Parametric families of probability distributions for functional data using quasi-arithmetic means with archimedean generators

Cuvelier, Etienne; Fraiture, Monique Noirhomme

Published in:
Functional and Operatorial Statistics

Publication date:
2008

Document Version
Early version, also known as pre-print

Link to publication

Citation for pulished version (HARVARD):
Cuvelier, E & Fraiture, MN 2008, Parametric families of probability distributions for functional data using quasi-arithmetic means with archimedean generators. in S Dabo-Niang, F Ferraty & PV HD (eds), *Functional and Operatorial Statistics.* pp. 127-133.

First International Workshop on
Functional and Operatorial Statistics.
Toulouse, June 19-21, 2008

Parametric families of probability distributions for functional data using quasi-arithmetic means with archimedean generators

Etienne Cuvelier*, Monique Noirhomme-Fraiture

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

Abstract

Parametric probability distributions are central tools for probabilistic modeling in data mining, and they lack in functional data analysis (FDA). In this paper we propose to build this kind of distribution using jointly Quasi-arithmetic means and generators of Archimedean copulas. We also define a density adapted to the infinite dimension of the space of functional data. We use these concepts in supervised classification.

1. QAMML distributions

Let $(\Omega, \mathcal{A}, P)$ a probability space and $\mathcal{D}$ a closed real interval. A *functional random variable (frv)* is any function from $\mathcal{D} \times \Omega \to \mathbb{R}$ such for any $t \in \mathcal{D}, X(t, .)$ is a real random variable on $(\Omega, \mathcal{A}, P)$. Let $L^2(\mathcal{D})$ be the space of square integrable functions (with respect to Lebesgues measure) $u(t)$ defined on $\mathcal{D}$.

If $f, g \in L^2(\mathcal{D})$, then the pointwise order between f and g on $\mathcal{D}$ is defined as follows :

$$\forall t \in \mathcal{D}, f(t) \leq g(t) \iff f \leq_{\mathcal{D}} g. \quad (1)$$

It is easy to see that the pointwise order is a partial order over $L^2(\mathcal{D})$, and not a total order. We define the *functional cumulative distribution function (fcdf)* of a *frv* $\underline{X}$ on $L^2(\mathcal{D})$ computed at $u \in L^2(\mathcal{D})$ by :

$$F_{\underline{X},\mathcal{D}}(u) = P[\underline{X} \leq_{\mathcal{D}} u]. \quad (2)$$

To compute the above probability, let us remark that, it is easy to compute the probability distribution of the value of $X(t)$ for a specific value of t, and this for any $t \in \mathcal{D}$. Then we define respectively the *surface of distributions* and the *surface of densities* as follow :

$$G : \mathcal{D} \times \mathbb{R} \to [0,1] : (t,y) \mapsto P[X(t) \leq y] \tag{3}$$

$$g : \mathcal{D} \times \mathbb{R} \to [0,1] : (t,y) \mapsto \frac{\partial}{\partial t} G\,(t,y) \tag{4}$$

We can use various methods for determining suitable g and G for a chosen value of $\underline{X}$. Thus for example, if $\underline{X}$ is a Gaussian process with mean value $\mu(t)$ and standard deviation $\sigma(t)$, then, for any $(t,y) \in \mathcal{D} \times \mathbb{R}$, we have : $G\,(t,y) = F_{\mathcal{N}(\mu(t),\sigma(t))}(y)$ and $g\,(t,y) = f_{\mathcal{N}(\mu(t),\sigma(t))}(y)$. In the following we will always use the function G with a function u of $L^2\,(\mathcal{D})$, so, for the ease of the notations, we will write : $G\,[t;u] = G\,[t, u\,(t)]$. We will use the same notation for g. In what follows we define our parametric families of probability distributions.

Let $\underline{X}$ be a frv, $u \in L^2(\mathcal{D})$ and G its *Surface of Distributions.* Let also ϕ be a continuous strictly decreasing function from $[0,1]$ to $[0,\infty]$ such that $\phi(0) = \infty$, $\phi(1) = 0$, where $\psi = \phi^{-1}$ must be 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. We define the *Quasi-Arithmetic Mean of Margins Limit (QAMML)* distribution of $\underline{X}$ by :

$$F_{\underline{X},\mathcal{D}}(u) = \psi\left[\frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \phi\,(G\,[t;u])\,dt\right]. \tag{5}$$

The function ϕ is called the *QAMML* generator. In fact the expression (5) can be seen as the limiting (or continuous) case of two other expressions. The first expression, which is obvious and gives its name to (5), use a quasi-arithmetic mean M :

$$F_{\underline{X},\mathcal{D}}(u) = \lim_{n\to\infty} M\,\{G[t_1;u], \ldots, G[t_n;u]\} \tag{6}$$

where $\{t_1, \ldots, t_n\} \subset \mathcal{D}$ is a subset of points in $\mathcal{D}$, preferably equidistant. In the discrete case, a quasi-arithmetic mean is a function $M : [a,b]^n \to [a,b]$ defined as follows:

$$M(x_1, \ldots, x_n) = \psi\left(\frac{1}{n}\sum_{i=1}^{n} \phi\,(x_i)\right) \tag{7}$$

where ϕ is a continuous strictly monotonic real function and $\psi = \phi^{-1}$.

