Classification de données fonctionnelles par décomposition de mélange

Cuvelier, Etienne; Fraiture, Monique Noirhomme

Published in:
Revue d'Intelligence Artificielle

Publication date:
2008

Document Version
Early version, also known as pre-print

Classification de données fonctionnelles par décomposition de mélange

Apports de la visualisation dans le cas des distributions de probabilité

Etienne Cuvelier — Monique Noirhomme-Fraiture

Faculté d'Informatique
Facultés Universitaires Notre-Dame de la Paix
rue Grandgagnage, 21, B-5000, Namur, Belgique
{ecu,mno}@info.fundp.ac.be

RÉSUMÉ. *Les données fonctionnelles peuvent résulter de mesures répétées, mais aussi avoir une origine statistique. Ainsi en analyse de données symboliques, un objet complexe peut être décrit par une variable s'exprimant comme une distribution de probabilité. La classification d'un ensemble d'objets symboliques décrits par ce type de variable, peut être obtenue en appliquant une décomposition de mélange de copules archimédiennes sur les valeurs des distributions calculées en un nombre q de points distincts, appelés coupures. Jusqu'à présent ces coupures ont été choisies arbitrairement. Dans cet article nous essayons de façon empirique à l'aide de visualisations de comprendre où se situent les coupures optimales et en quel nombre. Nous proposons ensuite quelques règles pour fixer visuellement ce paramètre de la classification.*

ABSTRACT. *Functional data can come from repeated measures, but also as result of statistical analysis. In symbolic data analysis a symbolic object can be described with a probability distribution. The clustering of such objects can be performed using a mixture decomposition with archimedean copulas on values of the distributions computed in q points, named intersection points. So far this points were chosen randomly. In this paper, using visualizations, we try, empirically, to understand what is the best choice for the number and the location of these intersections points. We propose also some rules to choose this parameter of the classification.*

MOTS-CLÉS : *Analyse de Données Symbolique, Analyse Fonctionnelle, Distributions de Probabilité, Copules Archimédiennes, Visualisation*

KEYWORDS: *Symbolic Data Analysis, Functional Data Analysis, Probability Distributions, Archimedean Copulas,*

RSTI - RIA – 22/2008. Visualisation et extraction des connaissances, pages 421 à 442

1. Introduction

Les données fonctionnelles deviennent de nos jours de plus en plus courantes. Ces données peuvent résulter de mesures répétées au cours du temps (comme dans le cas des séries temporelles) ou en faisant varier un paramètre (par exemple une longueur d'onde). Dans ce cas les données se présentent comme un vecteur dans lequel chaque coordonnée est une mesure. Néanmoins, en utilisant les fonctions splines ou les ondelettes, ces données peuvent être stockées sous une forme conforme à leur nature (Ramsay *et al.*, 2005). Ces données fonctionnelles peuvent aussi avoir une origine statistique, ainsi en analyse symbolique (Bock *et al.*, 2000) une variable peut être décrite classiquement par une valeur quantitative ou qualitative, mais aussi par un intervalle, un ensemble de valeurs qualitatives, un ensemble de valeurs quantitatives pondérées ou encore par une distribution de probabilité continue. Cette dernière possibilité est celle qui est la moins destructrice d'informations lorsqu'on veut matérialiser une propriété partagée par une population d'individus et exprimée par une variable quantitative continue. Ce type particulier de données fonctionnelles peut se présenter soit sous forme de vecteurs, considérant par exemple que l'on a mesuré la distribution empirique en un certain nombre de points, soit directement sous forme analytique s'il ressort de l'analyse que la distribution est d'un type connu (normale, exponentielle, bêta,...) dont les paramètres ont été estimés.

La classification en K groupes de N individus est un problème classique d'analyse de données. Pour mener à bien cette tâche dans le cas des données fonctionnelles, plusieurs approches ont été mises au point. L'approche la plus directe est, lorsque les données résultent de mesures successives, d'utiliser directement ces valeurs. En considérant ces mesures comme des vecteurs, on peut leur appliquer les algorithmes classiques de classification (Diday *et al.*, 1982) et (Hartigan, 1975). Mais cette approche suscite plusieurs reproches : premièrement si le nombre de mesures est important, cela impliquera de travailler avec des données ayant un grand nombre de dimensions. L'autre reproche est que cette approche directe ne tient pas compte des erreurs de mesures pouvant être présentes dans les données. La suppression de ce bruit afin de retrouver les fonctions originelles, peut se faire en effectuant un lissage des données et, en stockant le résultat à l'aide de splines ou d'ondelettes (Ramsay *et al.*, 2005). En utilisant des B-splines (Abraham *et al.*, 2003) propose d'utiliser l'algorithme des k-means sur les coefficients de ces polynômes. D'autres versions adaptées des k-means ont été proposées par (Tarpey *et al.*, 2003) et (Cuesta-Albertos *et al.*, 2007). D'autres techniques de clustering ont été adaptées comme la classification hiérarchique (Dabo-Niang *et al.*, 2006) et (Dabo-Niang *et al.*, 2007), et enfin les cartes de Kohonen (Rossi *et al.*, 2004). La décomposition de mélange de densités normales utilisant l'algorithme EM a aussi été proposée pour des données fonctionnelles mesurées à intervalles irréguliers (James *et al.*, 2003). Dans le cadre de l'analyse de données symboliques, et lorsque les données sont des distributions de probabilité, (Diday, 2002) a proposé d'utiliser les copules archimédiennes pour modéliser la relation entre les différentes variables générées par l'échantillonnage des distributions, et d'effectuer la classification par une décomposition de mélange en utilisant une version

adaptée de l'algorithme des nuées dynamiques. Cette approche a déjà été utilisée avec succès par (Vrac *et al.*, 2001) sur des données atmosphériques avec deux dimensions. Nous avons étendu cette approche en permettant d'utiliser un nombre quelconque de dimensions (Cuvelier *et al.*, 2005). Dans tous les cas, le choix des coupures se révèle déterminant pour la qualité de la classification, mais actuellement aucun critère de décision automatique n'existe pour effectuer ce choix. A l'instar des travaux proposant d'impliquer significativement l'utilisateur dans le processus de classification, via des techniques de visualisations ((Ankerst, 2002), (Fayyad *et al.*, 2001), (Keim, 2002) et (Do *et al.*, 2006)), nous nous proposons d'utiliser la visualisation avec un double objectif. Premièrement, mieux comprendre l'influence du choix des coupures sur la qualité de la classification obtenue, et ensuite proposer l'utilisation de règles visuelles pour choisir ces coupures.

Cet article est structuré comme suit : dans la section 2 nous exposons la version adaptée de l'algorithme des nuées dynamiques qui est utilisée. Nous définissons ensuite dans la section 3 la notion de distribution de fonction, nécessaire à l'utilisation de l'algorithme. Ensuite dans la section 4 nous expliquons comment calculer ces distributions de fonctions à l'aide des copules archimédiennes. Dans la section 5 nous exposons les tests qui ont été réalisés, les informations que la visualisation nous a permis d'en tirer et les recommandations pour le paramétrage visuel pour la classification de ce type de données. Enfin dans la section 6 nous livrons les résultats d'une validation expérimentale des recommandations exposées dans la section précédente. Nous terminons par les conclusions et perspectives.

