriques font l'hypothèse que le modèle des données prend une forme paramétrique fixée (e.g. gaussienne, student t, mélange fini de gaussiennes, etc.) et peut donc être défini par un paramètre de dimension finie.

Dans le cas de la classification non supervisée de données, on suppose que les données sont des réalisations indépendantes d'une distribution inconnue, par exemple un mélange fini de densités de probabilités gaussiennes, de paramètres (finis) inconnus à estimer. Chaque mode du mélange correspond à un cluster différent.

Dans le cas d'un modèle de Markov caché discret, on suppose que l'état caché peut sauter d'un mode à l'autre, le nombre total de modes étant fixé. Si elles sont inconnues, il est possible de définir un modèle de Dirichlet paramétrique sur les probabilité de transition d'un mode à l'autre afin d'estimer celles-ci à partir des données [2, Chap. 13].

Le choix d'un modèle paramétrique peut être assez restrictif car il contraint le modèle à une forme donnée, parfois difficile à spécifier. Afin de gagner en robustesse, il peut être souhaitable de considérer que la distribution inconnue a un support plus large que celui fourni par un modèle paramétrique donné. Au lieu de définir une distribution a priori sur un espace de dimension finie, les modèles bayésiens non paramétriques définissent de ce fait une distribution de probabilité sur des espaces fonctionnels (de dimension infinie). Un modèle non paramétrique peut ainsi être simplement considéré comme un modèle statistique avec un nombre infini de paramètres [13]. Une définition alternative est un modèle où la complexité du modèle augmente avec le nombre de données.

Les modèles bayésiens non paramétriques les plus populaires sont actuellement les processus gaussiens et les processus de Dirichlet. En particulier, le processus de Dirichlet à mélange (Dirichlet Process Mixture, DPM) est une distribution sur les distributions de probabilité. Le DPM dépend de deux paramètres et ses réalisations sont des mélanges infinis, par exemple de densités gaussiennes. Le nombre de clusters ne doit pas être défini a priori, mais est estimé à partir des données. Bien que les processus de Dirichlet soient des objets statistiques connus depuis le début des années 1970 [8], ces modèles ne sont vraiment devenus populaires que récemment grâce au développement des méthodes Markov Chain Monte Carlo (MCMC) permettant d'estimer de tels modèles. Ils ont depuis connu une grande

popularité pour l'estimation de densités et la classification non supervisée dans des domaines d'application variés. L'article [8] a reçu la plupart de ses 1900 citations (source Google Scholar) au cours des dix dernières années. Les méthodes bayésiennes non paramétriques ont connu un intérêt croissant ces quinze dernières années dans la communauté de l'apprentissage machine (machine learning) [18] et des statistiques [13], avec l'appropriation de modèles existants et le développement de nouveaux modèles pour la résolution de problèmes complexes. Ces méthodes sont cependant encore mal connues en traitement du signal, et l'objectif de ce tutoriel est de présenter plusieurs modèles bayésiens non paramétriques.

De la même façon que pour les modèles bayésiens paramétriques, la popularité d'un modèle bayésien non paramétrique dépend de ses propriétés de conjugaison. Ces propriétés sont particulièrement intéressantes pour les modèles non paramétriques, car dans le cadre de modèles hiérarchiques, elles permettent d'intégrer de façon analytique selon le paramètre de dimension infinie, et d'effectuer l'inférence dans un cadre paramétrique classique.

Nous présenterons dans la suite les processus du restaurant chinois, associé au processus de Dirichlet, le modèle dit du *buffet indien* qui permet de modéliser des objets possédant un nombre inconnu de caractéristiques [10]. Enfin, nous présentons les modèles de Markov cachés infinis permettant l'apprentissage d'une chaîne de Markov où le nombre d'éléments est potentiellement infini [17].

2 Processus de Dirichlet

Les processus de Dirichlet définissent une distribution sur l'ensemble des distributions de probabilité. Ils permettent de définir, dans un cadre bayésien, un a priori sur une distribution de probabilité inconnue.

