Clustering hiérarchique non paramétrique de données fonctionnelles

Marc Boullé *, Romain Guigourès *,**, Fabrice Rossi **

*Orange Labs
2 avenue Pierre Marzin
22300 Lannion
{prenom.nom}@orange.com

**SAMM, Université Paris 1
90 rue Tolbiac
75013 Paris
{prenom.nom}@univ-paris1.fr

Résumé. Dans cet article, il est question de clustering de courbes. Nous proposons une méthode non paramétrique qui segmente les courbes en clusters et discrétise en intervalles les variables continues décrivant les points de la courbe. Le produit cartésien de ces partitions forme une grille de données qui est inférée en utilisant une approche Bayésienne de sélection de modèle ne faisant aucune hypothèse concernant les courbes. Enfin, une technique de post-traitement, visant à réduire le nombre de clusters dans le but d'améliorer l'interprétabilité des clusters, est proposée. Elle consiste à fusionner successivement et de façon optimale les clusters, ce qui revient à réaliser une classification hiérarchique ascendante dont la mesure de dissimilarité correspond à la variation du critère. De manière intéressante, cette mesure est en fait une somme pondérée de divergences de Kullback-Leibler entre les distributions des clusters avant et après fusions. L'intérêt de l'approche dans le cadre de l'analyse exploratoire de données fonctionnelles est illustré par un jeu de données artificiel et réel.

1 Introduction

En analyse de données fonctionnelles (Ramsay et Silverman (2005)), les observations sont des fonctions (ou des courbes). Les données fonctionnelles sont présentes dans de nombreux domaines comme par exemple l'enregistrement des précipitations d'une station météorologique ou encore la surveillance de matériel, où chaque courbe est une série temporelle liée à une quantité physique enregistrée à fréquence spécifiée.
Les Méthodes d'analyse exploratoire pour les grandes bases de données fonctionnelles sont nécessaires dans de nombreuses applications pratiques comme par exemple, la surveillance de la consommation électrique (Hébrail et al. (2010)). Elles réduisent la complexité des données en combinant des techniques de clustering avec des méthodes d'approximation de fonction, modélisant par exemple un ensemble de données fonctionnelles par des courbes prototypiques, comme par exemple un ensemble de segments linéaires ou de splines. Dans ce type d'approches, à la fois le nombre de prototypes et le nombre de segments sont des paramètres utilisateur. D'un côté, cela limite pour l'utilisateur le risque d'obtenir des clusters trop complexes mais cela peut également induire un sous-apprentissage du modèle par rapport aux données.
Des approches Bayésiennes non paramétriques basées sur des processus de Dirichlet ont

également été appliquées au problème de clustering de courbes. Elles cherchent à déterminer une distribution de clustering sur un modèle infini de mélanges (Teh, 2010; Nguyen et Gelfand, 2011). Le modèle de clustering est obtenu en échantillonnant la distribution a posteriori par des méthodes d'inférence Bayésienne.
Cet article propose une nouvelle méthode d'analyse exploratoire non paramétrique de données fonctionnelles, basée sur les modèles en grilles (Boullé, 2010). La méthode ne fait d'hypothèse ni sur la distribution des données fonctionnelles ni sur le bruit lié aux mesures. La méthode ne requière pas de paramètre utilisateur et permet d'obtenir, de manière totalement autonome, un résumé optimal du jeux de données fonctionnelles, en utilisant une approche MAP (Maximum a posteriori) avec des termes de priors dépendants des données. Dans certains cas, particulièrement pour les jeux de données volumineux, le nombre optimal de clusters et d'intervalles peut être très important pour approximer finement les données, ce grand nombre de clusters se prêtant mal à une interprétation utilisateur en analyse exploratoire. C'est pourquoi un post-traitement est associé à la méthode. Ceci permet de réduire le nombre de clusters selon une procédure optimale sous condition d'emboîtement. Le nombre d'intervalles, quant à lui, est auto ajusté parallèlement au nombre de clusters.
Le post-traitement consiste à fusionner successivement les clusters de la manière la moins coûteuse, depuis le clustering le plus fin jusqu'à obtenir un unique cluster contenant toutes les courbes. Il apparait que le coût de fusion de deux clusters est une somme pondérée de divergences de Kullback-Leibler des clusters fusionnés au cluster formé, ce qui peut s'interpréter comme une mesure de dissimilarité entre les clusters fusionnés. Ainsi, le post-traitement peut être vu comme une classification hiérarchique ascendante Hastie et al. (2001). Des outils d'aide à la décision peuvent être utilisés, comme par exemple un Dendrogramme ou encore la courbe de Pareto du coût du modèle en fonction du nombre de clusters. Le reste de l'article est construit de la façon suivante : la section 2 présente le problème de clustering de courbes et positionne notre méthode par rapport à des approches alternatives. Ensuite, dans la Section 3, la méthode de clustering basée sur l'estimation de densité jointe est présentée. Puis, la technique de post-traitement est détaillée dans la Section 4. Enfin, des expérimentations sur des jeux de données artificiels et réels sont présentées dans la Section 5, avant de conclure par la Section 6.