2. Classification par décomposition de mélange

La décomposition de mélange est un outil important en classification non supervisée. Elle consiste en l'estimation de la densité de probabilité qui est supposée avoir gouverné la génération d'un échantillon de données consitué de plusieurs groupes :

$$f(u) = \sum_{i=1}^{K} p_i \cdot f(u, \beta_i) \qquad [1]$$

où les p_i représentent les proportions de chacun des groupes (leur somme étant égale à 1), et les fonctions $f(., \beta)$ les densités de ces groupes. Chaque composante du mélange correspondant en fait à un groupe. Pour trouver la partition $P = (P_1, ..., P_K)$ la mieux adaptée aux données deux grands algorithmes ont été proposés : EM (Estimation,Maximisation) par (Dempster *et al.*, 1977) et l'algorithme des nuées dynamiques par (Diday *et al.*, 1974). Dans le cadre de cet article nous avons choisi d'utiliser ce dernier car il avait déjà été proposé (Diday, 2002), et utilisé (Vrac *et al.*, 2001), dans le cadre de l'analyse symbolique, pour la classification non supervisée de distributions de probabilités.

L'algorithme utilisé est en fait une extension de la méthode des nuées dynamiques (Diday *et al.*, 1974) dans le cas d'un mélange. L'idée principale est, alternativement,

d'estimer au mieux la distribution de chaque classe, et ensuite de vérifier que chaque objet symbolique appartient à la classe de densité maximale. L'étape d'estimation est réalisée en maximisant un critère de qualité, ici la log-vraisemblance :

$$lvc(P,\beta) = \sum_{i}^{K} \sum_{u \in P_i} log(f(u,\beta_i)). \quad [2]$$

La classification commence avec une partition initiale aléatoire, et les deux étapes suivantes sont répétées jusqu'à stabilisation de la partition :

– **étape 1 : estimation des paramètres**
déterminer le vecteur $(\beta_1, ..., \beta_K)$ qui maximise le critère de qualité ;

– **étape 2 : distribution des objets symboliques dans les classes**
les classes $(P_i)_{i=1,...,K}$, dont les paramètres ont été calculés à l'étape 1, sont construites comme suit :

$$P_i = \{u : f(u,\beta_i) \geq f(u,\beta_m) \forall m\}.$$

On constate que cet algorithme, de même que l'algorithme EM, nécessite de pouvoir calculer la distribution de fonctions, ainsi que la densité associée.

3. Distribution de fonctions

Définition 3.0.1 *Soit un espace de probabilité $(\Omega, \mathcal{B}, P)$, où Ω est la catégorie d'épreuve, $\mathcal{B}$ est une $\sigma-$algèbre sur Ω, et P est une mesure de probabilité sur $\mathcal{B}$. Soit aussi $\mathcal{D} = [a,b] \subseteq \mathbb{R}$ un intervalle fermé de $\mathbb{R}$. Une* variable aléatoire fonctionnelle (vaf) $\underline{X}$ *sur $\mathcal{D}$ est une application de $\mathcal{D} \times \Omega$ dans la droite réelle achevée $\overline{\mathbb{R}} = [-\infty, +\infty]$ telle que, pour tout $t \in \mathcal{D}$, $X(t,.)$ soit une variable aléatoire réelle sur $(\Omega, \mathcal{B}, P)$:*

$$\underline{X} : \mathcal{D} \times \Omega \to \overline{\mathbb{R}} : (t,\omega) \mapsto X(t,\omega). \quad [3]$$

Les variables aléatoires réelle $\underline{X}(t,.)$ sont aussi notées $\underline{X}_t$. Chacune des fonctions $X(.,\omega)$ s'appelle une réalisation ou trajectoire de la vaf $\underline{X}$.

Les *vaf* sont aussi des processus stochastiques. Le cadre dans lequel nous travaillons est l'espace de Hilbert $\mathcal{L}_2(\mathcal{D})$.

Définition 3.0.2 *Soit $\mathcal{D} = [a,b] \subseteq \mathbb{R}$ un intervalle fermé de $\mathbb{R}$, et $C^0(\mathcal{D})$ l'ensemble des fonctions continues bornées de domaine $\mathcal{D}$. Soient $u, v \in C^0(\mathcal{D})$, on définit :*

– $\|u\|_2 = \left\{ \int_{\mathcal{D}} |u(x)|^2 \, dx \right\}^{1/2}$

– $\mathcal{L}_2(\mathcal{D}) = \left\{ u \in C^0(\mathcal{D}) \ : \ \|u\|_2 < \infty \right\}$.

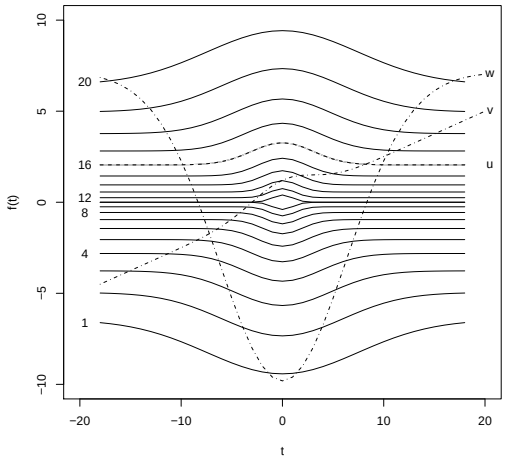

Figure 1. *Un exemple d'échantillon de 20 fonctions*

Dans ce cadre, chaque réalisation $X(.,\omega)$ est donc un élément de $\mathcal{L}_2(\mathcal{D})$:

$$\underline{X} : \Omega \to \mathcal{L}_2(\mathcal{D}) : \omega \mapsto X(\omega). \quad [4]$$

Définition 3.0.3 *Soient $f, g \in \mathcal{L}_2(\mathcal{D})$. L'ordre ponctuel entre f et g sur l'intervalle $\mathcal{D}$ est défini par :*

$$\forall t \in \mathcal{D}, f(t) \leq g(t) \iff f \leq_{\mathcal{D}} g. \quad [5]$$

Cet ordre ponctuel, n'est évidemment pas un ordre total, mais permet néanmoins, de définir une fonction de répartition fonctionnelle.

Définition 3.0.4 *La* fonction de répartition fonctionnelle *ou* distribution fonctionnelle *d'une* vaf $\underline{X}$ *pour l'intervalle $\mathcal{D}$ est la fonction définie sur $\mathcal{L}_2(\mathcal{D})$ par :*

$$\begin{aligned} F_{\underline{X},\mathcal{D}}(u) &= P\{\omega \in \Omega : X(\omega) \leq_{\mathcal{D}} u\} \\ &= P[\underline{X} \leq_{\mathcal{D}} u] \end{aligned} \quad [6]$$

où $u \in \mathcal{L}_2(\mathcal{D})$.

