Quantification de l'incertitude d'une partition issue d'un processus de Dirichlet à mélange

Aurore Lavigne [1] & Silvia Liverani [2]

[1] *Université Lille 3, LEM-CNRS (UMR 9221)*
[2] *Department of Mathematics, Brunel University London; Department of Epidemiology and Biostatistics, Imperial College London; MRC Biostatistics Unit, Cambridge*

Résumé. Nous présentons ici nos résultats sur la quantification de l'incertitude liée à une partition. Dans la littérature sur la classification, une unique partition est généralement identifiée comme "optimale" par rapport à un critère donné, et l'incertitude sur cette partition n'est en général pas discutée. En effet, l'espace des partitions est vaste et complexe, et quantifier cette incertitude reste une tâche difficile.

Nous nous intéressons à l'incertitude associée aux partitions obtenues à l'aide d'un processus de Dirichlet à mélange sous le paradigme bayésien. Nous proposons deux méthodes pour quantifier l'incertitude. L'une est basée sur la distribution marginale *a posteriori* de la variable d'allocation du processus de Dirichlet, l'autre sur la comparaison des probabilités d'appartenance de chaque individu à chaque classe dans le modèle de mélange estimé. Pour cette seconde méthode, nous fournissons aussi une représentation graphique de ces probabilités. Finalement, nous étudions comment ces méthodes sont liées, et nous les utilisons pour comparer certaines des stratégies utilisées pour définir la partition "optimale". Nous appliquons ces méthodes à un jeu de données en océanographie.

Mots-clés. Classification, partition, incertitude, processus de Dirichlet

Abstract. We report our results on the quantification of partition uncertainty. In the clustering literature, it is common practice to identify one partition as optimal with respect to some criteria and disregard the overall uncertainty associated with it. The partition space is vast and complex and quantifying this uncertainty is a challenging task. Usually uncertainty is reported for some parameters of a clustering model, but not for the cluster allocations.

Here we focus on the uncertainty of partitions obtained using a Dirichlet process Bayesian clustering model. We propose two methods for learning about uncertainty of partitions, one based on the marginal posteriori density of the Dirichlet process allocation variable, the other one is based on the comparison of probabilities of belonging of each individual to each cluster, conditionnally to the estimate mixture model. For the latter method, we also provide a visual representation of uncertainty. We discuss how these two methods relate to each other and we use them to compare some of the most popular strategies to identify the optimal partition. We apply these methods to an oceanography dataset.

Keywords. Clustering, partition, uncertainty, Dirichlet process

1 Introduction

Bien que le processus de Dirichlet ait initiallement été développé pour résoudre des problèmes non paramétriques sous le paradigme Bayésien (Ferguson, 1973), il est largement utilisé pour la classification, puisqu' il peut être utilisé comme distribution a priori sur les paramètres d'un modèle de mélange dont le nombre de composantes est infini. Des algorithmes de Monte-Carlo par chaînes de Markov ont été développés pour échantilloner les paramètres et la variable latente d'allocation dans la distribution *a posteriori*. De plus, ce modèle ne nécessite de pas spécifier le nombre de classes attendues, il est aussi estimé *a posteriori*. C'est pourquoi le processus de Dirichlet à mélange est utilisé dans de nombreux domaines comme l'épidémiologie (Liverani et al., 2014), la génomique ou le machine learning.

Dans la littérature sur la classification, une partition est généralement identifiée comme optimale et discutée sans égard aux autres classifications possibles alors même que l'algorithme MCMC fournit un échantillon de partitions issues de la distribution *a posteriori*. Plusieurs méthodes existent pour définir cette partition "optimale" et selon la méthode utilisée, des classifications trés différentes peuvent être obtenues. Il en est de même lorsqu'on utilise la même méthode sur deux tirages différents dans la loi *a posteriori*. Nous supposons que cette variabilité est due à l'incertitude sur la classification. Cependant l'espace des partitions est grand et complexe, et quantifier l'incertitude est une tâche difficile. Habituellement, seule l'incertitude sur quelques paramètres d'un modèle de classification est discutée.

Dans un premier temps, nous rappelons le modèle de Dirichlet à mélange et nous présentons certaines des méthodes utilisées pour obtenir une classification optimale, puis nous présentons nos deux méthodes pour quantifier l'incertitude liée à une partition. Enfin, nous appliquons ces méthodes pour déterminer l'incertitude sur la classification de biorégions en méditerranée.

