Processus de Dirichlet à mélange pour la classification non supervisée de données fonctionnelles multivariées

Asma Rabaoui[1], Hachem Kadri[2], Manuel Davy[3]

[1]Institut Fresnel, Groupe Phyti, UMR 6133 CNRS, Université Aix-Marseille, France

[2]LIF, Equipe Qarma, UMR 7279 CNRS, Université Aix-Marseille, France

[3]VEKIA SAS, LAGIS, UMR 8219 CNRS , Villeneuve d'ascq, France

asma.rabaoui@fresnel.fr, hachem.kadri@lif.univ-mrs.fr, manuel.davy@vekia.fr

Résumé − Cet article s'intéresse au clustering de données fonctionnelles multivariées par des méthodes bayésiennes non paramétriques. La plupart des travaux en analyse de données fonctionnelles s'appliquent au cas univarié où une donnée est considérée comme étant une fonction, une courbe par exemple. Cependant, dans plusieurs contextes applicatifs, un signal est caractérisé par différents descripteurs continus, et par conséquent par plusieurs fonctions pouvant être dépendantes ou indépendantes. Dans ce cas, à la différence des méthodes classiques, nous considérons un signal comme étant un vecteur de fonctions au lieu d'un vecteur discret regroupant des descripteurs de types variés et nous étendons la classification non supervisée par mélanges de processus de Dirichlet aux données fonctionnelles multivariées. Nous proposons un formalisme dédié au clustering de données fonctionnelles par des méthodes bayésiennes non paramétriques. Le formalisme permet de classer des signaux dont les descripteurs sont représentés par plusieurs fonctions qui n'ont pas forcément le même nombre de points d'évaluation et qui peuvent être observées à des instants différents. La méthodologie est illustrée sur des données simulées et des données réelles.

Abstract − This paper explores the potential use of functional data analysis (FDA) for modeling and classifying multivariate functional data. While most studies in the area of FDA have focused on univariate functions, the complexity constraints of characterizing signals in many applications impose using various feature sets and then multivariate functions. We provide a foundation for Bayesian multivariate functional data clustering using Dirichlet process mixtures. In this setting, the descriptive features of a signal are considered to be functions of time rather than a finite dimensional vector, and an extension of the Dirichlet Process Mixtures of Gaussian Process model to multivariate functional setting is proposed. As a practical test for the capabilities of the method we investigate the modeling of signals features which are a set of functions that do not have the same number of samples and are not evaluated at the same time points. We demonstrate that the model is appropriate for unsupervised classification on simulated and real multivariate functional data.

1 Introduction

Dans plusieurs applications de traitement du signal, les données sont collectées sur des grilles très fines pouvant être ainsi assimilées à des courbes ou à des surfaces (fonctions du temps ou de l'espace). L'analyse de données fonctionnelles (ADF) est un domaine de recherche très actif qui offre la possibilité d'exploiter pleinement des propriétés spécifiques aux fonctions décrivant ces données continues. Pour une introduction aux concepts et aux applications potentielles de l'analyse de données fonctionnelles, nous renvoyons le lecteur aux livres de Ramsay et Silverman [5, 6]. L'approche fonctionnelle présente clairement un certain nombre d'avantages par rapport aux méthodes vectorielles, en particulier elle permet de : (1) tenir compte des propriétés liées à la nature fonctionnelle des données (courbes lisses par exemple), (2) traiter le cas où les instants d'échantillonage sont différents d'une courbe à une autre, (3) contrôler et réduire les erreurs de mesure (observations bruitées).

La plupart des travaux en ADF s'appliquent au cas univarié, c'est-à-dire le cas où chaque donnée est considérée comme étant une fonction (une courbe par exemple). Cependant, dans plusieurs contextes applicatifs, un signal est caractérisé par différents descripteurs continus, et par conséquent, par plusieurs fonctions pouvant être dépendantes ou indépendantes. On peut citer par exemple le cas des données météorologiques où plusieurs variables climatiques, mesurées au cours du temps généralement, sont nécessaires pour caractériser un phénomène naturel observé. Aussi, en traitement de signaux audio, un signal est représenté par plusieurs descripteurs permettant d'avoir des informations variées : temporelles, spectrales et cepstrales [7].