2 Analyse exploratoire de données fonctionnelles

Dans cette Section, les données et l'objectif d'analyse sont décrits de manière formelle. Soit $\mathcal{C}$ un ensemble de n courbes ou fonctions, $c_i, 1 \leq i \leq n$, définies sur $[a, b]$ vers $[u, v]$, deux intervalles dans $\mathbb{R}$. Chaque courbe est composée de m_i valeurs, formant une série d'observations notées $c_i = (x_{ij}, y_{ij})_{j=1}^{m_i}$, avec $y_{ij} = c_i(x_{ij})$.

Comme dans toutes les problématiques d'analyse exploratoire, notre but ici est de réduire la complexité du jeux de données et de découvrir des motifs dans les données. Dans Chamroukhi et al. (2010); Hébrail et al. (2010), les motifs fonctionnels sont de simples fonctions telles que des fonctions indicatrices d'intervalles ou des polynômes simples : une fonction est approximée par une combinaison linéaire de ces fonctions simples dans Hébrail et al. (2010) ou générée par un processus logistique basé sur des polynômes de bas degrés dans Chamroukhi et al. (2010). Des B-splines peuvent également être utilisées, comme dans Abraham et al. (2003).

Notons k_C le nombre de clusters de courbes. La méthode, proposée par Hébrail et al. (2010), détermine une partition de l'ensemble des courbes $\mathcal{C}$ en k_C clusters modélisés par une

fonction simple $\mathcal{F}_k$ (fonction constante par morceaux composée de $\mathcal{P}$ segments par exemple), en minimisant

$$\sum_{k=1}^{k_C} \sum_{c_i \in \mathcal{C}_k} \sum_{j=1}^{m_i} (y_{ij} - f_k(x_{ij}))^2 , \tag{1}$$

avec $\mathcal{C}_k$ le $k^{ième}$ cluster. Ceci correspond à une sorte de k-means contraint par le choix des segments dans l'espace fonctionnel L^2. L'approche de Chamroukhi et al. (2010) optimise un critère similaire obtenu par maximisation de la vraisemblance des paramètres d'un modèle génératif.

Des approches Bayésiennes, comme celle présentée dans Nguyen et Gelfand (2011), considèrent que l'ensemble des courbes peut être représenté par des courbes moyennes générées suivant un processus Gaussien et organisées en clusters. Les clusters sont décrits par une fonction d'étiquetage qui suit la réalisation d'une distribution multinomiale avec un prior de Dirichlet. Alors que les modèles paramétriques utilisant un nombre fixe et fini de paramètres peuvent souffrir de sur/sous-apprentissage, des approches Bayésiennes non-paramétriques ont été proposées pour éviter ce problème. En utilisant un modèle de complexité non bornée, le sous-apprentissage est atténué, alors que l'approche Bayésienne de calcul ou d'approximation de la distribution a posteriori des paramètres réduit le risque de sur-apprentissage (Teh, 2010). Au final, la distribution des paramètres de clustering est obtenue en échantillonnant la distribution a posteriori des paramètres en utilisant des méthodes d'inférence Bayésienne comme les Chaines de Markov Monte Carlo (Neal, 2000) ou l'inférence variationnelle (Blei et Jordan, 2005). Suit un post-traitement permettant de choisir un clustering parmi leur distribution.

Le prior de Dirichlet nécessite deux paramètres utilisateur : le paramètre de concentration et la distribution de base. Pour un paramètre de concentration α et un jeu de données contenant n courbes, l'espérance du nombre de clusters $\bar{k}$ est $\bar{k} = \alpha \log(n)$ (Wallach et al., 2010). De ce fait, le paramètre de concentration a un impact significatif sur le nombre de clusters obtenus. Pour cette raison que, selon Vogt et al. (2010), il n'est pas possible d'estimer de manière fiable ce paramètre.

Notre méthode - appelée MODL et détaillée en Section 3 - est comparable aux approches basées sur les Processus de Dirichlet (DP) dans le sens où elles estiment une probabilité a posteriori basée sur la vraisemblance et la distribution a priori des paramètres d'un modèle. Les méthodes sont également non-paramétriques et de complexité non bornée, puisque le nombre de paramètres n'est pas fixé et croit en fonction de la quantité de données disponibles.

Cependant, MODL est intrinsèquement différent des méthodes basées sur les DP. D'abord, les approches basées sur les DP sont Bayésiennes et génèrent une distribution de clusterings, le clustering final étant obtenu par post-traitement consistant, par exemple, à choisir le mode de la distribution a posteriori ou encore en étudiant la matrice des co-occurrences. A contrario, MODL est une approche MAP, le modèle le plus probable est directement sélectionné en utilisant des algorithmes d'optimisation. Ensuite, MODL ne s'applique pas aux valeurs des données mais à leur rang. Ceci permet d'éviter les valeurs aberrantes et les problèmes d'échelle. En utilisant la statistique d'ordre, les modèles obtenus sont invariants par toute transformation monotone des données d'entrée, ce qui a du sens puisque la méthode se focalise sur les corrélations entre les variables, pas sur les valeurs des variables.

