A Probability Distribution of Functional Random Variable with a Functional Data Analysis Application

Cuvelier, Etienne; Fraiture, Monique Noirhomme

Published in:
The Second International Workshop on Mining Complex Data - MCD'06 - In Conjunction with IEEE ICDM'06, Hong-Kong, 2006

Publication date:
2006

Document Version
Early version, also known as pre-print

A Probability Distribution Of Functional Random Variable With A Functional Data Analysis Application

Etienne Cuvelier
Institut d'Informatique, FUNDP
Namur, Belgium
cuvelier.etienne@info.fundp.ac.be

Monique Noirhomme-Fraiture
Institut d'Informatique, FUNDP
Namur, Belgium
noirhomme.monique@info.fundp.ac.be

Abstract

Probability distributions are central tools for probabilistic modeling in data mining, and they lack in functional data analysis (FDA). In this paper we propose a probability distribution law for functional data. We build it using jointly the Quasi-arithmetic means and the generators of Archimedean copulas. We also define a density adapted to the infinite dimension of the space of functional data. For this we use the Gâteaux differential. We illustrate the utility of this tool in FDA, applying it in a mixture decomposition classification.

1 Introduction

Probability distributions are central tools for probabilistic modeling in data mining, and they lack in functional data analysis (FDA). A particular case of distributions of functions was first introduced by Diday (see [5]), in the Symbolic Data Analysis Framework, with the notion of distribution of distributions. In this work, after having defined the concept of distribution of functions, we propose to use the Quasi-arithmetic mean in conjunction with an Archimedean generator to build a probability distributions appropriate to the dimensional infinite nature of the functional data. Since a probability distribution is an incomplete tool without an associate density, we define also an appropriate density using the Gâteaux differential which is an extension of the directional differential. We finish with an application of our tools in a mixture decomposition classification on synthetic data.

2 Distribution of a functional random variable

Definition 2.1 *Let :*

- $\mathcal{D} \subseteq \mathbb{R}$ *be a closed interval of* $\mathbb{R}$*,*
- $C^n(\mathcal{D})$ *be the set of continuous, bounded functions of domain* $\mathcal{D}$ *for which their derivatives of order up to* n *are continuous on* $\mathcal{D}$*,*
- *for* $u \in C^0(\mathcal{D})$: $\|u\|_p = \left\{\int_{\mathcal{D}} |u(x)|^p \, dx\right\}^{1/p}$
- $L^p(\mathcal{D}) = \left\{u \in C^0(\mathcal{D}) \ : \ \|u\|_p < \infty\right\}$
- *for* $u, v \in L^p(\mathcal{D})$ *:* $d_p(u,v) = \|u - v\|_p$

Definition 2.2 *Let* Ω *be the set of objects whose properties can be described by a function of* $L^2(\mathcal{D})$*.*
Then a functional random variable (frv) *is any function from* Ω *to* $L^2(\mathcal{D})$ *such that:*

$$\underline{X} : \Omega \to L^2(\mathcal{D}) : \omega \mapsto X(\omega) \tag{1}$$

and, of course :

$$X(\omega) : \mathcal{D} \to \mathbb{R} : r \mapsto X(\omega)(r) \tag{2}$$

Definition 2.3 *Let* $f, g \in L^2(\mathcal{D})$*. The pointwise order between* f *and* g *on* $\mathcal{D}$ *is defined as follows :*

$$\forall x \in \mathcal{D}, f(x) \leq g(x) \iff f \leq_{\mathcal{D}} g \tag{3}$$

Definition 2.4 *The* functional cumulative distribution function (fcdf) *of a functional random variable* $\underline{X}$ *on* $\mathcal{D}$ *is given by :*

$$\begin{aligned} F_{\underline{X},\mathcal{D}}(u) &= P\{\omega \in \Omega : X(\omega)(x) \leq u(x), \forall x \in \mathcal{D}\} \\ &= P[\underline{X} \leq_{\mathcal{D}} u] \end{aligned} \tag{4}$$

where $u \in L^2(\mathcal{D})$*.*

How we can compute such a probability ? Let us start intuitively with a simple example where $\mathcal{D} = [-20, 20]$, and