2 Modèle

Dans un processus de Dirichlet à mélange, les données $\mathbf{X} = (X_1, \cdots, X_n)$ sont supposées distribuées selon le mélange infini dont la distribution de mélange $\mathbf{G}()$ est tirée dans un processus de Dirichlet $DP(\alpha, \mathbf{P}_{\theta_0})$ de distribution de base $\mathbf{P}_{\theta_0}$ et de paramètre d'échelle α. On suppose de plus que la composante c du mélange est une distribution paramètrique de $f(.|\Theta_c)$ définie à l'aide du paramètre Θ_c spécifique à la composante c.

$$\begin{array}{l} P(x) = \int f(x|\Theta)\mathbf{G}(\Theta, \Psi)d(\Theta, \Psi) \\ \mathbf{G}(\Theta, \Psi) = \sum_{c=1}^{\infty} \psi_c \delta_{\Theta_c} \\ \mathbf{G} \sim DP(\alpha, P_{\Theta_0}) \end{array} \tag{1}$$

Dans une formulation hiérarchique nous pouvons introduire une variable latente d'allocation

$\mathbf{Z} = (Z_1, \cdots, Z_n)$, telle que $Z_i = c$ si l'individu i est classé dans la classe c.

$$\begin{array}{l} X_i|Z_i, \Theta \sim f(X_i|\Theta_{Z_i}) \text{ pour } i = 1, \cdots, n \\ P(Z_i = c|\boldsymbol{\psi}) = \psi_c \\ \mathbf{G} = \sum_{c=1}^{\infty} \psi_c \delta_{\Theta_c} \text{ pour } i = 1, \cdots, n \\ \mathbf{G} \sim DP(\alpha, P_{\Theta_0}). \end{array} \tag{2}$$

Sous le paradigme bayésien, nous nous intéressons à la distribution *a posteriori* de la mesure $\mathbf{G}$ définie par la position des classes Θ et leur probabilité $\boldsymbol{\psi}$. Plusieurs algorithmes MCMC ont été développés (Iwasharan et James (2001)), ils fournissent des tirages de Θ et $\boldsymbol{\Psi}$ dans leur distribution *a posteriori*, mais aussi des tirages de la variable latente $\mathbf{Z}$, qui sont ensuite utilisés pour obtenir la partition "optimale". A cause du *label switching* et de la variation du nombre de classes d'un tirage à l'autre, il n'est pas possible d'extraire directement une classification *a posteriori* à partir de la variable $\mathbf{Z}$. La classification "optimale" est alors obtenue à partir d'une matrice de similarités, donnant pour chaque paire d'individus leur probabilité d'être classés ensembles. De nombreuses méthodes existent pour trouver la partition "optimale", on en distinguera deux types.

- Dans certains cas, la partition optimale est recherchée parmi les partitions échantillonnées par l'algorithme MCMC, elle doit minimiser un critère de perte ou être le maximum *a posteriori*.
- Dans d'autres cas, une nouvelle partition est obtenue en utilisant une méthode de classification à la matrice de similarités, cette partition n'appartient pas forcément à l'ensemble des partitions échantillonnées. C'est par exemple le cas de la méthode PAM (Partioning Around Medoids) utilisée dans l'application.

3 Quantification de l'incertitude sur la partition

Mesure d'incertitude dans l'espace des partitions Il est possible de quantifier la vraisemblance des partitions échantillonnées, en calculant la probabilité marginale *a posteriori* de chaque partition $P(\mathbf{Z}|\mathbf{X})$ lorsqu'on choisit une distribution de base P_{Θ_0} conjuguée avec la distribution paramétrique $f(.|\Theta_c)$ (Hastie et al, 2014). Dans l'application on considérera le prior Normal-Inverse Wishart conjugué avec le modèle de mélange gaussien multivarié,

$$P(\mathbf{Z}|\mathbf{X}) = \int P(\mathbf{X}|\mathbf{Z}, \Theta) P(\Theta) d\Theta P(\mathbf{Z}).$$

En pratique cette probabilité est calculée à une constante multiplicative près dépendant des données $\mathbf{X}$ et donc ne permet pas d'évaluer intrinsèquement l'incertitude d'une partition. Cependant, elle permet de permet de comparer des partitions tirées dans la distribution *a posteriori* et donc d'identifier un Maximum A Posteriori (MAP).