La nature non nécessairement comparable des données fonctionnelles rend le calcul de cette probabilité difficile, y compris de façon empirique. Essayons intuitivement avec un exemple simple de fonctions définies sur $\mathcal{D} = [-20, 20]$ (figure 1). Supposons que,

hormis v et w, les fonctions représentées constituent un échantillon A, de taille 20, représentatif d'une variable aléatoire fonctionnelle $\underline{X}$. On peut alors essayer d'estimer empiriquement la distribution de $\underline{X}$ en u, fonction issue de l'échantillon :

$$\widehat{F}_{\underline{X},\mathcal{D}}(u) = \frac{\#\{f \in A : f \leq_{\mathcal{D}} u\}}{\#A} = \frac{16}{20} = \frac{4}{5}.$$

De même nous pouvons estimer la distribution en v, fonction qui, visiblement n'est pas issue de l'échantillon :

$$\widehat{F}_{\underline{X},\mathcal{D}}(v) = \frac{\#\{f \in A : f \leq_{\mathcal{D}} v\}}{\#A} = \frac{2}{20} = \frac{1}{10}.$$

Mais pour w, nous pouvons difficilement accepter le résultat de cette estimation empirique, car si w n'est comparable à aucune fonction de l'échantillon :

$$\nexists f \in A : f \leq_{\mathcal{D}} w$$

on constate, par contre, sur la figure 1 que w est plus grande que $1/20^e$ de l'échantillon, et ce sur une grande majorité de $\mathcal{D}$. L'alternative à cette approche restrictive est d'approximer cette probabilité à l'aide d'une distribution multivariée.

4. Approximations multivariées

4.1. *Introduction*

Soient $n \in \mathbb{N}$ et $\{t_1, \ldots, t_q\}$, q points de $\mathcal{D}$. Si nous définissons les deux ensembles suivants :

$$\mathcal{A}_{\underline{X}}(u) = \{\omega \in \Omega : X(\omega) \leq_{\mathcal{D}} u\}$$

$$\mathcal{A}_{\underline{X},q}(u) = \bigcap_{i=1}^{q} \{\omega \in \Omega : X(\omega)(t_i) \leq u(t_i)\}$$

alors nous pouvons utiliser l'approximation suivante :

$$F_{\underline{X},\mathcal{D}}(u) = P[\mathcal{A}_{\underline{X}}(u)] \approx P[\mathcal{A}_{\underline{X},q}(u)] = H\left(u(t_1), \ldots, u(t_q)\right) \qquad [7]$$

où H est une distribution multivariée de dimension q. Le choix de la distribution, ou de la famille de distributions, a utiliser est évidemment important. Avant de préciser ce choix, remarquons que pour une valeur choisie $t \in \mathcal{D}$, il est très facile d'estimer la distribution des valeurs de $\underline{X}(t)$.

Définition 4.1.1 *Soit $\underline{X}$ une* vaf. *Les fonctions G et g, respectivement appelées* surface de distributions *et* surface de densités, *de domaines $\mathcal{D}$ sont définies par*

$$G(t, y) = P[\underline{X}(t) \leq y] \qquad [8]$$

$$g(t, y) = \frac{\partial}{\partial t} G(t, y). \qquad [9]$$

Si $\underline{X}$ est un processus Gaussien, alors ces deux fonctions peuvent être calculées pour une valeur donnée de t par la fonction de répartition et la densité de la loi $\mathcal{N}(\mu(t), \sigma(t))$.

$$G\left(t, y\right) = F_{\mathcal{N}(\mu(t),\sigma(t))}(y) \quad [10]$$

$$g\left(t, y\right) = f_{\mathcal{N}(\mu(t),\sigma(t))}(y). \quad [11]$$

Dans les cas où l'on ignore la loi suivie par $\underline{X}(t)$ on utilisera l'estimation empirique pour G :

$$\hat{G}\left(t, y\right) = \frac{\#\left\{X_i(t) \leq y\right\}}{N} \quad [12]$$

et l'estimation à noyaux pour g :

$$\hat{g}\left(t, y\right) = \frac{1}{N \cdot h(t)} \sum_{i=1}^{N} K\left(\frac{y - X_i(t)}{h(t)}\right) \quad [13]$$

où (Silverman, 1986) :

– K est une fonction noyau (Gaussienne, Epanechnikov, Triangulaire,...),

– h est le paramètre de lissage, qui peut être déterminé, notamment, en minimisant la MISE (Mean Integrated Square Error), par exemple en utilisant la "rule of thumb".

Les figures 2 et 3 montrent ces deux surfaces avec l'exemple de la figure 1, dans le cas Gaussien.

Etant donné qu'il est très facile de calculer les marges de la distribution H : $G\left(t_1, u\left(t_1\right)\right), \ldots, G\left(t_q, u\left(t_q\right)\right)$, l'idée de reconstruire cette distribution H à partir de ses marges en utilisant les copules, a été proposée par (Diday, 2002).

4.2. *Copules et copules archimédiennes*

Définition 4.2.1 *Une copule C est une distribution multivariée définie sur le cube $[0,1]^n$, dont toutes les marginales sont uniformes sur $[0,1]$.*

$$C : [0,1]^n \rightarrow [0,1] : (u_1, \ldots, u_n) \mapsto C(u_1, \ldots, u_n).$$

Les copules sont des outils précieux dans la modélisation des structures de dépendance grâce au théorème de Sklar (voir (Nelsen, 1999)).

Théoreme 4.1 (Sklar) *Si $H(x_1, \ldots, x_n)$ est une distribution multivariée de marges $F_1(x_1)$,...,$F_n(x_n)$, alors il existe une copule C telle que*

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \quad [14]$$

De plus, si $F_1, ..., F_n$ sont toutes continues, alors C est unique ; sinon C est unique seulement sur $domF_1 \times ... \times domF_n$.

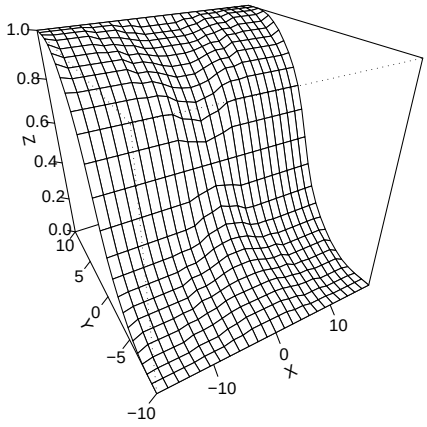

Figure 2. *La surface $G(x, y)$ de l'exemple de la figure 1*

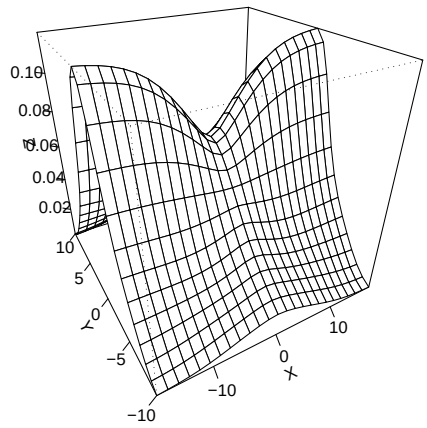

Figure 3. *La surface $g(x, y)$ de l'exemple de la figure 1*

En fait la copule associée à une distribution de probabilité capture la relation de dépendance qui relie les marginales. Les trois fonctions suivantes sont trois copules importantes :

$$M^n(u_1,\ldots,u_n) = min(u_1,\ldots,u_n), \quad [15]$$

$$\Pi^n(u_1,\ldots,u_n) = \prod_{i=1}^{n} u_i, \quad [16]$$

$$W^n(u_1,\ldots,u_n) = max\left(\sum_{i=1}^{n} u_i - n + 1, 0\right). \quad [17]$$

Les fonctions Π^n et M^n sont des copules $\forall n \geq 2$, alors que W^n n'est une copule que pour $n = 2$.