Définition 1 *Soit $(\mathcal{X}, \mathcal{A})$ un espace mesurable et $\mathbb{G}_0$ une mesure de probabilité sur $(\mathcal{X}, \mathcal{A})$. Soit α un réel positif. Une distribution de probabilité $\mathbb{G}$ est distribuée selon un processus de Dirichlet de distribution de base $\mathbb{G}_0$ et de facteur d'échelle $\alpha > 0$ si pour n'importe quelle partition $A_1, ..., A_k$ de $\mathcal{X}$, le vecteur de probabilité aléatoire $[\mathbb{G}(A_1), ..., \mathbb{G}(A_k)]$ suit une distribution de Dirichlet :*

$$[\mathbb{G}(A_1), ..., \mathbb{G}(A_k)] \sim \mathcal{D}(\alpha\mathbb{G}_0(A_1), ..., \alpha\mathbb{G}_0(A_k)) \quad (1)$$

où $\mathcal{D}$ est la distribution de Dirichlet standard. On le note

$$\mathbb{G} \sim DP(\mathbb{G}_0, \alpha) \quad (2)$$

Une réalisation d'un processus de Dirichlet est presque sûrement discrète, et prend la forme dite "stick-breaking" suivante

$$\mathbb{G} = \sum_{j=1}^{\infty} \pi_j \delta_{U_j} \quad (3)$$

avec $U_j \sim \mathbb{G}_0$, $\beta_j \sim \mathcal{B}(1, \alpha)$ et $\pi_j = \beta_j \prod_{l=1}^{j-1}(1 - \beta_l)$, où $\mathcal{B}(a, b)$ est la distribution Beta standard de paramètres a et b.

Les processus de Dirichlet possèdent des propriétés de conjugaison qui les rendent particulièrement attractifs dans un cadre d'estimation, car elles permettent la mise en œuvre de mécanismes simplifiés des principes d'inférence bayésienne. Considérons le modèle hiérarchique suivant :

$$\mathbb{G} \sim DP(\alpha, \mathbb{G}_0)$$

et pour $i = 1, \ldots, n$

$$\theta_i | \mathbb{G} \sim \mathbb{G}$$

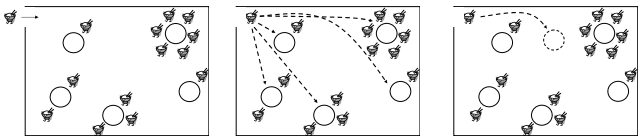
alors, la distribution a posteriori de $\mathbb{G}$ conditionnellement à $(\theta_1, \ldots, \theta_n)$ est toujours distribuée selon un processus de Dirichlet

$$\mathbb{G}|\theta_1, \ldots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n}\mathbb{G}_0 + \frac{1}{\alpha + n}\sum_{i=1}^{n}\delta_{\theta_i}\right)$$

De façon plus importante, il est possible d'obtenir analytiquement la loi de $\theta_{n+1}|\theta_1, \ldots, \theta_n$, pour obtenir la représentation en urne de Polya (également appelée représentation de Blackwell-Mac Queen) suivante :

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n}\mathbb{G}_0 + \frac{1}{\alpha + n}\sum_{i=1}^{n}\delta_{\theta_i} \quad (4)$$

On voit ainsi que, conditionnellement aux valeurs des variables latentes déjà échantillonnées, un effet de clustering apparait : la variable θ_{n+1} va prendre une valeur précédente avec probabilité $\frac{n}{n+\alpha}$, et une nouvelle valeur avec probabilité $\frac{\alpha}{n+\alpha}$. La distribution induite sur les partitions porte le nom de *Processus du Restaurant Chinois*, voir la figure 1.



(a) (b) (c)

FIGURE 1 – Métaphore du restaurant chinois [3]. (a) Un nouveau client entre dans un restaurant où se trouve un nombre infini de tables. Seul un nombre fini de tables est occupé par une ou plusieurs personnes. Le nouveau client a alors deux choix possibles. Soit (b) il se joint à une table déjà occupée avec une probabilité proportionnelle au nombre de personnes à cette table. Soit (c) il s'assoit à une nouvelle table, avec une probabilité proportionnelle à α.

Ces propriétés de conjugaison, et la représentation en urne de Polya ci-dessus, sont particulièrement importantes pour l'inférence : bien que l'on traite de façon sous-jacente des paramètres de dimension infinie, il est possible de marginaliser selon ces paramètres, et de travailler ainsi uniquement avec des paramètres de dimension finie.

Pour des applications de clustering ou d'estimation de densité, on considère généralement le modèle hiérarchique suivant [7], pour $i = 1, \ldots, n$

$$\theta_i|\mathbb{G} \sim \mathbb{G} \text{ et } y_i|\theta_i \sim f(.|\theta_i) \quad (5)$$

où $\mathbb{G} \sim DP(\mathbb{G}_0, \alpha)$ et f est une distribution paramétrique connue (e.g. gaussienne). Cette formulation étend les modèles de mélange finis [12], et permet d'estimer le nombre de clusters à partir des données. Le paramètre $\alpha > 0$ définit de façon implicite un a priori sur le nombre de clusters observés pour n donné. Une valeur de α faible va favoriser un faible nombre de clusters et inversement.