Ensuite, les méthodes basées sur les DP étudient la distribution des paramètres définis sur $\mathcal{R}$, dont la mesure est par conséquent continue. Quant à MODL, les corrélations entre les variables sont modélisées sur un échantillon. Dans le cas de clusters de courbes, ces variables sont le

point de mesure, la valeur de la courbe en ce point et l'identifiant de la courbe. Ceci permet de travailler sur un espace discret et donc de simplifier le réalisation du problème, qui se résume ainsi principalement à un problème de dénombrement.

Au final, l'approche MODL est clairement dépendante des données. Dans une première étape, l'échantillon de données est utilisé pour construire le prior et l'espace de modélisation en regardant les variables de façon indépendantes : seuls la taille de l'échantillon et les valeurs (rangs empiriques) de chaque variables sont exploitées. Le modèle de corrélation est inferré dans une seconde phase, en utilisant une approche MAP. Par conséquent, prouver la consistance de cette technique de modélisation dépendante des données demeure un problème ouvert. En fait, des résultats expérimentaux obtenant à la fois des motifs précis et fiables montrent la cohérence de la méthode.

3 L'approche MODL appliquée aux données fonctionnelles

Dans cette Section, les principes des modèles en grilles de données, détaillés dans Boullé (2010), sont résumés et appliqués aux données fonctionnelles.

3.1 Modèles en grille de données

Les modèles en grilles de données sont basés sur le partitionnement de chaque variable en intervalles dans le cas numérique et en groupement de valeurs dans le cas catégoriel. Le produit Cartésien des partitions univariées forme une partition multivariée de l'espace de représentation dans un ensemble de cellules. Cette partition multivariée, appelée grille de données, est un estimateur non paramétrique et constant par morceaux de la probabilité jointe ou conditionnelle. La meilleure grille de données est obtenue en utilisant une approche Bayésienne de sélection de modèle et des algorithmes combinatoires efficaces.

3.2 Application aux données fonctionnelles

L'ensemble $\mathcal{C}$ des n courbes est représenté par un jeu de données contenant $m = \sum_{i=1}^{n} m_i$ instances et trois variables : C représentant l'identifiant de la courbe, X et Y les coordonnées des points des courbes. Les modèles en grilles sont appliqués pour estimer la probabilité jointe $p(C, X, Y)$ entre les trois variables. La variable C est segmentée en clusters de courbes, alors que chacune des variables continues X et Y est discrétisée en intervalles. Le produit Cartésien de ces partitions univariées forme une grille de données. Comme $p(X, Y|C) = \frac{p(C,X,Y)}{p(C)}$, la méthode peut aussi être interprétée comme un estimateur de densité jointe entre les variables continues de coordonnées des points (X et Y), qui est constante par cluster de courbes.

La Definition 3.1 introduit la notion de modèle de clustering de données fonctionnelles.

Definition 3.1. *Un modèle de clustering de données fonctionnelles est défini par :*
- *un nombre de clusters de courbes,*
- *un nombre d'intervalles des variables continues X et Y,*
- *la répartition des courbes dans les clusters,*
- *la distribution des points du jeu de données sur les cellules de la grilles de données.*
- *la distribution des points de chaque cluster sur les courbes de ce même cluster.*

Notation.
- $\mathcal{C}$ *: ensemble des courbes, de taille* $n = |\mathcal{C}|$.
- $\mathcal{P}$ *: ensemble des points définis sur 3 dimensions formant* $\mathcal{C}$ *de taille* $m = |\mathcal{P}|$.
- C *: variable identifiant la courbe*
- X, Y *: variables de coordonnées des points*
- k_C *: nombre de clusters de courbes*
- k_X, k_Y *: nombre d'intervalles des variables* X *and* Y
- $k = k_C k_X k_Y$ *: nombre de cellules de la grille de données*
- n_{i_C} *: nombre de courbes du cluster* i_C
- m_i *: nombre de points de la courbe* i
- m_{i_C} *: nombre de points dans le cluster* i_C
- m_{j_X}, m_{j_Y} *: nombre de points dans les intervalles* j_X *de* X *et* j_Y *de* Y
- $m_{i_C j_X j_Y}$ *: nombre de points de la cellule* (i_C, j_X, j_Y).

Nous considérons que les nombres de courbes n et de points m sont connus par avance et souhaitons modéliser la distribution jointe des m points sur leurs courbes et coordonnées associées. Pour sélectionner le meilleur modèle, nous employons une approche MAP, utilisant la distribution a priori des paramètres du modèle décrit dans la Définition 3.2.