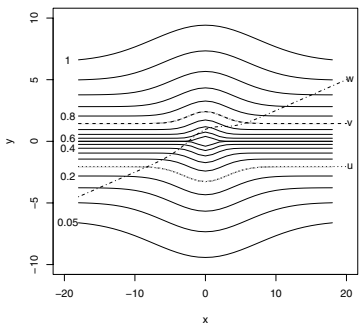

Figure 1. A example with 20 sample functions

suppose (Fig. 1) that the drawn functions form a representative sample A of a functional random variable $\underline{X}$. We can try to empirically estimate the distribution of $\underline{X}$ at u :

$$\widehat{F}_{\underline{X},\mathcal{D}}(u) = \frac{\#\{f \in A : f \leq_{\mathcal{D}} u,\}}{\#A} = \frac{1}{4}$$

In the same manner $\widehat{F}_{\underline{X},\mathcal{D}}(v) = 0.75$. But for w, if we do the same we find $\widehat{F}_{\underline{X},\mathcal{D}}(w) = 0.1$, and this in spite of the fact that w is greater than 20% of the functions of A for most of the values of $\mathcal{D}$. And thus, this way is perhaps too restrictive. That is why we propose to construct a special distribution law dedicated to this type of random variable in the next section.

3 The QAMM and QAMML distributions

Let $n \in \mathbb{N}$, $q = 2^n + 1$, and $\left\{x_1^n, \ldots, x_q^n\right\}$, q equidistant points of $\mathcal{D}$ such that $x_1^n = \inf(\mathcal{D})$ and $x_q^n = \sup(\mathcal{D})$, and $\forall i \in \{1, \ldots, q-1\}$ we have

$$\left|x_{i+1}^n - x_i^n\right| = \frac{|\mathcal{D}|}{2^n} = \frac{|\mathcal{D}|}{q} \tag{5}$$

Let :

$$\mathcal{A}_n(u) = \bigcap_{i=1}^{q} \{\omega \in \Omega : X(\omega)(x_i^n) \leq u(x_i^n)\}$$

and,

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

We will use the following approximation:

$$\begin{aligned} F_{\underline{X},\mathcal{D}}(u) = P[\mathcal{A}(u)] &\approx P[\mathcal{A}_n(u)] \\ &= H\left(u(x_1^n), \ldots, u(x_q^n)\right) \end{aligned} \tag{6}$$

where $H(\cdot, \ldots, \cdot)$ is a joint distribution of dimension q. We will use a sequence of multivariate distributions to approximate the *fcdf*, and to find a limit distribution.

Now the question is :"which H distributions are suitable for a such sequence ?". Distributions which have a matrix as parameter, like normal or Student laws, have $(q^2 - q)$ real parameters, and do not seem adapted for the evaluation of the limit $n \to \infty$. In previous works (see [5],[13],[3]) the Archimedean copulas was used for the approximation with small value of q.

Definition 3.1 *A copula is a multivariate cumulative distribution function defined on the n-dimensional unit cube* $[0,1]^n$ *such that every marginal distribution is uniform on the interval [0, 1] :*

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

The power of copulas comes from the following theorem (see [11]).

Theorem 3.1 (Sklar's theorem) *Let H be an n-dimensional distribution function with margins $F_1, ..., F_n$. Then there exists an n-copula C such that for all $x \in \bar{\mathbb{R}}^n$,*

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \tag{7}$$

If $F_1, ..., F_n$ are all continuous, then C is unique; otherwise, C is uniquely determined on Range of $F_1 \times ... \times$ Range of F_n.

An important class of copulas is the class of Archimedean copulas.

Definition 3.2 *An Archimedean copula is a function from* $[0,1]^n$ *to* $[0,1]$ *given by*

$$C(u_1, ..., u_n) = \psi\left[\sum_{i=1}^{n} \phi(u_i)\right] \tag{8}$$

where ϕ, called the generator, is a function from $[0,1]$ *to* $[0,\infty]$ *such that:*

- *ϕ is a continuous strictly decreasing function,*
- *$\phi(0) = \infty$ and $\phi(1) = 0$,*
- *$\psi = \phi^{-1}$ is completely monotonic on* $[0,\infty[$ *i.e.*

$$(-1)^k \frac{d^k}{dt^k}\psi(t) \geq 0$$

for all t in $[0,\infty[$ *and for all k.*

Before using copulas, we define a function that gives the distribution of the values of $\underline{X}(x)$ for a chosen $x \in \mathcal{D}$.