Incertitude sur la classification des individus, à l'intérieur d'une partition Une seconde approche consiste à considérer le modèle de mélange fini issu de la partition "optimale" et de considérer, comme dans l'algorithme EM, la probabilité que l'individu i appartienne à une classe l, conditionnellement au modèle de mélange estimé. En effet, à partir de la classification "optimale" sélectionnée, notée $\mathcal{C} = (s_1, s_2, \cdots, s_k)$, la distribution des paramètres de la classe l, $(\widetilde{\Theta}_l, \widetilde{\Psi}_l)$ peut être obtenue en mélangeant les distributions $(\Theta_{Z_i}, \Psi_{Z_i})$ des individus de la classe l,

$$P(\widetilde{\Theta}_l, \widetilde{\Psi}_l|\mathbf{X}) = \frac{1}{n_l} \sum_{i \in s_l} \sum_{c=1}^{\infty} P(\Theta_c, \psi_c|\mathbf{X}, Z_i = c) P(Z_i = c|\mathbf{X}).$$

On peut alors approximer la distribution prédictive *a posteriori* par un modèle de mélange fini,

$$P(X_{n+1}|\mathbf{X}) \approx \sum_{l=1}^{k} \int \widetilde{\Psi}_l f(X_{n+1}|\widetilde{\Theta}_l) P(\widetilde{\Theta}_l, \widetilde{\Psi}_l|\mathbf{X}) d(\widetilde{\Theta}_l, \widetilde{\Psi}_l).$$

Finalement, on pourra calculer p_i^l la probabilité conditionnelle que l'individu i appartienne à la classe l,

$$p_i^l = \frac{\int \widetilde{\Psi}_l f(X_i|\widetilde{\Theta}_l) P(\widetilde{\Theta}_l, \widetilde{\Psi}_l|\mathbf{X}) d(\widetilde{\Theta}_l, \widetilde{\Psi}_l)}{\sum_{j=1}^{k} \int \widetilde{\Psi}_j f(X_i|\widetilde{\Theta}_j) P(\widetilde{\Theta}_j, \widetilde{\Psi}_j|\mathbf{X}) d(\widetilde{\Theta}_j, \widetilde{\Psi}_j)}$$

Cette méthode très simple à l'avantage de fournir une mesure d'incertitude pour chacun des individus, mais a le défaut de ne considérer qu'une seule des partitions. Nous proposons de présenter ces données sous la forme d'une matrice de taille $n \times k$, dans laquelle la probabilité conditionnelle p_i^l que l'individu i appartienne à la classe l connaissant le modèle de mélange est figurée sur la i^e ligne et la l^e colonne par un niveau de gris. Pour une représentation plus simple, les individus sont ordonnés, d'abord selon leur classe, puis selon leur probabilité d'appartenance à la classe dans laquelle ils ont été classés.

4 Application

La mer méditerranée est reconnue comme un point sensible du changement climatique, et dans ce contexte les chercheurs recherchent une partition de la mer en régions spatiales homogènes pour mettre en place une stratégie de surveillance de l'environnement marin. Ici, nous nous intéressons uniquement au bassin ouest qui regroupe des zones très hétérogènes, notamment à la fin de l'hiver. Nous cherchons une classification de cette région, à partir des moyennes mensuelles de surface de février 2013 des principaux paramètres physiques (température et salinité) et biogéochimiques (concentration en nitrate et chlorophylle).

Le processus de Dirichlet à mélange et la méthode PAM sont appliqués à ces données. A l'aide du package PReMiuM (Silverani et al, 2014), nous avons procédé à 10000 tirages dans la distribution *a posteriori*, après une période de chauffe de 10000 itérations. La

partition "optimale" du bassin ouest de la méditerranée est présentée sur la figure 1 et l'incertitude quantifiée avec la second méthode est présentée Fig. 2.

Le nombre de classes obtenues (9) est relativement grand en comparaison avec d'autres études. La classe 1, correspondant à toute la partie sud du bassin, parait très certaines, car elle est majoritairement conposée d'individus dont la probablité d'appartenance à cette classe est forte. Au contraire, les classes 2 et 3 paraissent peu certaines et pourraient être fusionnées. Enfin, les autres classes sont composées à la fois par des individus ayant une forte probabilité et une faible probabilité d'appartenance à la classe, ce qui suggère que les frontières de ces classes sont floues.