Definition 3.2. *L'a priori sur les paramètres d'un modèle de clustering de données fonctionnelles est choisi hiérarchiquement et uniformément à chaque niveau :*
- *les nombres de clusters* k_C *et d'intervalles* k_X, k_Y *sont indépendants les uns des autres, et uniformément distribués entre* 1 *et* n *pour les courbes et entre* 1 *et* m *pour les coordonnées* X *et* Y*,*
- *pour un nombre* k_C *de clusters donné, toutes les partitions des* n *courbes en* k_C *clusters sont équiprobables,*
- *pour un modèle de taille* (k_C, k_X, k_Y)*, toutes les distributions des* m *points sur les* $k = k_C k_X k_Y$ *cellules de la grille de données sont équiprobables,*
- *pour un cluster de courbes donné, toutes les distributions des points sur les courbes au sein du cluster sont equiprobables,*
- *pour un intervalle donné* X *(resp.* Y*), toutes les distributions des rangs des valeurs des points sur* X *(resp.* Y*) sont equiprobables.*

En prenant le logarithme négatif de la probabilité a posteriori d'un modèle connaissant les données, un critère d'évaluation est obtenu et est donné par le théorème 3.3, qui est adapté aux données fonctionnelles.

Theorem 3.3. *Un modèle* M *de clustering de données fonctionnelles distribué selon un a priori hiérarchique uniforme est optimal au sens de Bayes si la valeur du critère suivant est minimal*

$$
\begin{aligned}
c(M) &= -\,log(P(M)) - log(P(\mathcal{P}|M)) \\
&= \log n + 2\log m + \log B(n, k_C) + \log \binom{m+k-1}{k-1} + \sum_{i_C=1}^{k_C} \log \binom{m_{i_C} + n_{i_C} - 1}{n_{i_C} - 1} \\
&\quad + \log m! - \sum_{i_C=1}^{k_C} \sum_{j_X=1}^{k_X} \sum_{j_Y=1}^{k_Y} \log m_{i_C j_X j_Y}! + \sum_{i_C=1}^{k_C} \log m_{i_C}! - \sum_{i=1}^{n} \log m_i! \\
&\quad + \sum_{j_X=1}^{k_X} \log m_{j_X}! + \sum_{j_Y=1}^{k_Y} \log m_{j_Y}!
\end{aligned}
\tag{2}
$$

$B(n,k)$ est le nombre de partitions de n éléments en k sous parties (avec éventuellement des parties vides). Lorsque $n = k$, $B(n,k)$ correspond au nombre de Bell. Dans le cas général, $B(n,k)$ peut être écrit $B(n,k) = \sum_{i=1}^{k} S(n,i)$, où $S(n,i)$ est le nombre de Stirling de seconde espèce (Abramowitz et Stegun, 1970), ce qui revient au nombre de façons de partitionner un ensemble de n éléments en i sous parties non vides.

Étant donné qu'un logarithme négatif de probabilités est une longueur de codage (Shannon, 1948), la technique de sélection de modèle est similaire à l'approche MDL (Minimum Description Length) introduite par Rissanen (1978). La première ligne de la Formule 2 correspond à la distribution a priori des nombres de clusters k_C et d'intervalles k_X et k_Y et au partitionnement des courbes en clusters. La seconde ligne représente la spécification des paramètres de la distribution multinomiale des m points dans les k cellules de la grille de données, ainsi que des points de chaque cluster sur les courbes de ce même cluster. La troisième ligne correspond à la vraisemblance de la distribution des points sur les cellules. Quant à la dernière ligne, il s'agit de la vraisemblance de la distribution des points de chaque cluster sur les courbes de ce cluster, suivie par la vraisemblance de la distribution des rangs des valeurs de X (resp. Y) dans chaque intervalle.

3.3 Algorithme d'optimisation

Dans cet article, des heuristiques d'optimisation (détaillées dans Boullé (2010)) ont été utilisées. Elles possèdent des propriétés de scalabilité, avec une complexité spatiale en $O(m)$ et temporelle en $\mathrm{O}(m\sqrt{m}\log m)$. L'heuristique principale est une heuristique gloutonne ascendante, qui démarre avec un modèle fin avec peu de points par intervalle de X et Y et peu de courbes par cluster. Elle considère toutes les fusions entre les clusters et les intervalles adjacents, et réalise la meilleure fusion si cela permet de faire décroitre le critère. Cette heuristique est améliorée avec une étape de post-optimisation (déplacement des bornes des intervalles et changement de clusters pour les courbes) et englobée dans une métaheuristique de type VNS (Hansen et Mladenovic, 2001) qui tire profit de plusieurs lancements de l'algorithme avec des initialisations aléatoires différentes.
L'algorithme d'optimisation résumé ci-dessus a été évalué dans de nombreux cas de figures dans Boullé (2010) où la vraie distribution sous-jacente est connue. Au final, la méthode est à la fois résistante au bruit et capable de détecter des motifs fins et complexes. Elle est en mesure d'approximer n'importe quelle distribution de données, sous condition d'avoir suffisamment d'instances dans l'ensemble d'apprentissage.