Definition 3.3 *Let $\underline{X}$ a* frv. *We define respectively the* surface of distributions *and the* surface of densities *as follow :*

$$G : \mathcal{D} \times \mathbb{R} \to [0,1] : (x,y) \mapsto P[\underline{X}(x) \leq y] \tag{9}$$

$$g : \mathcal{D} \times \mathbb{R} \to [0,1] : (x,y) \mapsto \frac{\partial}{\partial x} G\,(x,y) \tag{10}$$

Table 1. Families of completely monotonic generators

Name	Generator	Dom. of θ
Clayton	$t^{\theta}-1$	$\theta>0$
Frank	$-\ln\frac{e^{-\theta\cdot t}-1}{e^{-\theta}-1}$	$\theta>0$
Gumbel-Hougaard	$(-\ln t)^{\theta}$	$\theta\geq 1$

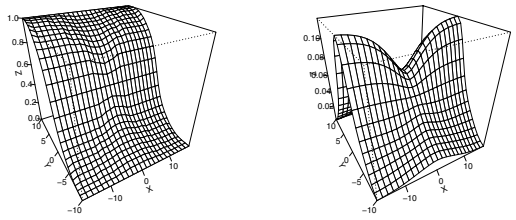

Figure 2. The surfaces $G(x,y)$ and $g(x,y)$ for the example of Fig. 1

We can use various methods for determining suitable g and G for a chosen value of x. Thus for example, if $\underline{X}$ is a Gaussian process with mean value $\mu(x)$ and standard deviation $\sigma(x)$, then we have :

$$G\left(x,y\right)=F_{\mathcal{N}(\mu(x),\sigma(x))}(y) \tag{11}$$

$$g\left(x,y\right)=f_{\mathcal{N}(\mu(x),\sigma(x))}(y) \tag{12}$$

In other cases we can use the empirical cumulative distribution function and the kernel density estimation to estimate $\hat{G}$ and $\hat{g}$:

$$\hat{G}\left(x,y\right)=\frac{\#\left\{X_i(x)\leq y\right\}}{N} \tag{13}$$

$$\hat{g}\left(x,y\right)=\frac{1}{N\cdot h(x)}\sum_{i=1}^{N}K\left(\frac{y-X_i(x)}{h(x)}\right) \tag{14}$$

The Fig. 2 shows these surfaces for the example of the Fig. 1, using the expressions (13) and (14).
If we use expression (9) in conjunction with (8), then we can use the following approximation :

$$\begin{aligned}P[\mathcal{A}_n(u)] &= C\left(G\left[x_1^n,u\left(x_1^n\right)\right],...,G\left[x_q^n,u\left(x_q^n\right)\right]\right)\\ &= \psi\left(\sum_{i=1}^{q}\phi\left(G\left[x_i^n,u\left(x_i^n\right)\right]\right)\right)\end{aligned} \tag{15}$$

The distribution (15) with the Clayton generator was already used for clustering of functional data coming from the symbolic data analysis (see [13] and [3]). Unfortunately the above limit is almost always null for Archimedean copulas!

Proposition 3.2 *If for $u\in L^2(\mathcal{D}):G(x,u(x))<1,\ \forall x\in\mathcal{D}$, then*

$$\lim_{q\to\infty}\psi\left[\sum_{i=1}^{q}\phi\left(G\left[x_i^n,u\left(x_i^n\right)\right]\right)\right]=0 \tag{16}$$

Proof Let $p=max\left\{G(x,u(x))|x\in\mathcal{D}\right\}$, and so $\forall x\in\mathcal{D}$

$$\begin{aligned}1>p\geq G\left[x,u(x)\right]\\ 0<\phi(p)\leq\phi\left(G\left[x,u(x)\right]\right)\end{aligned}$$

and thus

$$0<q\cdot\phi(p)\leq\sum_{i=1}^{q}\phi\left(G\left[x_i^n,u\left(x_i^n\right)\right]\right),$$

And then

$$\lim_{q\to\infty}\sum_{i=1}^{q}\phi\left(G\left[x_i^n,u(x_i^n)\right]\right)=\infty \qquad \blacksquare$$

Another objection to the use of this type of joint distributions is something which we could call volumetric behavior.