Dans ce contexte, utilisant les méthodes classiques, un signal est considéré comme étant un vecteur de dimension finie construit en concaténant les échantillons des

différents descripteurs. Ceci a l'inconvénient de : (1) ne pas considérer les dépendances éventuelles entre les paramètres d'un descripteur, (2) ne pas différencier les paramètres provenant de deux descripteurs différents. Afin de remédier à ce problème, il est important d'étendre les méthodes fonctionnelles au cas multivarié permettant ainsi de traiter des signaux dont les descripteurs sont représentés par plusieurs fonctions pouvant être mesurées à des instants différents et qui n'ont pas forcément la même longueur.

Dans ce travail, l'objectif est de montrer comment la classification non supervisée des signaux peut être appliquée dans le cas où plusieurs variables fonctionnelles sont disponibles pour chaque signal. En particulier, nous proposons une approche bayésienne non paramétrique utilisant un processus Gaussien multivarié pour le modèle générant les données et un processus de Dirichlet à mélange pour apprendre les paramètres de chaque cluster. Outre le fait qu'une telle modélisation non paramétrique va pouvoir apporter un maximum de flexibilité et de s'affranchir de paramètres utilisateur pour le clustering de données, on peut d'ores et déjà juger de son intérêt au travers les constatations suivantes. D'un coté, si nous revenons à l'a priori sur les données fonctionnelles, on peut affirmer que notre modèle permet de traiter des données pour lesquelles les instants d'observation peuvent différer d'un individu à l'autre et l'échantillonnage peut ne pas être régulier pour chaque individu ce qui est très difficile dans l'approche classique (problème de données non directement comparables et de données manquantes). Par ailleurs, il est à préciser que le bruit gaussien additif peut être facilement intégré dans le modèle à travers les fonctions de covariance que nous détaillerons ultérieurement. En plus comme nous considérons des données pouvant être caractérisées par plusieurs fonctions, nous étendons les représentations fonctionnelles classiques à une représentation plus générale. D'un autre coté, les modèles des clusters ne sont pas contraints à prendre une forme paramétrique donnée. Nous considérons un modèle de mélange avec un nombre infini de paramètres. 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). Ce modèle peut ainsi être simplement considéré comme un modèle statistique avec un nombre infini de paramètres. Une définition alternative est un modèle dont la complexité augmente avec le nombre de données. Ceci évite ainsi de fixer la complexité ou l'ordre du modèle, le nombre de paramètres pouvant augmenter dynamiquement avec le nombre de données.

2 Formalisme

Nous proposons un formalisme bayésien pour le clustering de données fonctionnelles multivariées. Nous disposons de n signaux à classer $\{\mathbf{f}^k\}_{k=1}^n$ tel que chaque signal $\mathbf{f}^k$ est représenté par D fonctions. Dans ce qui suit, les données et l'objectif de l'analyse sont décrits de manière formelle.

2.1 Processus Gaussien Multivarié

Dans cette étude, chaque individu (signal) est caractérisé par un ensemble de D descripteurs où chaque descripteur est une fonction observée sur un nombre fini de points. Ainsi, chaque individu est représenté par plusieurs fonctions regroupées dans un vecteur que nous notons $\mathbf{f}(\cdot) = [\mathbf{f}_1(\cdot), \ldots, \mathbf{f}_D(\cdot)]$. Supposons que nous disposons des observations de la fonction $\mathbf{f}(\cdot)$ à différents instants $\{\mathbf{f}(t_p), p = 1, \cdots, P\} \in \mathbb{R}$ où $t_p \in \mathbb{R}^m$ ($m = 1$ pour les courbes et $m = 2$ pour les images). En l'absence de modèle génératif permettant de définir un a priori sur ce type de données, nous proposons d'utiliser un processus Gaussien multivarié (Multivariate Gaussian Process (MGP)) qui généralise un processus Gaussien au cas multivarié. Nous supposons donc qu'un vecteur de fonctions $\mathbf{f}(\cdot) = \{\mathbf{f}_d(\cdot)\}_{d=1}^D$ est une réalisation d'un MGP et peut ainsi s'écrire sous la forme suivante :

$$\mathbf{f}(\cdot) \sim \mathcal{MGP}(\mathbf{f}(\cdot); \mathbf{m}(\cdot), \mathbf{K}_{\mathbf{f},\mathbf{f}}(\cdot,\cdot))$$