5 Perspectives

Nous présentons ici deux méthodes pour quantifier l'incertitude liée à une partition obtenue par un processus de Dirichlet à mélange. La première méthode à l'avantage de s'appliquer dans l'espace des partitions, mais ne permet que de comparer les partitions. La seconde méthode, permet d'étudier plus finement les points d'incertitudes d'une partition, mais s'applique conditionnellement au modèle de mélange estimé. Par la suite, nous projetons de combiner les résultats des deux méthodes, pour sélectionner la partition la plus certaine tout en précisant son incertitude.

Bibliographie

[1] Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2), 209–230.

[2] Liverani, S., Hastie D. I., et Richardson S. (2014), PReMiuM: An R Package for Profile Regression Mix- ture Models using Dirichlet Processes. *A paraitre dans the Journal of Statistical Software.* Preprint available at arXiv:1303.2836.

[3] Ishwaran, H., James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453), 161–173

[4] Hastie, D. I., Liverani, S., et Richardson, S. (2014). Sampling from Dirichlet process mixture models with unknown concentration parameter: Mixing issues in large data implementations. *A paraitre dans Statistics Computing.* Preprint available at arXiv:1304.1778.

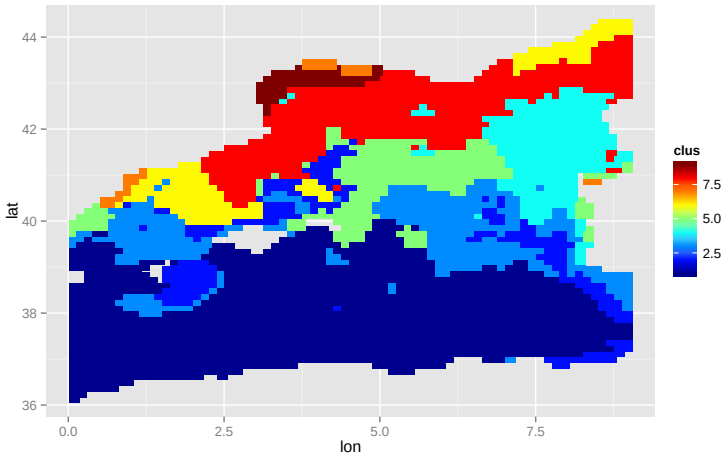

Figure 1: Classification du bassin ouest de la méditerranée en 9 zones par le processus de Dirichlet à mélange.

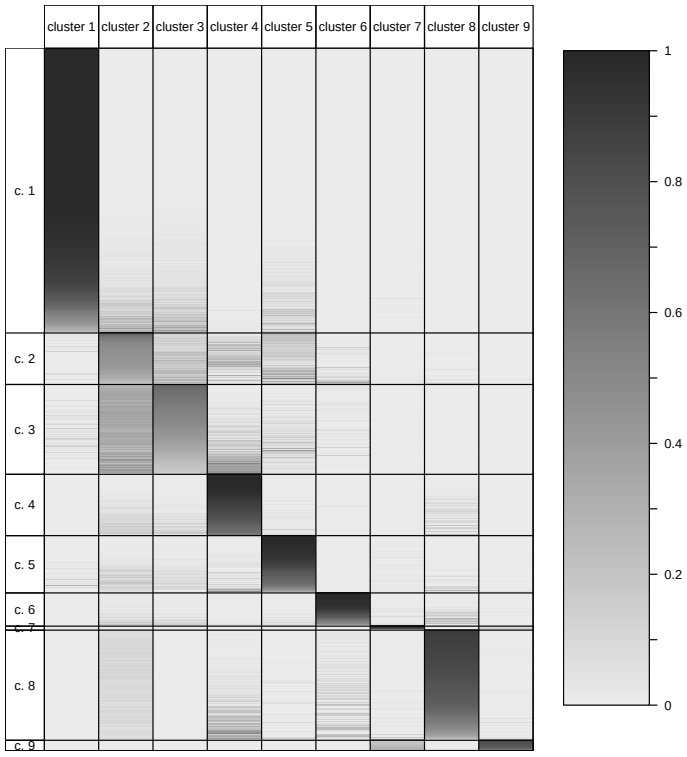

Figure 2: Graphique des probabilités conditionnelles. Le niveau de gris de la i^e ligne et la j^e colonne correspond à la probabilité d'appartenance du i^e individu à la j^e classe du modèle estimé. Les individus sont ordonnés selon leur classe d'affectation et leur probabilité d'appartenance à cette classe.