Les fonctions M^n et W^n sont appelées *bornes de Fréchet-Hoeffding*, et ce car quelle que soit la copule C on a toujours la relation suivante :

$$W^n(u_1,\ldots,u_n) \leq C(u_1,\ldots,u_n) \leq M^n(u_1,\ldots,u_n). \quad [18]$$

Même en n'étant pas une copule pour $n > 2$, W^n est la meilleure borne inférieure pour l'ensemble des copules de dimensions supérieures à deux. Ces copules W^n, Π^n et M^n correspondent aussi à trois cas particuliers de dépendance. Si $X_1, X_2, \ldots, X_n$ sont des variables aléatoires continues alors (Nelsen, 1999) :

– $X_1, X_2, \ldots, X_n$ indépendantes si et seulement si leur copule associée est Π^n ;

– chaque variable aléatoire $X_1, X_2, \ldots, X_n$ est presque sûrement l'image par une fonction strictement croissante de toutes les autres variables si et seulement si leur copule associée est M^n ; ces variables sont dites comonotoniques quand $n = 2$;

– si $n = 2$: chaque variable aléatoire est presque sûrement l'image par une fonction strictement décroissante de l'autre variable si et seulement si leur copule associée est W^n ; ces variables sont dites anticomonotoniques.

Nous avons indiqué que les *vaf* sont aussi des processus stochastiques, or il existe une classe importante de ce type de processus : les processus strictement stationnaires. Un processus stochastique $\underline{X}_t$ est dit strictement stationnaire ((Burril, 1972) et (Cox *et al.*, 1965)) si $\forall t_1, \ldots, t_n$ et pour tout h, la distribution conjointe de $\left(\underline{X}_{t_1+h}, \ldots, \underline{X}_{t_n+h}\right)$ ne dépend pas de h. Dans le cadre de l'analyse de données fonctionnelles nous proposons d'utiliser comme modèle une classe plus large de processus stochastiques : les processus stationnaires par copule. Un processus stochastique $\underline{X}_t$ sera dit stationnaire par copule si $\forall t_1, \ldots, t_n$ et pour tout h, la copule associée à $\left(\underline{X}_{t_1+h}, \ldots, \underline{X}_{t_n+h}\right)$ ne dépend pas de h. Une importante famille de copules est bien adaptée à ce cas de figure : les copules archimédiennes.

Définition 4.2.2 *Pour $n \geq 2$ les copules archimédiennes sont définies par*

$$C(\underline{u}) = C(u_1, ..., u_n) = \psi\left(\sum_{i=1}^{n} \phi(u_i)\right) \quad [19]$$

Nom	$\phi_\theta(\mathbf{t})$	**Dom.** θ	**Cas limites**
Clayton	$t^{-\theta}-1$	$[-1,\infty[\setminus\{0\}$	$C_{-1}=W, C_0=\Pi, C_\infty=M$
Frank	$-\ln\frac{e^{-\theta\cdot t}-1}{e^{-\theta}-1}$	$]-\infty,\infty[\setminus\{0\}$	$C_{-\infty}=W, C_0=\Pi, C_\infty=M$
Gumbel	$(-\ln t)^\theta$	$[1,\infty[$	$C_1=\Pi, C_\infty=M$

Tableau 1. *Trois générateurs de copules archimédiens bivariées*

où ϕ, le générateur, est une fonction continue strictement décroissante de $[0,1]$ vers $[0,\infty]$ telle que :

(*) $\phi(0)=\infty$ *et* $\phi(1)=0$

(*) $\psi=\phi^{-1}$ *soit complètement monotonique sur* $[0,\infty[$ *c-à-d que* $\forall t\in[0,\infty[$ *et* $\forall k\geq 0$

$$(-1)^k\psi^{[k]}(t)\geq 0$$

où $\psi^{[k]}$ représente la dérivée d'odre k de ψ.

Notez bien, que si $n=2$, il n'est pas nécessaire que le générateur ϕ soit complètement monotonique, il suffit alors qu'il soit convexe, les autres conditions restant inchangées. Il existe un nombre assez important de générateurs de copules archimédiennes lorsque $n=2$, mais, malheureusement, peu de ces générateurs sont encore valables lorsque $n>2$. Dans le tableau 4.2 nous montrons trois familles de générateurs Archimédiens qui définissent des copules pour $n\geq 2$. Nous donnons le domaine et les cas limites dans le cas bivarié.

REMARQUE. — Pour $n>2$, si C est une copule archimédienne, alors on a que $\Pi^n\leq C\leq M^n$, c'est-à-dire que les générateurs de copules archimédiennes ne peuvent modéliser que des dépendances positives. Ceci est une conséquence de l'exigence de complète monotonicité, qui se traduit aussi par une réduction du domaine de définition du générateur. Ainsi, par exemple, pour $n>2$, le domaine de définition du générateur de la copule de Clayton sera réduit à $]0,\infty[$ au lieu de celui mentionné dans le tableau 4.2.

Une des propriétés des copules archimédiennes est que pour k fixé (avec $2\leq k\leq n$) toutes les marges de dimension k d'une copule sont identiques. Ainsi les marges bidimensionnelles, obtenues à partir de l'expression (19) de la façon suivante,

$$\phi^{-1}\left(\phi(u_i)+\phi(u_j)+\sum_{k\neq i,j}\phi(1)\right)=\phi^{-1}\left(\phi(u_i)+\phi(u_j)\right)=C(u_i,u_j) \qquad [20]$$

sont toutes modélisées de la même façon, c'est-à-dire avec le même générateur ϕ. Cette propriété nous permet d'utiliser les copules archimédiennes pour notre modéli-

sation à l'aide de processus stationnaires par copule. Dans le cadre de cet article nous n'avons utilisé que la copule de Clayton :

$$C_\theta(\underline{u}) = \left(\sum_{i=1}^{n} (u_i^{-\theta}) - n + 1 \right)^{-\frac{1}{\theta}}. \qquad [21]$$

Si nous utilisons conjointement les *surfaces de distributions* et les copules archimédiennes, alors notre approximation [7] peut directement se récrire :

$$F_{\underline{X},\mathcal{D}}(u) \approx P[\mathcal{A}_q(u)] = \psi \left(\sum_{i=1}^{q} \phi \left(G \left[t_i, u \left(t_i \right) \right] \right) \right). \qquad [22]$$

La densité conjointe étant donnée par l'expression suivante :

$$\frac{\partial^q}{\partial u_1 ... \partial u_q} C(G \left[t_1, u(t_1) \right], \ldots, G \left[t_q, u(t_q) \right]) \cdot \prod_{i=1}^{q} g \left[t_i, u(t_i) \right]. \qquad [23]$$

L'expression [22] a déjà été utilisée avec l'algorithme proposé et la copule de Frank par (Vrac *et al.*, 2001) dans le cadre de l'analyse symbolique, mais seulement en deux dimensions ($q = 2$) . Dans (Cuvelier *et al.*, 2005) nous avons proposé d'utiliser la copule de Clayton, car la densité de cette dernière est facile calculer pour un nombre quelconque de dimensions, ce qui n'est pas le cas de le copule de Frank. A la suite de ces travaux deux questions restaient sans réponse :

1) Quelle valeur de q choisir ?

2) Où choisir les valeurs $t_1, \ldots, t_q$?

Il est difficile de répondre à ces questions dans l'absolu, mais dans la cadre de la classification, on peut dire, en ce qui concerne les coupures, que les valeurs calculées des fonctions en ces points doivent permettre de distinguer les K groupes recherchés. En ce qui concerne le nombre de coupures, la valeur de q doit à la fois être suffisamment grande pour permettre la distinction évoquée ci-dessus, mais suffisamment petite pour ne pas engendrer un temps calcul déraisonnable. La qualité de la classification va clairement résulter du choix de ces paramètres. En l'absence de critères objectifs pour cette paramétrisation nous nous sommes tournés vers une utilisation des informations fournies par la visualisation des données.