Les propriétés de conjugaison des processus de Dirichlet discutées ci-dessus permettent de définir des algorithmes de Monte Carlo par chaîne de Markov pour échantillonner selon la loi a posteriori $P(\theta_1, \ldots, \theta_n|y_1, \ldots, y_n)$, voir [14] et [1, 6] pour différentes applications en traitement du signal.

3 Modèle du *buffet indien*

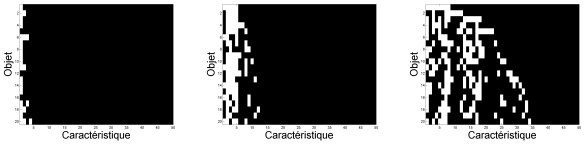
On s'intéresse ici à des problématiques où l'on observe plusieurs objets, chaque objet possédant un ensemble de caractéristiques inconnues. Par exemple, les objets peuvent être des images, dans lesquelles on cherche à extraire des formes (voiture, arbre, etc.). De façon mathématique, ceci peut être représenté par une matrice binaire Z, où $Z_{i,j} = 1$ si l'objet i possède la caractéristique j. Dans un cadre bayésien, on cherche à définir un a priori sur cette matrice Z. Dans le cas où le nombre de caractéristiques est inconnu, le modèle dit du *buffet indien* [10] permet de définir un a priori sur une matrice binaire de dimension infinie.

De la même façon que pour les processus de Dirichlet, le modèle peut être introduit à l'aide de mesures aléatoires, en utilisant un modèle hiérarchique basé sur un processus de beta [19]. Il est possible de marginaliser analytiquement selon la distribution inconnue pour se ramener au processus dit du *buffet indien* [1]. Nous nous intéressons ici uniquement à ce processus, en définissant la loi de $Z_{n+1}|Z_1, \ldots, Z_n$ où $Z_i = (Z_{i,1}, Z_{i,2}, \ldots)$ est un vecteur binaire de dimension infinie.

Dans la métaphore du buffet indien, les objets sont des clients, et les caractéristiques des plats. Le premier client choisit un nombre $N_1 \sim Poisson(\alpha)$ de plats. Pour chaque plat j déjà choisi par au moins un client précédent, le client i choisit le plat j avec probabilité $\frac{m_j}{j}$ où m_j est le nombre de fois où le plat j a été choisi. Puis il choisit un certain nombre $N_i \sim Poisson(\frac{\alpha}{i})$ de nouveaux plats. Ce processus induit un phénomène de renforcement, les plats/caractéristiques les plus populaires ayant plus de chances d'être choisis. Chaque nouveau client peut choisir un certain nombre de nouveaux plats, et le nombre total de plats n'est pas fixé à l'avance.

Des réalisations de ce processus sont représentées sur la figure 2 pour différentes valeurs du paramètre α. Ce dernier règle l'a priori sur le nombre de caractéristiques observées, une valeur faible favorisant peu de caractéristiques et inversement.

1. Son nom provient du fait qu'il est au processus Beta-Bernoulli ce que le processus du restaurant chinois est au processus de Dirichlet



(a) $\alpha = 1$ (b) $\alpha = 3$ (c) $\alpha = 10$

FIGURE 2 – Réalisations selon un processus du buffet indien pour différentes valeur du coefficient α.

De la même façon que pour le restaurant chinois, cette formulation permet de définir des algorithmes MCMC simples d'implémentation pour estimer la matrice Z à partir des données.

4 Modèles pour les séries temporelles

4.1 Modèles de Markov cachés infinis

Soit (s_t) une chaine de Markov discrète, de matrice de transition P. Dans un cadre bayésien complet, on cherche à estimer à la fois les états s et la matrice de transition P [2].

Basé sur une hiérarchie de processus de Dirichlet, le modèle de Markov caché infini [17] permet de définir un a priori sur une matrice stochastique P de dimension infinie. Il est possible d'intégrer analytiquement selon la matrice de dimension infinie P afin de se ramener à un processus de renforcement en urne pour la loi de $s_t|s_1, \ldots, s_{t-1}$ [17]. Ces modèles permettent d'estimer le nombre d'états différents, qui n'est pas fixé à l'avance, en évitant des méthodes de type MCMC à sauts réversibles [2].

4.2 Systèmes dynamique linéaires

En utilisant les techniques de Kalman, il est possible d'effectuer l'estimation optimale dans les modèles d'état linéaires à bruits gaussiens. Lorsque la distribution des bruits est inconnue, et n'a pas de forme paramétrique donnée, un modèle bayésien non paramétrique basé sur les processus de Dirichlet a été dérivé dans [5] afin d'estimer conjointement les états et les distributions inconnues des bruits. Il a notamment été appliqué pour l'estimation des erreurs de mesures d'un signal GPS [16], la classification et le suivi d'objets dans une séquence vidéo [11] ou le traitement de la parole [15].