où $\mathbf{m}(\cdot)$ est un vecteur contenant les fonctions moyennes $\{\mathbf{m}_d(\cdot)\}_{d=1}^D$ et $\mathbf{K}_{\mathbf{f},\mathbf{f}}(\cdot,\cdot)$ est une fonction de covariance, sachant que chaque fonction $\mathbf{f}_d(\cdot)$ est générée par un processus Gaussien

$$\mathbf{f}_d(\cdot) \sim \mathcal{GP}(\mathbf{f}_d(\cdot); \mathbf{m}_d(\cdot), \mathbf{K}_{\mathbf{f}_d,\mathbf{f}_d}(\cdot,\cdot)),$$

$\forall d \in \{1, \ldots, D\}$. Maintenant si on considère un nombre fini de points d'observations $\mathcal{T} = \{t_p\}_{p=1}^P$, la loi a priori sur le vecteur de fonctions $\mathbf{f}(\cdot)$ est donnée par

$$\mathbf{f} \sim \mathcal{N}(\mathbf{f}; \mathbf{m}, \mathbf{K})$$

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_D \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_D \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mathbf{f}_1,\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1,\mathbf{f}_2} & \cdots & \mathbf{K}_{\mathbf{f}_1,\mathbf{f}_D} \\ \mathbf{K}_{\mathbf{f}_2,\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2,\mathbf{f}_2} & \cdots & \mathbf{K}_{\mathbf{f}_2,\mathbf{f}_D} \\ \vdots & \vdots & \ldots & \vdots \\ \mathbf{K}_{\mathbf{f}_D,\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_D,\mathbf{f}_2} & \cdots & \mathbf{K}_{\mathbf{f}_D,\mathbf{f}_D} \end{bmatrix} \right)$$

où chaque vecteur $\mathbf{f}_d = [f_d(t_1), f_d(t_2), \cdots, f_d(t_P)]^\top$ et $\mathbf{m} = [\mathbf{m}_1^\top, \mathbf{m}_2^\top, \cdots, \mathbf{m}_D^\top]^\top$ avec $\mathbf{m}_d = [\mathbf{m}_d(t_1), \mathbf{m}_d(t_2), \cdots, \mathbf{m}_d(t_P)]^\top$. La matrice de covariance $\mathbf{K} \in \mathcal{R}^{DP \times DP}$ est une matrice par blocs où chaque bloc $\mathbf{K}_{\mathbf{f}_d,\mathbf{f}_{d'}} \in \mathcal{R}^{P \times P}$ $(d, d' = 1, \cdots, D)$ est obtenu à partir du noyau $\mathbf{K}_{\mathbf{f}_d,\mathbf{f}_{d'}}(\cdot,\cdot)$ (fonction de covariance) évalué sur l'ensemble $\mathcal{T}$. Dans le cas où les descripteurs représentant les données sont indépendants, on a $\forall t, t' \in \mathcal{T}$, $\mathbf{K}_{\mathbf{f}_d,\mathbf{f}_{d'}}(t, t') = 0$ si $d \neq d'$. Ainsi, la matrice de covariance $\mathbf{K}$ devient une matrice diagonale par blocs contenant D matrices de covariance associées aux processus Gaussiens qui ont généré les fonctions $\{\mathbf{f}_d(\cdot)\}_{d=1}^D$. Dans ce qui suit, afin de classer ces données de manière non supervisée, nous procédons à un apprentissage de la loi des clusters avec un modèle de mélange infini utilisant les processus de Dirichlet. Notons θ l'ensemble des paramètres de la fonction de covariance $\mathbf{K}_{\mathbf{f},\mathbf{f}}(\cdot,\cdot)$ que nous allons apprendre à partir des données.

2.2 Processus de Dirichlet à mélange pour le clustering

Modèle Hiérarchique. Dans des travaux récents [2, 3], les auteurs ont proposé un mélange infini de type processus de Dirichlet (DP) pour apprendre les paramètres des clusters dans le cas où chaque individu est décrit par une seule courbe. Nous généralisons cette approche aux données fonctionnelles mutilivariées. Un individu est considéré comme étant une réalisation d'un MGP, nous proposons un a priori de type DP sur les paramètres θ de ce modèle. Nous supposons ainsi un modèle hiérarchique représenté par le schéma suivant :