Definition 3.4 *A function $u\in L^2(\mathcal{D})$ is called a* functional quantile *of value p if*

$$G(x,u(x))=p,\ \forall x\in\mathcal{D} \tag{17}$$

We write $u=Q_p$ and so $G(x,Q_p(x))=p,\ \forall x\in\mathcal{D}$

The graph of the functional quantile Q_p can be seen as the level curve of value p. Now let us remark that

$$\begin{aligned}P[\mathcal{A}_n(Q_p)] &= \psi\left[\sum_{i=1}^{q}\phi\left(G\left[x_i^n,Q_p(x_i^n)\right]\right)\right]\\ &= \psi\left[\sum_{i=1}^{q}\phi(p)\right]\\ &= \psi\left(q\cdot\phi\left(p\right)\right)<p\end{aligned}$$

Then, the more we try to have a better approximation for a *functional quantile* of value p, the more we move away from these reference value toward zero.
A simple way to avoid these two problems is to use the notion of quasi-arithmetic mean, concept which was studied by Kolmogorov [8] , Nagumo [10] and Aczél [1].

Definition 3.5 *Let $[a,b]$ be a closed real interval, and $q\in\mathbb{N}_0$. A quasi-arithmetic mean is a function $M:[a,b]^q\to[a,b]$ defined as follows:*

$$M(x_1,\ldots,x_q)=\psi\left(\frac{1}{q}\sum_{i=1}^{q}\phi\left(x_i\right)\right) \tag{18}$$

where ϕ is a continuous strictly monotonic real function.

If we use the generator for Archimedean copulas in (18), we define a cumulative distribution function built from one-dimensional distributions.

Lemma 3.3 *Let $q \in \mathbb{N}_0$, F be a one dimensional cdf, and ϕ a generator of Archimedean copula , then*

$$F^*(x) = \psi\left(\frac{1}{q} \cdot \phi\left(F(x)\right)\right)$$

is also a cdf.

Proof Like F and ϕ are both continuous functions, it is easy to see that $F^*(x)$ is continuous, monotone increasing and :

$$\lim_{x \to -\infty} F^*(x) = 0$$
$$\lim_{x \to +\infty} F^*(x) = 1 \quad \blacksquare$$

Proposition 3.4 *Let $q \in \mathbb{N}_0$, $\{F_i | 1 \leq i \leq q\}$ be a set of one dimensional cdf, and ϕ a generator of Archimedean copula , then*

$$H\left(x_1, \ldots, x_q\right) = \psi\left(\frac{1}{q}\sum_{i=1}^{q} \phi\left(F_i(x_i)\right)\right) \qquad (19)$$

is a multivariate cdf.

Proof By the above lemma we have that the functions $F_i^*(x)$ are cdf, and as ϕ is an "Archimedean generator" so :

$$\psi\left(\sum_{i=1}^{q} \phi(y_i)\right)$$

is a copula,

$$\psi\left(\sum_{i=1}^{q} \phi(F_i^*(x_i))\right)$$

is a multivariate cdf. $\blacksquare$

We call the distributions given by the expression (19) the *Quasi-Arithmetic Mean of Margins (QAMM)* distributions. In fact these distributions result of the conjunction of an Archimedean copula and the transformation $\psi\left(\frac{1}{q}\phi(u)\right)$. Now we can use *QAMM* for our approximation :

$$\begin{aligned} P[\mathcal{A}_n(u)] &= \psi\left[\frac{1}{q}\sum_{i=1}^{q} \phi\left(G\left[x_i^n, u(x_i^n)\right]\right)\right] \qquad (20) \\ &= \psi\left[\frac{1}{|\mathcal{D}|}\sum_{i=1}^{q} \frac{|\mathcal{D}|}{q} \cdot \phi\left(G\left[x_i^n, u(x_i^n)\right]\right)\right] \end{aligned}$$

And if we note $\left|x_{i+1}^n - x_i^n\right| = \Delta_x$ (recall 5) $\forall i \in \{1, \ldots, q\}$, we can now take the limit.