5. Choix des coupures et domaines

Pour essayer de répondre aux deux questions précédentes nous avons procédé à plusieurs expérimentations sur des ensembles artificiels de données. Générant nos ensembles de données, nous en connaissons donc la répartition des individus entre les différents clusters. Ils nous est possible, démarche non conventionnelle en classification non supervisée, de comparer les résultats de la classification avec les groupes a

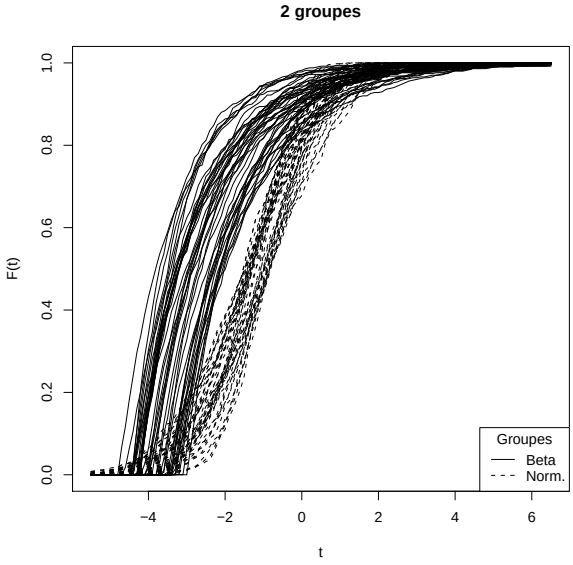

Figure 4. *Ensemble à 2 groupes*

priori. L'idée n'étant pas ici de se concentrer sur les résultats de la classification, mais bien sur l'impact du choix des coupures sur ceux-ci.

Pour illustrer notre propos nous avons utilisé 2 ensembles de données constitués de 2 à 3 groupes de distributions. Ces dernières étant choisies parmi les trois types suivants : lois normales, lois bêta et lois exponentielles.

Pour que ces ensembles ne soient pas trop éloignés de ceux résultant d'une situation réelle, nous avons, pour chaque distribution, généré 500 nombres aléatoires suivant la loi et les paramètres choisis, ensuite nous avons estimé la distribution empirique. L'environnement statistique de (R Development Core Team, 2007) permettant de stocker ces distributions empiriques sous forme fonctionnelle, il nous est alors loisible de calculer les valeurs prises par ces fonctions de répartitions aux coupures choisies.

Les 2 ensembles de données sont illustrés dans les figures 4 et 5. Nous avons ensuite généré tous les ensembles de 2 à 7 coupures équidistantes (d'un multiple de 0.25) et ayant comme première coupure au moins -5.5, et comme dernière coupure au plus 6 :

$$\{T_k = d + 0.25k : q \in \{2, \ldots, 7\}, k \in \{0, \ldots, q-1\}, -5.5 \leq T_0, d, T_{q-1} \leq 4\}. \quad [24]$$

Ensuite nous avons effectué les classifications sur base de chacun de ces ensembles de 2 à 7 coupures.

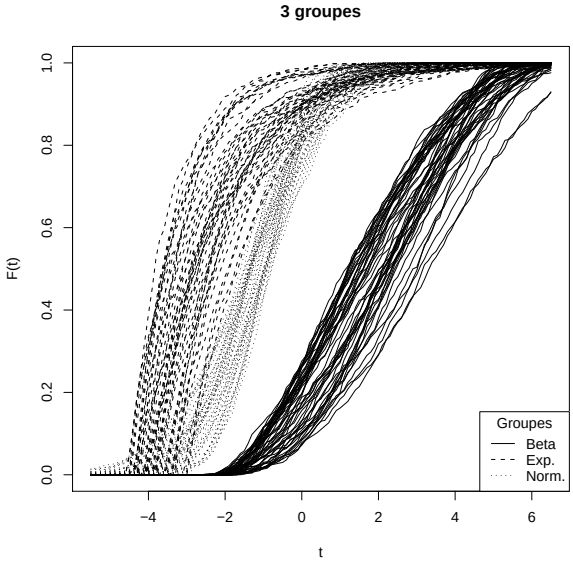

Figure 5. *Ensemble à 3 groupes*

La qualité de la classification de type nuées dynamiques dépend directement des optima de la fonction [2] permettant de calculer le critère de qualité. La paramétrisation, c'est-à-dire la partition initiale aléatoire des données, à une influence directe sur l'optima qui sera atteint. Comme il n'existe aucun critère permettant de distinguer un optimum local d'un optimum global, la technique la plus communément utilisée pour tenter d'atteindre l'optimum global, sans garantie sur son existence, est d'utiliser successivement des partitions initiales aléatoires différentes, et de retenir la classification procurant la plus grande valeur pour le critère choisi.

q	N	Moy.	Min.	Max.	N<10%	Moy.10%	$[T_{min}, T_{max}]$
2	741	22,2%	0%	50,0%	230	5,1%	[-4,50, 4,00]
3	361	18,1%	0%	50,0%	125	4,1%	[-4,50, 4,00]
4	234	17,9%	0%	50,0%	82	3,2%	[-4,50, 4,00]
5	171	17,2%	0%	48,8%	64	3,4%	[-4,50, 4,00]
6	133	18,2%	0%	50,0%	43	5,0%	[-4,50, 4,00]
7	108	18,5%	0%	47,7%	36	6,2%	[-5,00, 4,00]

Tableau 2. *Résultats pour 2 groupes*

Dans notre expérimentation, comme nous utilisons des ensembles synthétiques de données, nous en connaissons donc la répartition a priori. Pour tenter de mesurer l'im-

q	N	Moy.	Min.	Max.	N<10%	Moy.10%	$[T_{min}, T_{max}]$
2	741	23,2%	0,0%	42,1%	214	4,0%	[-3,00, 1,50]
3	361	19,3	2,1%	37,9%	115	5,1%	[-3,25, 2,00]
4	234	20,0	2,1%	38,6%	47	5,7%	[-4,00, 3,00]
5	171	18,8	2,1%	37,9%	38	5,8%	[-4,00, 2,75]
6	133	20,4	3,6%	39,3%	28	6,5%	[-4,00, 2,50]
7	108	22,9	0,7%	42,1%	23	7,1%	[-4,00, 2,75]

Tableau 3. *Résultats pour 3 groupes*

pact du choix des coupures nous avons décidé de figer la répartition initiale aléatoire des données, et de ne faire varier que l'emplacement des coupures. Nous ne nous sommes pas directement préoccupés de la valeur finale du critère de qualité, car nous avons ensuite comparé le résultat de toutes ces classifications avec notre connaissance a priori de l'appartenance aux différents groupes. Nous nous sommes ensuite concentrés sur les coupures fournissant un taux d'erreur considéré comme acceptable. Nous avons fixé ce taux à 10 %.