4 Clustering Hierarchique Ascendant

Alors que le modèle obtenu par la méthode détaillée dans la section 3 est optimal selon le critère introduit par le théorème 3.3, nous proposons ici une technique de post-traitement qui vise à simplifier le clustering tout en minimisant la perte d'informations. Dans un premier temps, l'impact d'une fusion sur le critère est étudié, puis les propriétés de la mesure de dissimilarité proposée sont détaillées et enfin la méthode de clustering hiérarchique ascendant est décrite. Notons que les paramètres de modélisation utilisés pour la construction du clustering initial et pour la fusion des clusters dans la méthode agglomérative sont les mêmes.

4.1 Le Coût de Fusion de deux Clusters

Soient $M_{1_C,2_C}$ et M_{γ_C} deux modèles de clustering, le premier correspondant au modèle MAP avant fusion des clusters 1_C et 2_C, le second au modèle après fusion, contenant un nouveau cluster $\gamma_C = 1_C \cup 2_C$. Nous notons $\Delta c(1_C, 2_C)$ le coût de fusion de 1_C et 2_C, défini comme :

$$\Delta c(1_C, 2_C) = c(M_{\gamma_C}) - c(M_{1_C,2_C})$$

Il résulte du Theorème 3.3 que le modèle de clustering M_{γ_C} est une explication moins probable selon MODL du jeu de données $\mathcal{P}$ que $M_{1_C,2_C}$ suivant un facteur basé sur $\Delta c(1_C, 2_C)$.

$$p(M_{\gamma_C}|\mathcal{P}) = e^{-\Delta c(1_C,2_C)} p(M_{1_C,2_C}|\mathcal{P}) \tag{3}$$

Nous nous focalisons maintenant sur le comportement asymptotique de $\Delta c(1_C, 2_C)$, c'est-à-dire lorsque le nombre de points m du jeu de données tend vers l'infini.

Theorem 4.1. *La variation du critère est asymptotiquement égale à une somme pondérée de divergences de Kullback-Leibler des clusters 1_C et 2_C à γ_C, estimée sur la discrétisation bivariée $k_X \times k_Y$.*

$$\Delta c(1_C, 2_C) = m_{1_C} D_{KL}(1_C||\gamma_C) + m_{2_C} D_{KL}(2_C||\gamma_C) + O(log(m_{\gamma_C})) \tag{4}$$

La preuve complète n'est pas détaillée ici pour des raisons de concision. Pour résumer, le calcul de $\Delta c(1_C, 2_C)$ élimine certains termes de prior (les deux premières lignes de la Formule 2) et borne les autres par $O(log(m_{\gamma_C}))$. Ensuite, en utilisant l'approximation de Stirling $log(m!) = m(log(m) - 1) + O(log(m))$, la variation de la vraisemblance (les deux dernières lignes de la Formule 2) peuvent être réécrites comme une somme pondérée de divergences de Kullback-Leibler.

La variation du critère, liée à la fusion de deux clusters, est basée sur des divergences de Kullback-Leibler. Cette mesure est non-symétrique et caractérise la différence entre deux distributions (Cover et Thomas, 1991). Le principe du clustering hiérarchique ascendant est de fusionner successivement les clusters dans le but de construire un arbre appelé dendrogramme. Nous le construisons ici en utilisant la variation du critère Δc. De part les propriétés de cette mesure de dissimilarité, le dendrogramme que nous construisons est équilibré. En effet, étant donné que nous faisons un compromis entre la fusion de deux clusters similairement distribués et la fusion d'un petit cluster avec un gros lors de la classification, nous obtenons des clusters de tailles comparables à chaque niveau hiérarchique du dendrogramme. Notons que lors du processus agglomératif, la meilleure fusion peut aussi bien être effectuée sur les clusters que sur les intervalles des variables X et Y. Ainsi, la granularité de la représentation des courbes se dégradera en même temps que le nombre de clusters diminuera.

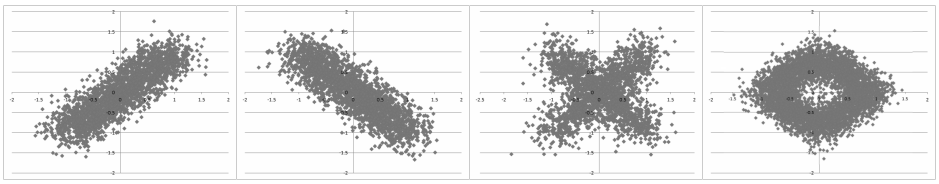
5 Expérimentations

Dans cette partie, les propriétés de notre approche sont dans un premier temps illustrées en utilisant un jeu de données artificiel. Ensuite la méthode est appliquée à un jeu de données réel, puis les clusters sont fusionnés successivement et enfin des exemples d'analyses exploratoires sont présentés.