Cette approche a été étendue dans [9] en considérant des modèles linéaires à saut, se basant sur un modèle de Markov infini.

5 Conclusion

Les méthodes bayésiennes non paramétriques offrent un cadre de recherche très dynamique : d'un point de vue pratique, elles permettent d'étendre la gamme des modèles bayésiens, et de s'affranchir de techniques complexes pour la sélection de mo-

dèle ; d'un point de vue méthodologique, elles permettent le développement de nouveaux modèles, pour résoudre des problématiques typiques du traitement du signal ; d'un point de vue théorique, ces méthodes sont basées sur des mesures aléatoires, et offrent des challenges nouveaux concernant la consistence a posteriori de ces modèles.

Il est à noter qu'un workshop est organisé tous les deux ans sur ce thème rassemblant des personnes d'horizons différents. Plusieurs tutoriels en anglais existent également pour aller plus loin sur ces méthodes.

Références

[1] E. Barat, C. Comtat, T. Dautremer, T. Montagu et R. Trebossen *IEEE Nuclear Science Symposium Conference Record, 2007.*

[2] O. Cappé, E. Moulines et T. Ryden. *Inference in Hidden Markov Models*. Springer Series in Statistics, 2005.

[3] F. Caron. *Inférence bayésienne pour la détermination et la sélection de modèles stochastiques*. Mémoire de thèse de l'Ecole Centrale de Lille, 2006.

[4] F. Caron, M. Davy et A. Doucet. *Generalized Polya Urn for Time-varying Dirichlet Process Mixtures*. 23rd Conference on Uncertainty in Artificial Intelligence (UAI'2007), Vancouver, Canada, July 2007.

[5] F. Caron, M. Davy, A. Doucet, E. Duflos et P. Vanheeghe. *Bayesian inference for linear dynamic models with dirichlet process mixtures*. IEEE Transactions on Signal Processing, 56(1) :71 - 84, 2008.

[6] M. Davy et J.Y. Tourneret. *Generative Supervised Classification Using Dirichlet Process Priors*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 10, pp. 1781-1794, 2010.

[7] M. Escobar et M. West. *Bayesian Density Estimation and Inference using Mixture*. Journal of the American Statistical Association, volume 90, pp 577-588, 1995.

[8] T.S. Ferguson. *A Bayesian analysis of some nonparametric problems*. The Annals of Statistics, 1(2) :209 - 230, 1973.

[9] E.B. Fox, E.B. Sudderth, M.I. Jordan et A.S. Willsky. *Bayesian Nonparametric Methods for Learning Markov Switching Processes*. In IEEE Signal Processing Magazine, vol. 27, pp. 43-54, 2010.

[10] T.L. Griffiths et Z. Ghahramani. *Infinite latent feature models and the Indian Buffet Process*. In Advances in Neural Information Processing Systems 18, 2006.

[11] K. Ishiguro, T. Yamada et N. Ueda. *Simultaneous clustering and tracking unknown number of objects*. IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[12] J.-M. Marin et C. Robert *Bayesian Core : A practical approach to computation Bayesian Statistics*. Springer Texts in Statistics, 2007.

[13] P. Müller et F.A. Quintana. *Nonparametric Bayesian Data Analysis*. Statistical Science, 19(1) :95 - 110, 2004.

[14] R. Neal. *Markov Chain Sampling Methods for Dirichlet Process Mixtures Models*. Journal of Computational and Graphical Statistics, volume 9, pp 249-265, 2000.

[15] K. Ota, E. Duflos, P. Vanheeghe et M. Yanagida. *Speech recognition with speech density estimation by the Dirichlet Process Mixture*. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008. pp 1553 - 1556, 2008.

[16] A. Rabaoui, N. Viandier, J. Marais et E. Duflos. *Selecting the hyperparameters of the DPM Models for the Density Estimation of Observation Errors*. ICASSP 2011, Prague, Mai 2011.

[17] Y.W. Teh, M.J. Beal, M.I. Jordan et David M. Blei. *Hierarchical Dirichlet Processes* Journal of the American Statistical Association, 101(476), pp 1566-1581, 2006.

[18] Y.W. Teh et M.I. Jordan. *Hierarchical Bayesian Nonparametric Models with Applications*. In Bayesian Nonparametrics : Principles and Practice. Edité par N. Hjort, C. Holmes, P. Müller et S. Walker. Cambridge University Press, 2010.

[19] R. Thibaux et M.I. Jordan. *Hierarchical Beta Processes and the Indian Buffet Process*. In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, 2007.