Les tableaux 2 et 3 donnent les résultats de ces 3 496 tests. Ces tableaux contiennent les informations suivantes :

q : indique le nombre de coupures utilisées (càd la dimension de la copule utilisée),

N : donne le nombre d'ensembles de coupures testés,

Moy. : moyenne de mauvaise classification pour tous les ensembles générés,

Min. : minimum de mauvaise classification pour tous les ensembles générés,

Max. : maximum de mauvaise classification pour tous les ensembles générés,

$\mathbf{N \leq 10\%}$ **:** nombre d'ensembles de coupures permettant une mauvaise classification inférieure à 10 %,

Moy.10% : moyenne de mauvaise classification pour les ensembles de coupures permettant une mauvaise classification inférieure à 10 %,

$[\mathbf{T_{min}}, \mathbf{T_{max}}]$ **:** intervalles où se situent les coupures des ensembles permettant mauvaise classification inférieure à 10 %.

Comme on peut le voir dans ces tableaux, si on choisit les coupures de manière arbitraire on peut obtenir des cas favorables et avoir un taux d'erreur proche de 0, mais aussi obtenir des cas très défavorables et avoir jusqu'à 50 % d'erreur. Nous avons dès lors besoin d'heuristiques pour bien choisir les coupures.

Concentrons-nous maintenant sur les ensembles de coupures qui permettent d'effectuer la classification avec un taux d'erreur acceptable (10 % maximum). La première constatation, concerne le nombre de coupures, et on voit que l'augmentation du nombre de coupures n'améliore pas nécessairement le taux d'erreur. En effet quel que soit le nombre de coupures, la copule archimédienne modélisera la dépendance

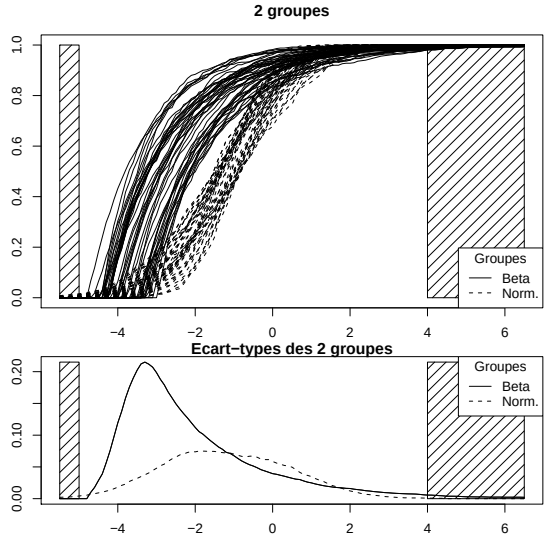

Figure 6. *Comparatif pour 2 groupes*

entre marges de la même façon, toutes les marges de de dimension $2 \leq k \leq q$ étant identiques.

En ce qui concerne la localisation des coupures nous avons généré deux graphiques (figures 6 et 7) dans lesquelles nous avons mis en évidence les intervalles où se situent les coupures des ensembles permettant une mauvaise classification inférieure à 10 %, et ce en hachurant les parties complémentaires (au sens ensembliste). Nous avons aussi superposé les données (fonctions de répartitions) et les écarts types des différents groupes. On constate que les coupures permettant une classification intéressante ne sont jamais choisies là où au moins un des groupes est d'écart type nul ou quasi nul. Cela est dû au fait que les différents groupes peuvent être considérés comme faisant partie de sous-espaces différents. En effet si, pour un groupe, on note $N = \{t_i \in \mathcal{D} : \sigma(t_i) = 0\}$, alors si $t_i \in N$, $G[t_i, u(t_i)]$ ne peut être égale qu'à 0 ou 1. Si G vaut 0 pour au moins un t_i alors l'expression (22) est égale à 0. Par contre si G est non nulle pour tous les t_i et égale à 1 pour les coupures appartenant à N, alors la même expression (22) devient :

$$\psi\left(\sum_{i \notin N} \phi\left(G\left[t_i, u\left(t_i\right)\right]\right)\right)$$

c'est-à-dire qu'elle est seulement évaluée dans le sous-espace où l'écart type du groupe est non nul. Il n'est de toute façon pas possible d'évaluer l'expression (23) hors de ce sous-espace, cette densité conjointe étant alors nulle ou infinie.

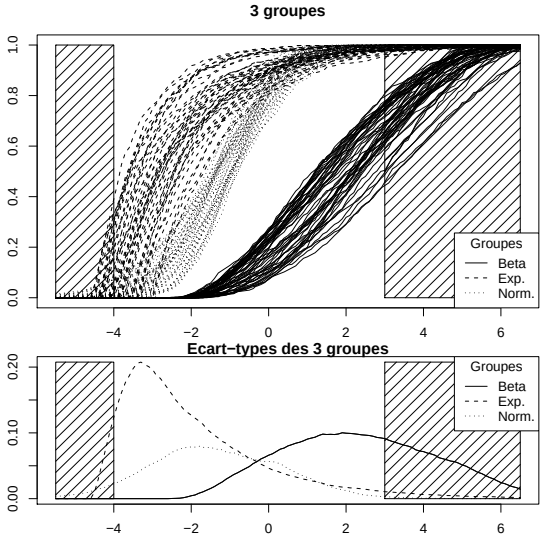

Figure 7. *Comparatif pour 3 groupes*

Enfin on constate que dans l'espace non hachuré on peut choisir les coupures de façon à ce que, premièrement un maximum de groupes soit distinguable le long de ces coupures, et deuxièmement, que les groupes non distincts sur un groupe de coupures, le soient sur son complément, au sens ensembliste.

Pour l'ensemble à 2 groupes nous avons, dans la figure 8, utilisé la densité pour illustrer la répartition des coupures pour les dimensions de 2 à 6 (nous nous sommes limités à ces dimensions pour une question de lisibilité du graphique). Pour rechercher les coupures qui semblent optimales nous avons recherché dans le tableau 2 la dimension pour laquelle nous obtenons le taux d'erreur moyen le plus bas, ensuite pour ce nombre de dimensions nous avons recherché les maxima des densités illustrées. Ces maxima se situent en $t_1 = -3.42$, $t_2 = -0.24$, $t_3 = 1.15$ et $t_4 = 3.09$. La figure 9 montre les valeurs des fonctions en ces coupures ainsi que la densités des valeurs calculées pour chacune des coupures. Nous constatons que dans chacune des cases de la matrice graphique, 2 groupes plus ou moins distincts se détachent, alors que dimension par dimension il n'y a que pour la 1re et 3e coupures que deux groupes semblent distinguables à l'aide des densités. Nous avons répété la même recherche pour l'ensemble à 3 groupes (figures 10 et 11), et dans ce cas les maxima des densités se situent en $t_1 = -1.65$ et $t_2 = -1.1$. On distingue bien les 3 groupes dans la partie centrale de la figure 11, de même au niveau de la densité de la première coupure.