5.1 Expérimentations sur un jeu de données artificiel

Une variable z est échantillonnée suivant une loi uniforme : $Z \sim \mathcal{U}(-1, 1)$. Nous notons ε_i un bruit blanc Gaussien : $E \sim \mathcal{N}(0, 0.25)$. Soient les distributions suivantes :

- $f_1 : x = z + \varepsilon_x$, $y = z + \varepsilon_y$
- $f_2 : x = z + \varepsilon_x$, $y = -z + \varepsilon_y$
- $f_3 : x = z + \varepsilon_x$, $y = \alpha z + \varepsilon_y$ avec $\alpha \in \{-1, 1\}$ et $p(\alpha = -1) = p(\alpha = 1)$
- $f_4 : x = (0.75 + \varepsilon_x)cos(\pi(1 + z))$, $y = (0.75 + \varepsilon_y)sin(\pi(1 + z))$



(a) f_1 (b) f_2 (c) f_3 (d) f_4

FIG. 1: Distributions générées aléatoirement

Un ensemble de 40 courbes (10 par distribution) est généré en utilisant les distributions précédemment définies. Un ensemble $\mathcal{P}$ de 10^5 points est également généré. Chaque point est un triplet de valeurs avec un identifiant de courbe choisi parmi les 40, une valeur de x et de y générée suivant la distribution correspondant à la courbe choisie.
Nous appliquons notre méthode de clustering de données fonctionnelles introduite en Section 3 sur des sous-ensembles de $\mathcal{P}$ de tailles croissantes. L'expérience est renouvelée 10 fois pour chaque sous-ensemble de points ré-échantillonné à chaque fois. Pour les petits sous-ensembles (en-dessous de 400 points), il n'y pas suffisamment de données pour découvrir des motifs significatifs, et notre méthode produit un unique cluster contenant toutes les courbes et un seul intervalle pour X et Y. À partir de 400 points, le nombre de clusters et d'intervalles commence à croitre. Finalement pour 25 points par courbe en moyenne, c'est-à-dire 1000 points au total, notre méthode retrouve les motifs sous-jacents et génère 4 clusters de courbes qui correspondent aux distributions f_1, f_2, f_3 et f_4.

Bien que la méthode retrouve le véritable nombre de clusters, en-dessous de 2000 points, les clusters peuvent ne pas être totalement purs et certaines courbes mal placées. Dans nos expériences, pour 1000 points, en moyenne 2% des courbes sont mal placées, alors qu'avec 2000 points, les courbes sont systématiquement classées dans le bon cluster.
Notons qu'en augmentant la taille du sous-ensemble de points au-delà de 2000 points, le nombre de motifs obtenu est constant et égal à 4. A contrario, le nombre d'intervalles croit avec le nombre de points. Ceci montre le bon comportement asymptotique de la méthode : on retrouve le bon nombre de motifs et la méthode exploite la quantité croissante des données pour mieux approximer la forme des motifs.

Cette expérience permet de mettre en évidence une propriété intéressante : la méthode ne nécessite pas que la position des valeurs de la variable x soit la même pour toutes les courbes. De plus, au delà d'un clustering de courbes, la méthode s'applique aux distributions. Ainsi il est

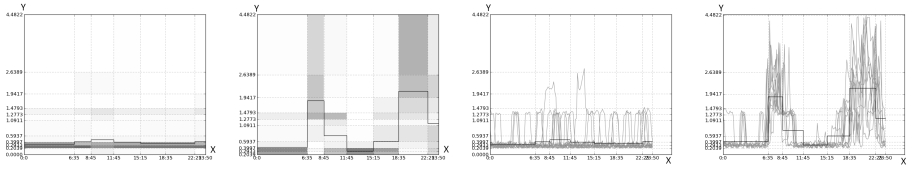
possible de détecter des clusters de distributions multimodales comme celles générées par f_3 et f_4.

5.2 Analyse d'un jeu de données de consommation électrique

Le jeu de données est un enregistrement de la consommation électrique d'un foyer français pendant un an (données disponibles pour 349 jours) (Hébrail et al., 2010). Chaque courbe est composée de 144 mesures qui donnent la consommation électrique journalière enregistrée toutes les 10 minutes. Il y a au total 50256 points de mesure et 3 variables : la mesure temporelle X, la consommation électrique Y et l'identifiant de la journée C. L'étude a pour but de regrouper les jours ayant le même profil de consommation électrique.

La Grille Optimale. La grille optimale est définie par 57 clusters, 7 intervalles sur X et 10 sur Y. Cela signifie que les 349 courbes ont été classées dans 57 clusters, chaque jour a été discrétisé en 7 plages horaires et 10 plages de consommation électrique. Ce résultat permet de déterminer des profils caractéristiques de jours, tels que les jours de travail, les jours chômés ou encore les jours où personne n'est au domicile.

Les prototypes moyens, représentés par des fonctions continues par morceaux, permettent d'apprécier la consommation moyenne par plage horaire. La probabilité conditionnelle des intervalles de consommation sachant les plages horaires est représentée par des cellules grisées, où le niveau de gris modélise la probabilité conditionnelle associée à la cellule. La première représentation a été choisie pour simplifier l'interprétation des clusters de courbes, alors que la seconde permet de détecter des multimodalités dans les plages horaires.