Table 2. ***fcdf*** **with several values for the parameter of Clayton's generator**

Parameter	u	v	w
0.5	0.2382	0.7010	0.3732
2	0.2380	0.7010	0.2545
8	0.2373	0.7007	0.1408

Definition 3.6 *Let : $\underline{X}$ be a frv, $u \in L^2(\mathcal{D})$, G its* Surface of Distributions *and ϕ a generator of Archimedean Copulas. We define the* Quasi-Arithmetic Mean of Margins Limit (QAMML) *distribution of $\underline{X}$ by :*

$$\begin{aligned} F_{\underline{X},\mathcal{D}}(u) &= \lim_{n \to \infty} P[\mathcal{A}_n(u)] \\ &= \psi\left[\frac{1}{|\mathcal{D}|} \cdot \int_{\mathcal{D}} \phi\left(G\left[x, u\left(x\right)\right]\right) dx\right] \qquad (21) \end{aligned}$$

De Finetti (see [4] & [7]) was the first to extend the results of Nagumo and Kolmogorov to the case of a continous probability distribution . It is easy to see that the *QAMML* distribution preserves the *functional quantiles*.

Proposition 3.5 *If $Q_p \in L^2(\mathcal{D})$ is a* functional quantile *of value p, then $F_{\underline{X},\mathcal{D}}(Q_p) = p$*

The table 2 shows the values of the *fcdf* for the functions u, v and w from our example. But how can we choose the parameter for the *fcdf* ? The maximum likelihood is a usual method for this, but we need the notion of density.

4 The Gâteaux Density

A *fcdf* is an incomplete tool without an associate density. As long as we use finite values of n in expression (20), then we can use the classical multivariate density function:

$$h(x_1, ..., x_q) = \frac{\partial^q}{\partial x_1 \ldots \partial x_q} H(x_1, \ldots, x_q) \qquad (22)$$

But what is the matter when $n \to \infty$? Can we hope to find a limit for (22)? It seems difficult. We have seen that the *QAMML* distribution "preserve" the value of a *functional quantile*. We need to find a derivative operator which preserves this property, i.e. suppose that $0 \leq p < q \leq 1$, we search an operator D such that :

$$F_{\underline{X},\mathcal{D}}(Q_q) \approx F_{\underline{X},\mathcal{D}}(Q_p) + DF_{\underline{X},\mathcal{D}}(Q_p) d_2\left(Q_q, Q_p\right) \qquad (23)$$

As the considered functions belong to the vector space $L^2(\mathcal{D})$, this operator D must take into account the direction between Q_p and Q_q. Thus we will call upon a concept of functional analysis : the *Gâteaux differential* which is a generalization of directional derivative (see [2]).

Definition 4.1 *Suppose V and W are normed vector spaces, and F an operator from V to W. The* Gâteaux differential *$DF(u;s)$ of F at u in the direction $s \in V$ is given by:*

$$DF(u;s) = \lim_{\epsilon \to 0} \frac{F(u+\epsilon \cdot s) - F(u)}{\epsilon} \quad (24)$$

$$= F'(u) \cdot s \quad (25)$$

If (24) exists $\forall s \in L^2(\mathcal{D})$ then F is Gâteaux differentiable *and the map $F'(u)$ is the* Gâteaux derivative *of F at u.*

If we use this kind of derivative, we need to find the direction between two *functional quantiles*. Suppose that the surface of distributions G, and the surface of densities g follow a known distribution with a location parameter l and a scale parameter s. If, if we note $l(x)$ and $s(x)$ the functions which give the location and scale parameter at x, we can write

$$Q_p(x) = Q(p; l(x), s(x)) = s(x) \cdot Q(\alpha; 0, 1) + l(x) \quad (26)$$

And then the searched direction in (23) comes easily:

$$Q_q(x) - Q_p(x) = s(x) \cdot \epsilon$$

Where $\epsilon = Q(q;0,1) - Q(p;0,1)$ is a constant. And thus, we use the scale parameter as direction if the used distribution for the surfaces has such a parameter, and more generally we can use any statistical dispersion function s, like the standard deviation σ.

Definition 4.2 *Let $\underline{X}$ be a* frv*, $F_{\underline{X},\mathcal{D}}$ its* fcdf *and u a function of $L^2(\mathcal{D})$. If $s \in L^2(\mathcal{D})$ is a function such that $s(x)$ measure the statistical dispersion of the values $\underline{X}(x)$, then we define the* Gâteaux density of $F_{\underline{X},\mathcal{D}}$ at u and in direction of s *by:*

$$f_{\underline{X},\mathcal{D},s}(u) = \lim_{\epsilon \to 0} \frac{F_{\underline{X},\mathcal{D}}(u + s \cdot \epsilon) - F_{\underline{X},\mathcal{D}}(u)}{d_2(u + s.\epsilon, u)}$$

$$= \frac{DF_{\underline{X},\mathcal{D}}(u;s)}{\|s\|_2} \quad (27)$$

Where $DF_{\underline{X},\mathcal{D}}(u;s)$ is the Gâteaux differential *of $F_{\underline{X},\mathcal{D}}$ at u in the direction $s \in V$.*

To give the *Gâteaux density* for the *QAMML* distribution we need a result coming from functional analysis [9].