$$\begin{aligned} \mathbf{f}(\cdot) &\sim \mathcal{MGP}(\mathbf{f}(\cdot); \mathbf{m}(\cdot), \mathbf{K}_{\mathbf{f},\mathbf{f}}(\cdot,\cdot)) \\ \mathbf{m}(\cdot), \mathbf{K}_{\mathbf{f},\mathbf{f}}(\cdot,\cdot) &\sim \mathbb{G}(d(\mathbf{m}(\cdot), \mathbf{K}_{\mathbf{f},\mathbf{f}}(\cdot,\cdot))) \\ \mathbb{G} &\sim \mathcal{DP}(\mathbb{G}; \mathbb{G}_0, \alpha) \end{aligned} \quad (1)$$

où $\mathbf{f}(\cdot)$ suit un MGP et $\mathbb{G}$ est une réalisation d'un DP. Avec ce schéma, on ne contraint pas le modèle d'un cluster à une forme donnée qui peut être difficile à spécifier en pratique. On gagne ainsi en robustesse en considérant que la distribution inconnue a un support plus large que celui fourni par un modèle paramétrique. Ce sont les DPs que nous détaillons dans ce qui suit qui permettent une telle représentation.

Processus de Dirichlet et Inférence Bayésienne. Les DPs définissent une distribution sur l'ensemble des distributions de probabilité. Ils permettent de définir, dans un cadre bayésien, un a priori sur une distribution inconnue. Une réalisation d'un DP est presque sûrement discrète, et prend la forme dite "stick-breaking" suivante $\mathbb{G} = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$, avec $\theta_k \sim \mathbb{G}_0$, $\pi_k = \beta_k \prod_{j=1}^{k-1}(1-\beta_j)$ et $\beta_k \sim \mathcal{B}(1,\alpha)$, où $\mathcal{B}$ est une distribution Beta, α est un paramètre de précision et δ_{θ_k} est un Dirac centré en θ_k. Il en découle que cet a priori flexible peut être adapté pour représenter une densité inconnue. Les DPs possèdent des propriétés de conjugaison qui les rendent particulièrement attractifs car elles permettent la mise en oeuvre de schémas simplifiés pour réaliser l'inférence bayésienne [4]. Si on considère le modèle hiérarchique (1), dans lequel on suppose pour des raisons de simplification que $\mathbf{m} = \mathbf{0}$, alors les paramètres θ_k des clusters à estimer sont ceux de la fonction de covariance $\mathbf{K}^k_{\mathbf{f},\mathbf{f}}$, θ_k et $\mathbf{K}^k_{\mathbf{f},\mathbf{f}}$ sont respectivement le vecteur des paramètres et la fonction de covariance associés au signal $\mathbf{f}^k$ à classer. Il s'en suit d'après (1) que $\theta_k|\mathbb{G} \sim \mathbb{G}, \quad k = 1, \ldots, n$. Grâce à la représentation en urne de Polya [1], la loi conditionnelle $\theta_{k+1}|\theta_{1:k}$ peut s'écrire analytiquement comme suit :

$$\theta_k|\theta_{1:k-1}, \psi_k \sim \frac{1}{\alpha+k-1}\sum_{j=1}^{k-1}\delta_{\theta_j} + \frac{\alpha}{\alpha+k-1}\mathbb{G}_0(\varphi_k)$$

où $\psi_k = \{\alpha, \varphi_k\}$ est un vecteur d'hyperparamètres. On constate ainsi que, conditionnellement aux valeurs des variables déjà échantillonnées, un effet de clustering apparaît. Cette formulation étend les modèles de mélange finis 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. On note que $\mathcal{DP}(\mathbb{G}_0(\varphi_k), \alpha)$ peut être représentée de façon équivalente par un couple ($\mathbf{U}$, $\mathbf{z}$) représenté par des variables latentes $\mathbf{z}$ dites d'allocations et par les valeurs des clusters $\mathbf{U}$. On introduit alors la notion de label de classe z_k pour chaque hyperparamètre θ_k. On pose $\mathbf{z} = \{z_k, k = 1, \cdots, n\}$ et on désigne par $\mathcal{I}(\mathbf{z})$ l'ensemble des valeurs prises par les labels. Les valeurs des clusters sont $\mathbf{U} = \{\mathbf{U}_l, l = 1, \cdots, L\}$ tel que $\theta_k = \mathbf{U}_{z_k}$. Les $\mathbf{U}_l$ représentent les hyperparamèteres du modèle, et les z_k indiquent leurs labels correspondant tel que $z_k = l$. A partir de la représentation en Urne de Polya, les variables d'allocation sont obtenues suivant le schéma suivant :