(a) (b) (c) (d)

FIG. 2: Deux exemples parmi les 57 clusters, les traits représentent les prototypes et les cellules grisées les probabilités conditionnelles. La Figure(a) représente le plus gros cluster, caractéristique des jours où personne n'est à domicile : la consommation est quasi constante et très basse. La Figure (b), qui représente le second plus gros cluster, montre un jour de travail avec une consommation basse pendant la nuit et les heures de bureau et avec des pics de consommation le matin et le soir. Les Figure (c) et (d) sont les prototypes et les courbes des clusters des Figures (a) et (b)

Distributions multimodales. La Figure 2.(b) présente une multimodalité pour la 3^{eme} plage horaire : le prototype est situé entre deux cellules denses. Cela signifie que la plupart des mesures de consommation électrique ont été prises dans ces deux modalités et rarement dans l'intervalle dans lequel passe le prototype. Ceci est illustré par la Figure 2.(d). Notons que la Figure 2.(a) présente une autre illustration de distribution multimodale pour laquelle les

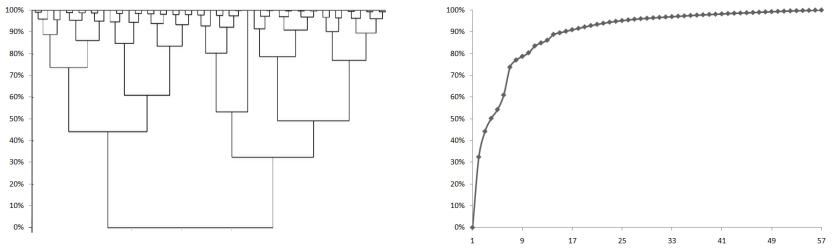
points sont majoritaires dans la modalité inférieure. De manière générale, la méthode étend le clustering de courbes au clustering de distributions.

Fusion des clusters. Alors que le clustering le plus fin produit un clustering riche avec des clusters caractéristiques précis, une étude plus synthétique et plus interprétable de la consommation électrique annuelle peut être souhaitable pour certaines applications. C'est pourquoi des fusions successives ont été effectuées et représentées sur la Figure 3 par un dendrogramme et une courbe de Pareto présentant le pourcentage d'information conservée en fonction du nombre de clusters.

Definition 5.1. *Soit $M_\emptyset$ le modèle nul avec un cluster de courbes et un intervalle temporel et de consommation. La grille de données n'est composée que d'une unique cellule. Ses propriétés sont détaillées dans Boullé (2008). M_{opt} est le modèle optimal selon le critère optimisé défini dans le Théorème 3.3 et M_k résultant des fusions successives jusqu'à k clusters. Le pourcentage τ_k d'information conservée pour k clusters est défini par*

$$\tau_k = \frac{c(M_k) - c(M_\emptyset)}{c(M_{opt}) - c(M_\emptyset)}$$

Le dendrogramme est équilibré et la courbe de Pareto est concave, ce qui permet de diviser par trois le nombre de clusters en gardant 90% de l'information initiale.



(a) (b)

FIG. 3: Dendrogramme et courbe de Pareto de la quantité d'information conservée en fonction du nombre de clusters.

Des expérimentations détaillées ont été réalisées à différents niveaux de hiérarchie. Ici nous étudions le cas d'une grille de données simplifiée où 4 clusters et 50% de l'information ont été conservés. En affichant le calendrier avec différentes couleurs pour les 4 clusters, il est possible de mettre en évidence une certaine saisonnalité comme l'illustre la Figure 4. En effet, la manière dont les courbes ont été groupées présente un lien entre la météo et les températures en France cette année là. Deux clusters caractérisent la saison estivale (de Juin à Septembre) où la consommation électrique est plus basse. Le reste de l'année, les températures sont plus basses et donc la consommation électrique plus importante. La période de fin Avril à début Mai était particulièrement chaude cette année là, ce qui explique qu'elle ait été classée avec les clusters estivaux. De manière intéressante, les clusters de la Figure 2.(a) où personne n'était au domicile sont regroupés avec les clusters d'été. On les retrouve du 23 Février au 2 Mars et du 29 Octobre au 3 Novembre.

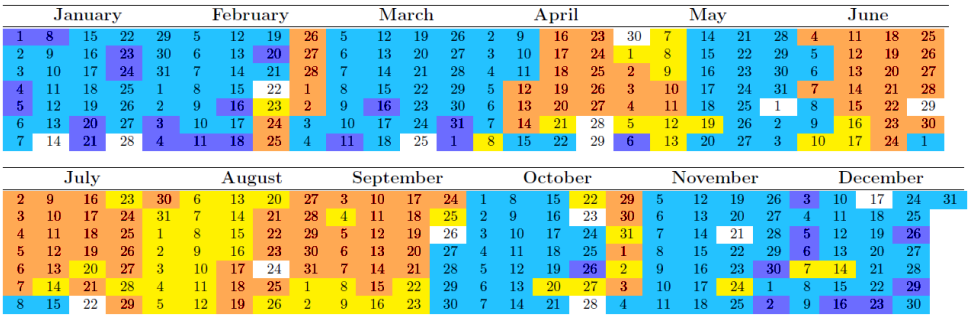


FIG. 4: Calendrier de l'année 2007. Chaque ligne représente un jour de la semaine. Il y a 4 couleurs (une par cluster), les jours blancs correspondent aux données manquantes.