Les coupures ayant une influence directe sur la qualité des résultats, nous ne pouvons nous satisfaire d'un choix aléatoire. En nous inspirant des constatations ci-dessus

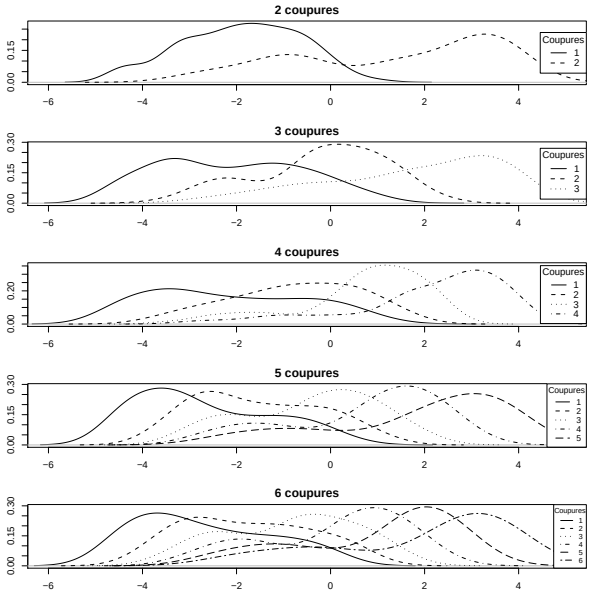

Figure 8. *Distributions des coupures pour 2 groupes*

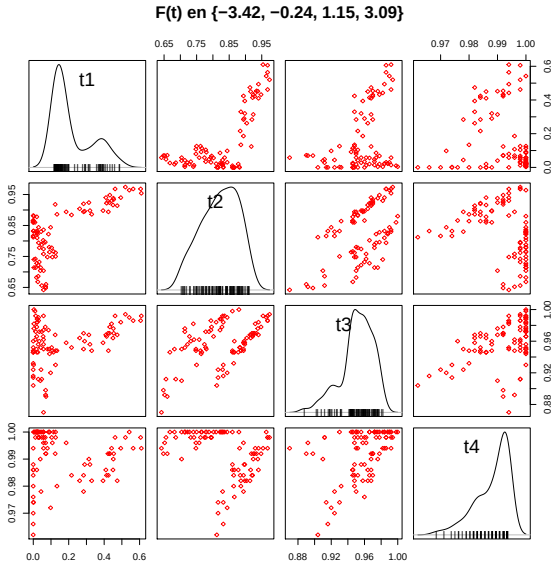

Figure 9. *Valeurs des fonctions aux 4 coupures optimales pour 2 groupes*

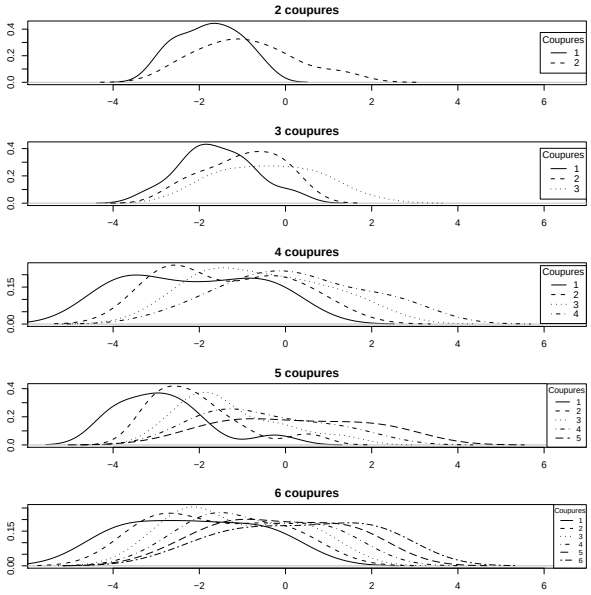

Figure 10. *Distributions des coupures pour 3 groupes*

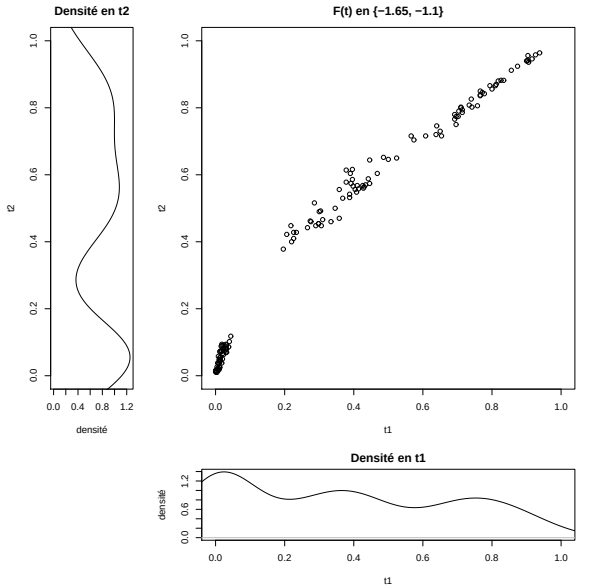

Figure 11. *Valeurs des fonctions aux 2 coupures optimales pour 3 groupes*

nous pouvons émettre les heuristiques suivantes à mettre en œuvre à partir de la visualisation des données :

1) restreindre les coupures aux domaines où un maximum de groupes sont d'écart type non nul,

2) choisir des coupures qui maximisent le nombre de groupes discernables de valeurs le long de ces coupures, en veillant à ce que les groupes qui ne sont pas discernés sur un sous-ensemble de coupures le soient sur son complément (au sens ensembliste),

3) minimiser le nombre de coupures choisies.

6. Validation expérimentale des heuristiques

Pour illustrer, et valider expérimentalement, nos heuristiques nous avons utilisé un ensemble à 4 groupes, non utilisés dans les tests précédents. Cet ensemble est illustré dans la partie supérieure de la figure 12. Nous avons choisi nos coupures à l'aide des règles énoncées :

1) on distingue, sur la figure 12, 4 groupes, et un maximum de groupes semblent d'écart type non nul entre 1,5 et 5,2 ;

2) nous choisissons comme 1,5 comme première coupure où deux groupes semblent être distinguables, les deux autres étant confondus aux valeurs nulles ; nous choisissons 3,5 comme deuxième coupure où les deux premiers groupes semblent toujours distinguables et où un troisième groupe se détache ; enfin nous choisissons 5,2 comme dernière coupure car le quatrième groupe, non encore distingué des autres sur les deux premières coupures, est enfin distinguable, alors que les deux premiers groupes commencent à se confondre ;

3) nous fixons notre choix à ces trois coupures car tous les groupes qui semblent distinguables peuvent être distingués sur l'ensemble des coupures.

Les valeurs des fonctions en ces coupures sont visibles dans la figure 13, où l'on distingue trois groupes dans chacune des cases de la matrice, mais jamais quatre. Seule la troisième coupure semble, permettre de distinguer trois groupes rien qu'à l'aide des la densité. Nous avons ensuite effectué 20 classifications avec des partitions initiales différentes. Nous avons ensuite comparé ces classifications à notre connaissance a priori des groupes. Les résultats sont les suivants : la moyenne des taux d'erreur est de 12,13 %, avec un écart type de 10,55 %, ce qui est plus que le taux de 10 % maximum recherché, mais si nous ne retenons que la classification donnant la meilleure log-vraisemblance, alors les taux d'erreur est 0 %.