$$z_k \sim \Pr(z_k|\mathbb{G}) = \begin{cases} \dfrac{N_{-k,l}(\mathbf{z})}{\alpha+t-1} \text{ for } l \in \mathcal{I}(\mathbf{z}) \\ \dfrac{\alpha}{\alpha+t-1} \text{ for a new } l \in \mathcal{I}(\mathbf{z}) \end{cases} \quad (2)$$

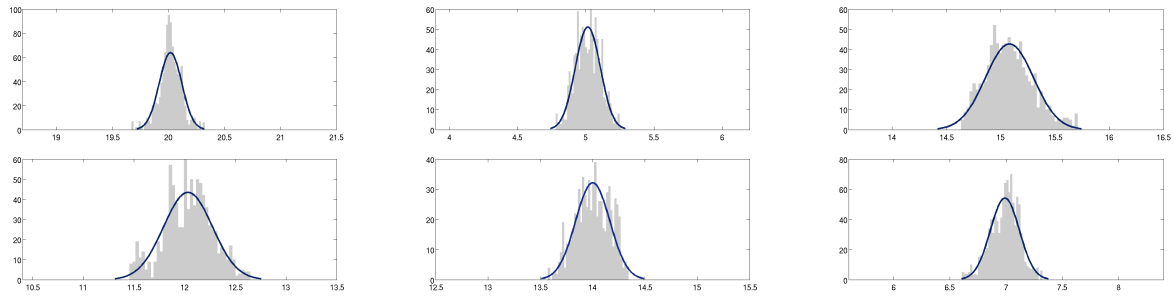
où $N_{-k,l}(\mathbf{z}) = \sum_{k'=1,k'\neq k}^{t} \delta_{l,z'_k}$ est le nombre de z'_k ($k' \neq k$) égaux à l. L'échantillonnage des θ peut se faire en deux étapes : tout d'abord les variables d'allocation et puis les valeurs des clusters.

Les propriétés des DPs permettent de définir des algorithmes de Monte Carlo par chaîne de Markov [8] pour échantillonner selon la loi a posteriori $p(\theta_1, \ldots, \theta_n|\mathbf{f}^1, \cdots, \mathbf{f}^n)$ où $\mathbf{f}^k$ est un descripteur formé par D courbes. Nous appliquons un échantillonneur de Gibbs permettant de générer des échantillons de manière séquentielle à partir des lois conditionnelles. Le calcul de la distribution a posteriori nécessite le calcul de la loi de vraisemblance qui s'écrit pour chaque $\mathbf{f}^k$ comme une loi normale multivarié

$$p(\mathbf{f}^k|\theta_k) = \mathcal{N}(\mathbf{f}^k; \mathbf{0}, \mathbf{K}^k).$$

3 Résultats expérimentaux

Données synthétiques. Pour illustrer l'intérêt de l'approche proposée, nous avons considéré un jeu de données de 60 exemples (15 exemples pour chaque classe) où chaque exemple est représenté par 3 courbes. Ces données sont générées à partir de 4 processus Gaussiens multivariés (4 classes de données) de fonctions moyennes supposées nulles et de fonctions de covariance Gaussienne à 2 paramètres. Nous avons fixé les composantes du vecteur θ à des valeurs arbitraires. Un nombre total d'itérations pour l'algorithme d'échantillonnage fixé à 2000 a permis d'atteindre la convergence. Tous les exemples ont été correctement classés. Les histogrammes reportés dans la figure 1 représentent la probabilité a posteriori des 2 composantes du vecteur θ pour chacune des



$\widetilde{\theta}_1 \approx \{20, 12\}$ $\widetilde{\theta}_2 \approx \{5, 14\}$ $\widetilde{\theta}_3 \approx \{15, 7\}$

FIGURE 1 – Les histogrammes des paramètres calculés pour chacune des 3 courbes à partir des échantillons $\theta^{(r)}$ pour $r \in \{1001, \cdots, 2000\}$.