6 Conclusion

Dans cet article, nous nous sommes concentrés sur l'analyse exploratoire de données fonctionnelles, et plus particulièrement de clustering de courbes. La méthode proposée ne considère pas un ensemble de courbes mais de points décrits par trois variables, deux continues, la position du point et une catégorielle, l'identifiant de la courbe. En groupant les courbes et en discrétisant les variables continues en sélectionnant le meilleur modèle selon une approche MAP, la méthode se comporte comme un estimateur non paramétrique de densité jointe à la fois des courbes et des coordonnées des points. Dans le cas de données volumineuses, le meilleur modèle tend à être trop précis pour en faire une interprétation simple. Pour éviter ce problème, un post-traitement est proposé. Cette technique a pour but de fusionner successivement les clusters jusqu'à obtenir un clustering simplifié en perdant le moins de précision possible. Ce processus est équivalent à un clustering hiérarchique ascendant dont la mesure de dissimilarité serait la variation du critère, qui correspond à une somme pondérée de divergences de Kullback-Leibler des clusters fusionnés au cluster généré. Des expérimentations ont été menées sur un jeu de données réel, la consommation électrique annuelle d'un foyer. D'un côté, le clustering le plus fin met en évidence des phénomènes intéressants tels des distributions multimodales pour certaines plages horaires. Quant au post-traitement, un dendrogramme équilibré et une courbe de Pareto concave soulignent la possibilité de simplifier le modèle le plus fin en perdant un minimum d'information, et ainsi d'obtenir un clustering plus interprétable. Au-delà du clustering de courbes, la méthode proposée est capable de générer un clustering de distributions. Dans de prochains travaux, il est prévu d'étendre la méthode aux distributions multidimensionnelles en considérant plus de deux dimensions.

Références

Abraham, C., P. Cornillon, E. Matzner-Løbe, et N. Molinari (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics 30*(3), 581–595.

Abramowitz, M. et I. Stegun (1970). *Handbook of mathematical functions*. New York : Dover Publications Inc.

Blei, D. M. et M. I. Jordan (2005). Variational inference for dirichlet process mixtures. *Bayesian Analysis 1*, 121–144.

Boullé, M. (2008). Multivariate data grid models for supervised and unsupervised learning. Technical Report NSM/R&D/TECH/EASY/TSI/5/MB, France Telecom R&D. http ://perso.rd.francetelecom.fr/-boulle/publications/BoulleNTTSI5MB08.pdf.

Boullé, M. (2010). Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hands on pattern recognition*. Microtome. in press.

Chamroukhi, F., A. Samé, G. Govaert, et P. Aknin (2010). A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing 73*(7-9), 1210–1221.

Cover, T. et J. Thomas (1991). *Elements of information theory*. New York, NY, USA : Wiley-Interscience.

Hansen, P. et N. Mladenovic (2001). Variable neighborhood search : principles and applications. *European Journal of Operational Research 130*, 449–467.

Hastie, T., R. Tibshirani, et J. Friedman (2001). *The elements of statistical learning*. Springer.

Hébrail, G., B. Hugueney, Y. Lechevallier, et F. Rossi (2010). Exploratory Analysis of Functional Data via Clustering and Optimal Segmentation. *Neurocomputing 73*(7-9), 1125–1141.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS 9*(2), 249–265.

Nguyen, X. et A. Gelfand (2011). The dirichlet labeling process for clustering functional data. *Sinica Statistica 21*(3), 1249–1289.

Ramsay, J. et B. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica 14*, 465–471.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal 27*, 379–423.

Teh, Y. W. (2010). Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer.

Vogt, J. E., S. Prabhakaran, T. J. Fuchs, et V. Roth (2010). The translation-invariant Wishart-Dirichlet process for clustering distance data. In J. Fürnkranz et T. Joachims (Eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, Israel, pp. 1111–1118. Omnipress.

Wallach, H. M., S. Jensen, L. Dicker, et K. A. Heller (2010). An alternative prior process for nonparametric bayesian clustering. *Journal of Machine Learning Research - Proceedings Track 9*, 892–899.

Summary

In this paper, we deal with the problem of curves clustering. We propose a nonparametric method which partitions the curves into clusters and discretizes the dimensions of the curve points into intervals. The cross-product of these partitions forms a data-grid which is obtained using a Bayesian model selection approach while making no assumption regarding the curves. Finally, a post-processing technique, aiming at reducing the number of clusters in order to improve the interpretability of the clustering, is proposed. It consists in optimally merging the clusters step by step, which corresponds to an agglomerative hierarchical classification whose dissimilarity measure is the variation of the criterion. Interestingly this measure is none other than the sum of the Kullback-Leibler divergences between clusters distributions before and after the merges. The practical interest of the approach for functional data exploratory analysis is presented on an artificial and a real world dataset.