Proposition 4.1 *Let the following integral transform :*

$$\mathcal{T}(f) = \int_a^b K(t,s) \cdot g[s, f(s)] \, ds \quad (28)$$

where the kernel $K(s,t)$ is continuous on $[a,b]^2$, and $g(s,t)$ is a function of two variables, defined and continuous on $[a,b] \times]-\infty, +\infty[$. Then for any function $h \in C[a,b]$ we have

$$D\mathcal{T}(f,h) = \int_a^b K(t,s) \cdot g'_v[s, f(s)] \cdot h(s) \, ds \quad (29)$$

Where $D\mathcal{T}(f,h)$ is the Gâteaux differential *of $\mathcal{F}$ at f in the direction h.*

Theorem 4.2 *Let $F_{\underline{X},\mathcal{D}}$ a* fcdf*, u a function of $L^2(\mathcal{D})$. If $s \in L^2(\mathcal{D})$ is a functional measure of the statistical dispersion of the values $\underline{X}(x)$, then the* Gâteaux density of $F_{\underline{X},\mathcal{D}}$ in u and in direction of s *is given by:*

$$f_{\underline{X},\mathcal{D},s}(u) = \frac{1}{\|s\|_2 \cdot |\mathcal{D}|} \psi' \left[\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \phi(G[t,u(t)]) \, dt \right] \left\{ \int_{\mathcal{D}} \phi'(G[t,u(t)]) \, g[t,u(t)] \, s(t) \, dt \right\} \quad (30)$$

Proof It is sufficient to use (29) with $K(t,s) = \frac{1}{|\mathcal{D}|}$ $g(s,t) = \phi(G(s,t))$ and then

$$F_{\underline{X},\mathcal{D}}(u) = \psi[\mathcal{T}(u)] = (\psi \circ \mathcal{T})(u)$$

$$DF_{\underline{X},\mathcal{D}}(u;s) = \psi[\mathcal{T}(u)]' \cdot D\mathcal{T}(u;s) \quad \blacksquare$$

Now we can use a *maximum likelihood estimation* to estimate the parameter of the generator ϕ. Thus for our simple example of Fig. 1 we find the following parameter : 1.589. Let us note that in the expression (30) of the *Gâteaux density* for the *QAMML* distribution we need to calculate $g(x,y)$, and it is impossible if there is no dispersion. We can say that the domain of the *QAMML* model is the set of reals such that the statistical dispersion is greater then zero. Thus, from here, we suppose that $\mathcal{D} \subseteq \{x \in \mathbb{R} : s(x) > 0\}$, and we call $\mathcal{D}$ the *model's domain*.

5 A classification version and use

We propose now to illustrate the utility of the *QAMML* distributions in classification with a mixture decomposition classification of functional data coming from the *Symbolic Data Analysis* framework. The *Symbolic Data Analysis* summarizes concepts contained in databases using numbers, intervals, histograms and, also, probability distributions. Of course we are, here, only interested by this last case. For our test we use a synthetic dataset which contains exponential, normal and beta (see Fig. 3) distributions. The used clustering algorithm, proposed by Diday [5] is an extension of the *dynamical clustering* method [6] for density mixtures. The main idea is to estimate at each step, the density which describes at best the clusters of the current

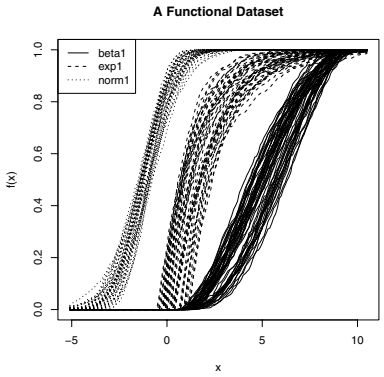

Figure 3. A cdf dataset

partition *P*, according by a given quality criterion. We considered the classifier log-likelihood :

$$lvc(P,\beta) = \sum_{i}^{K} \sum_{u \in P_i} log(f_{\underline{X},\mathcal{D}_i,\sigma}(u)) \qquad (31)$$

where σ is the *standard deviation function* and $\mathcal{D}_i$ the *model's domain* of the ith cluster. The classification starts with a random partition, then the two following steps are repeated:

- **Step 1 : Parameters estimation**
  Find the vector $(\beta_1, ..., \beta_K)$ which maximizes the chosen criterion;
- **Step 2 : Distribution of units in new classes**
  Build new classes $(P_i)_{i=1,...,K}$ with parameters found at step 1 :

$$P_i = \{u : f_{\underline{X},\mathcal{D}_i,\sigma}(u,\beta_i) \geq f_{\underline{X},\mathcal{D}_m,\sigma}(u,\beta_m) \forall m\}$$

until stabilization of the partition. We also use a classification version of the *QAMML* law. Indeed, this classification of functional data try to distinguish different probability distributions. But, two distributions are not distinguishable when they both have values equal to 1 (or both to 0). So we need to restrict the use of the *QAMML* distribution on a domain where the considered distribution is distinguishable of the others. For this we use a function $\tau : \cup_{i=1}^{K} \mathcal{D}_i \rightarrow [0,1]$ that we will call the *trust function*. And thus, expression (21) becomes :

$$\psi \left[\frac{1}{\int_{\mathcal{D}} \tau(t)dt} \int_{\mathcal{D}} \phi\left(G\left[x, u\left(x\right)\right]\right) \cdot \tau(u(x))dx \right] \qquad (32)$$

We use the following trust function $\tau : \cup_{i=1}^{K} \mathcal{D}_i \rightarrow \{0,1\}$:

$$\begin{aligned} \tau(x) &= 1 \ if \ 0 < x < 1 \\ &= 0 \ otherwise \end{aligned}$$

With this *trust function* the same changes are made to the *Gâteaux density*. The implementation of the algorithm and the *QAMML* laws was made with the R-project [12]. We choose the *Clayton* generator (see Table 1), and estimations $\hat{G}$ and $\hat{g}$ (see expressions (13) & (14)). We run the method five times and we retained the result which has the best criterion, and we obtain a misclassification rate of 0.71%, i.e. only one functional data misclassified over 135.

6 Conclusion

We have not presented here a new method for data mining, but a new mathematical tool which can be used with existing probabilistic methods. We used the classification of functional data coming from the *Symbolic Data Analysis* framework to illustrate the utility of our tool. Several ways to improve the model exist. By example let us note that the *QAMML* definition (see (21)) uses a uniform distribution over $\mathcal{D}$: other distributions can be considered (see [4]) .

References

[1] J. Aczel. *Lectures on Functional Equations and Their Applications*. Mathematics in Science and Engineering. Academic Press, New York and London, 1966.

[2] K. Atkinson and W. Han. *Theoretical Numerical Analysis*. texts in Applied Mathematics. Springer, New-York, 2001.

[3] E. Cuvelier and M. Noirhomme-Fraiture. Clayton copula and mixture decomposition. In *ASMDA 2005*, pages 699–708, 2005.

[4] B. De Finetti. Sul concetto di media. *Giornale dell' Instituto Itialiano degli Attuari*, 2:369–396, 1931.

[5] E. Diday. Mixture decomposition of distributions by copulas. In *Classification, Clustering and Data Analysis*, pages 297–310. Springer, Verlag, 2002.

[6] E. Diday, A. Schroeder, and Y. Ok. The dynamic clusters method in pattern recognition. In *IFIP Congress*, pages 691–697, 1974.

[7] G. H. Hardy, J. E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, Cambridge, 1934.

[8] A. Kolmogorov. Sur la notion de moyenne. *Rendiconti Accademia dei Lincei*, 12(6):388–391, 1930.

[9] L. A. Lusternik and V. J. Sobolev. *Elements of Functional Analysis*. Hindustan Publishing Corpn., Delhi, 1974.

[10] M. Nagumo. Uber eine klasse der mittelwerte. *Japan Journal of Mathematics*, 7:71–79, 1930.

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

[12] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[13] M. Vrac, E. Diday, A. Chédin, and P. Naveau. Mélange de distributions de distributions, décomposition de mélange de copules et application à la climatologie. In *Actes du VIIIème congrés de la Société Francophone de Classification*, pages 348–355, 2001.