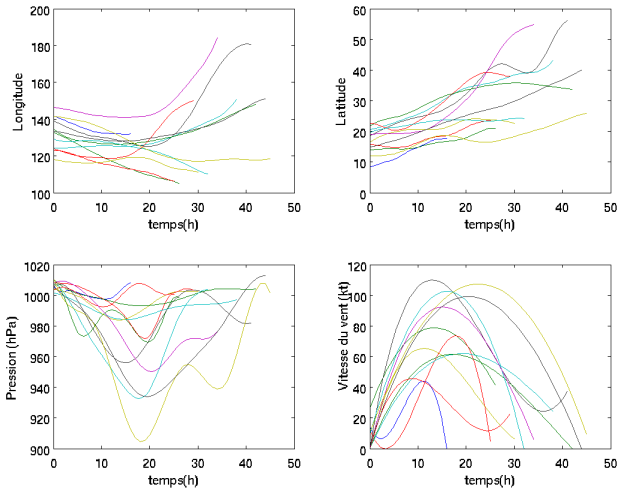


FIGURE 2 – Données mesurées au coeur des ouragans observés dans l'océan Pacifique en 2010 : longitude, latitude, pression et vitesse du vent.

3 courbes. Nous illustrons les paramètres estimés d'une classe choisie arbitrairement parmi les 4 classes.

Données réelles. Dans cette partie on s'intéresse à la classification d'ouragans selon leurs intensités. Nous utilisons des données disponibles sur le site web "Digital Typhoon : Typhoon Images et information [1]". L'ensemble des données contient des informations sur des ouragans qui ont été observés dans l'océan Pacifique depuis 1951. Un ouragan est caractérisé par 4 paramètres indiquant le positionnement (latitude et longitude), la pression atmosphérique et la vitesse du vent (voir figure 2). Chaque ouragan a été observé depuis sa génération jusqu'à sa disparition, et par conséquent on se retrouve avec des signaux qui ne sont pas mesurées aux mêmes instants et qui n'ont pas la même longueur. Contrairement aux méthodes vectorielles, notre approche est adaptée à ce genre de données et permet d'identifier correctement les 3 classes d'intensité pour ces ouragans (classe 1 : ouragan de faible intensité, classe 2 : d'intensité moyenne, classe 3 : grande intensité).

1. URL : http ://www.digital-typhoon.org/

4 Conclusion

Dans ce papier, nous avons introduit une méthode Bayésienne non-paramétrique pour la classification non-supervisée de données fonctionnelles multivariées. Les expérimentations réalisées sur des données simulées et réelles ont montré que notre méthode est bien adaptée pour le clustering de signaux décrits par plusieurs descripteurs continus et qui n'ont pas forcément le même nombre de points d'évaluation. Nos travaux futurs porteront en particulier sur l'accélération de la convergence de l'algorithme MCMC utilisé dans cet article en adoptant une approche de type Monte Carlo Séquentiel, ainsi que l'application de la méthodologie présenté ici à la classification d'autres catégories de signaux tels que les signaux musicaux.

Références

[1] D. Blackwell et J. MacQueen. *Ferguson Distributions Via Polya Urn Schemes*. The Annals of Statistics (1), 353-355, 1973.

[2] E. Jackson, M. Davy, A. Doucet, et W. Fitzgerald. *Bayesian unsupervised signal classification by dirichlet process mixtures of gaussian processes*. In ICASSP, 1077–1080, 2007.

[3] A. Rabaoui, H. Kadri, et M. Davy. *Nonparametric Bayesian Supervised Classification of Functional Data*. In ICASSP, 3381-3384, 2012.

[4] R.M. Neal. *Markov Chain Sampling Methods for Dirichlet Process Mixture Models*. Journal of Computational and Graphical Statistics (9), 249-265, 2000.

[5] J.O. Ramsay et B.W. Silverman. *Functional Data Analysis*. Springer-Verlag, New York, 2005.

[6] J.O. Ramsay et B.W. Silverman. *Applied functional data analysis*. Springer-Verlag, New York, 2002.

[7] G. Tzanetakis, et P. Cook, *Musical genre classification of audio signals*. IEEE Transactions on Speech and Audio Processing, 10(5), 293-302, 2002.

[8] A. Doucet et X. Wang. *Monte Carlo methods for signal processing : a review in the statistical signal processing context*. IEEE Signal Processing Magazine, 22(6), 152-170, 2005.