Nous avons ensuite appliqué la méthode exposée dans la section précédente pour rechercher les 3 coupures optimales dans l'ensemble suivant :

$$\{T_k = d + 0.25k : k \in \{0, 1, 2\}, -5 \leq T_0, d, T_2 \leq 10\}. \qquad [25]$$

Les densités correspondantes sont illustrées dans le bas de la figure 12. Nous constatons que nos deux dernières coupures sont proches des maxima des deux premières

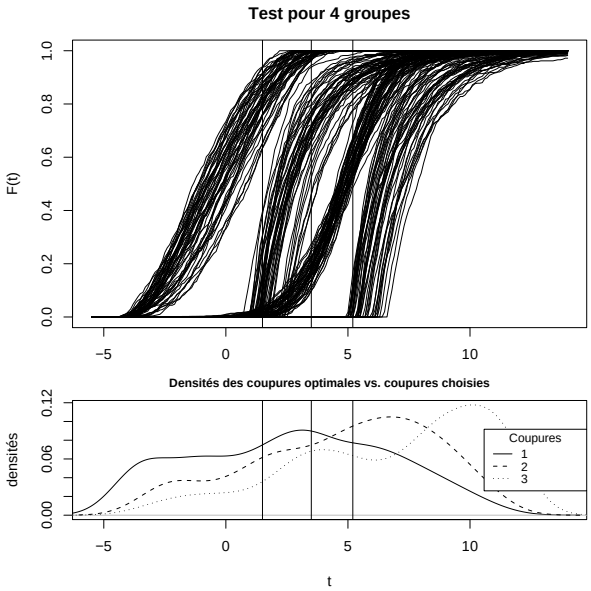

Figure 12. *Ensemble de validation à 4 groupes, distribution des coupures optimales et coupures choisies*

coupures optimales. La dernière coupure “optimale” n’apportant manifestement aucune information supplémentaire par rapport aux deux premières, nous aurions vraisemblablement pu épargner une coupure. Néanmoins, la démarche nous a permis de trouver une classification de bonne qualité, et ce en appliquant la démarche classique qui consiste à ne retenir que la classification donnant le critère de qualité le plus grand.

7. Conclusions et perspectives

Dans cet article nous avons montré le double apport de la visualisation dans le choix des coupures nécessaires à la classification. Le premier de ces apports est de nous avoir permis de comprendre et synthétiser les résultats de nos expérimentations. Le second apport est d’utiliser ces conclusions pour émettre des heuristiques se basant sur la visualisation des données pour paramétrer la classification. Une première validation expérimentale a donné des résultats encourageants dans le cadre d’un processus classique de classification. Les résultats des tests exposés en section 5, pourraient encore être affinés en répétant les mêmes procédures à partir de plusieurs partitions initiales. De même, les observations de cet article ne concernant que des coupures équidistantes, nous expérimentons actuellement la même démarche pour des coupures distribuées aléatoirement. Cette démarche est particulièrement gourmande en temps car

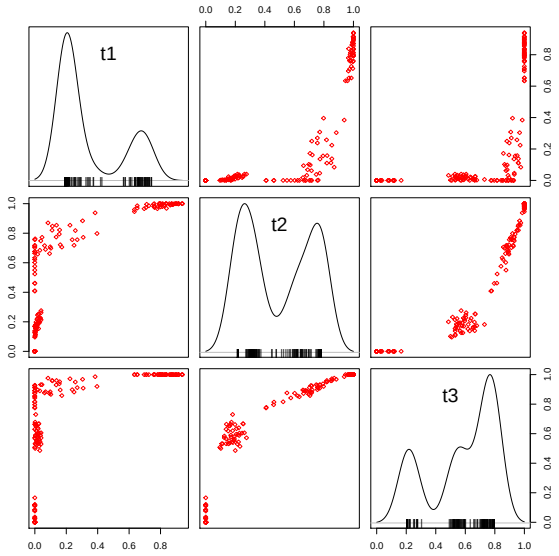

Figure 13. *Valeurs des fonctions aux 3 coupures choisies pour l'ensemble de validation*

dans certains cas la proportion de coupures intéressantes par rapport à l'ensemble des coupures générées est inférieure à 1 pour 10 000. Néanmoins nous espérons pouvoir, grâce à la visualisation, en tirer des conclusions similaires et affiner nos heuristiques pour la détermination du nombre optimal de coupures et leurs emplacements. Enfin une utilisation sur des données issues du monde réel est envisagée.

8. Bibliographie

Abraham C., Cornillon P. A., Matzner-Løber E., Molinari N., « Unsupervised Curve Clustering using B-Splines », *Scandinavian Journal of Statistics*, vol. 30, n° 3, p. 581-595, 2003. available at http ://ideas.repec.org/a/bla/scjsta/v30y2003i3p581-595.html.

Ankerst M., « Report on the SIGKDD-2002 Panel The Perfect Data Mining Tool : Interactive or Automated », *SIGKDD Explorations*, vol. 4, n° 2, p. 110-111, 2002.

Bock H., Diday E., *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*, Springer Verlag, 2000.

Burril C. W., *Measure, integration and probability*, McGraw-Hill, New-York, 1972.

Cox D. R., Miller H., *The theory of stochastic processes*, Methuen, London, 1965.

Cuesta-Albertos J. A., Fraiman R., « Impartial trimmed k-means for functional data », *Comput. Stat. Data Anal.*, vol. 51, n° 10, p. 4864-4877, 2007.

Cuvelier E., Noirhomme-Fraiture M., « Clayton copula and mixture decomposition », *in* J. Janssen, P. Lenca (eds), *Applied Stochastic Models and Data Analysis ASMDA 2005*, Brest, p. 699-708, 2005.

Dabo-Niang S., Ferraty F., Vieu P., « Mode estimation for functional random variable and its application for curves classification », *Far East J. Theoret. Statist.*, vol. 18, n° 1, p. 93-119, 2006.

Dabo-Niang S., Ferraty F., Vieu P., « On the using of modal curves for radar waveforms classification », *Comput. Stat. Data Anal.*, vol. 51, n° 10, p. 4878-4890, 2007.

Dempster A. P., Laird N. M., Rubin D. B., « Maximum likelihood from incomplete data via the EM algorithm (with discussion) », *Journal of the Royal Statistical Society (Series B)*, vol. 39, n° 1, p. 1-38, 1977.

Diday E., « Mixture decomposition of distributions by copulas », *Classification, Clustering and Data Analysis*, p. 297-310, 2002.

Diday E., Lemaire J., Pouget J., Testu F., *Éléments d'analyse de données*, Dunod, 1982.

Diday E., Schroeder A., Ok Y., « The Dynamic Clusters Method in Pattern Recognition. », *IFIP Congress*, p. 691-697, 1974.

Do T.-N., Poulet F., « SVM incrémental, parallèle et distribué pour le traitement de grandes quantités de données », *EGC*, p. 47-52, 2006.

Fayyad U., Grinstein G. G., Wierse A., *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

Hartigan J. A., *Clustering Algorithms*, John Wiley & Sons, 1975.

James G., Sugar C., « Clustering for Sparsely Sampled Functional Data », *Journal of the American Statistical Association*, vol. 98, p. 397-408, January, 2003. available at http ://ideas.repec.org/a/bes/jnlasa/v98y2003p397-408.html.

Keim D. A., « Information Visualization and Visual Data Mining », *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, n° 1, p. 1-8, 2002.

Nelsen R., *An introduction to copulas*, Springer, London, 1999.

R Development Core Team, *R : A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. 2007.

Ramsay J. O., Siverman B. W., *Functionnal Data Analysis*, Springer Series in Statistics, Springer, New-York, 2005.

Rossi F., Conan-Guez B., El Golli A., « Clustering Functional Data with the SOM algorithm », *Proceedings of ESANN 2004*, Bruges, Belgium, p. 305-312, April, 2004.

Silverman B. W., *Density estimation for statistics and data analysis*, Chapman and Hall, London, 1986.

Tarpey T., Kinater K. K., « Clustering Functional Data », *Journal of Classification*, vol. 20, n° 1, p. 93-114, 2003.

Vrac M., Diday E., Chédin A., Naveau P., « Mélange de distributions de distributions, décomposition de mélange de copules et application à la climatologie. », *Actes du VIIIème congrés de la Société Francophone de Classification*, p. 348-355, 2001.