The second limiting case links the *QAMML* distributions to the classical approximation : $P\,[\underline{X} \leq_{\mathcal{D}} u] = H\,(u(t_1), \ldots, u(t_n))$, using the archimedean copulas:

$$F_{\underline{X},\mathcal{D}}(u) = \lim_{n\to\infty} \psi\left[\sum_{i=1}^{n} \phi\,(G^*\,[t_i;u])\right] \tag{8}$$

where * is the following transformation, applied to margins:

$$G^*(x) = \psi\left(\frac{1}{n}\phi\,(G(x))\right). \tag{9}$$

Let us remind that 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]. The interest of copulas comes from the fact that (Sklar's theorem), if H is an n-dimensional distribution function with margins $F_1, ..., F_n$, then there exists an n-copula C such that for all $x \in \mathbb{R}^n$,

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \tag{10}$$

The copula captures the dependence structure of the distribution. An important family of copulas is the family of Archimedean copula, given by the following expression :

$$C(u_1, ..., u_n) = \psi\left[\sum_{i=1}^{n} \phi(u_i)\right]. \tag{11}$$

where ϕ, called the generator, has the same properties that a *QAMML* generator.
This second limiting case shows that *QAMML* shares the properties and limitations of archimedeans copulas in the modeling of an *frv* $\underline{X}$ (see the GQAMML section).

2. Gateaux density

A *fcdf* is an incomplete tool without an associate density, but as the *QAMML* distributions deal directly with infinite nature of functional data, we cannot use the classical multivariate density function:

$$h(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} H(x_1, \dots, x_n). \tag{12}$$

To solve this problem we propose to use a concept of the functional analysis : the *Gâteaux differential* which is a generalization of directional derivative. Let $\underline{X}$ be a *frv*, $F_{\underline{X},\mathcal{D}}$ its *fcdf* and u a function of $L^2(\mathcal{D})$. Then for $h \in L^2(\mathcal{D})$ we define the *Gâteaux density of* $F_{\underline{X},\mathcal{D}}$ at u and in the direction of h by:

$$f_{\underline{X},\mathcal{D},h}(u) = \lim_{\epsilon \to 0} \frac{F_{\underline{X},\mathcal{D}}(u + h \cdot \epsilon) - F_{\underline{X},\mathcal{D}}(u)}{\epsilon} = DF_{\underline{X},\mathcal{D}}(u; h) \tag{13}$$

where $DF_{\underline{X},\mathcal{D}}(u; h)$ is the *Gâteaux differential* of $F_{\underline{X},\mathcal{D}}$ at u in the direction $h \in V$.
It is easy to show that, if $F_{\underline{X},\mathcal{D}}$ is a *QAMML fcdf*, u and h are two functions of $L^2(\mathcal{D})$, then the corresponding *Gâteaux density of* $F_{\underline{X},\mathcal{D}}$ computed in u, in direction of h is given by:

$$f_{\underline{X},\mathcal{D},h}(u) = \frac{1}{|\mathcal{D}|} \cdot \psi' \left[\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \phi\left(G\left[t; u\right]\right)\, dt\right] \cdot \left\{\int_{\mathcal{D}} \phi'\left(G\left[t; u\right]\right) \cdot g\left[t; u\right] \cdot h(t)\, dt\right\}. \tag{14}$$

We can show that, if we use the statistical dispersion $\sigma(t)$ of the functional data, then $f_{\underline{X},\mathcal{D},\sigma}(u) = P[\underline{X} = u]$.

3. GQAMML distributions

QAMML shares the limitations of archimedeans copulas (see section 1), but the archimedean copulas of dimension $n > 2$, can capture dependence structures from independence until the complete positive dependence between variables. Thus, if for $s, t \in \mathcal{D}$, there is a negative dependence between $X(s)$ and $X(t)$, the *QAMML* will not be able to model the situation. But the bidimensional archimedean copulas can deal with this kind of dependence, using the same generator, but with larger domain for the parameter. Then we define the *Generalized Quasi-Arithmetic Mean of Margins Limit (GQAMML)* $\mathbb{F}_{\underline{X},\mathcal{D}}(u)$ as follows. Let $\underline{X}$ be a *frv* defined on $\mathcal{D}$, $u \in L^2(\mathcal{D})$, $\{\mathcal{D}_p, \mathcal{D}_n\}$ a partition of $\mathcal{D}$ such :

- $\forall s, t \in \mathcal{D}_p$, there is a positive dependence between $X(s)$ and $X(t)$,
- $\forall s, t \in \mathcal{D}_n$, there is a positive dependence between $X(s)$ and $X(t)$,
- $\forall s \in \mathcal{D}_p$ and $\forall t \in \mathcal{D}_n$, there is a negative dependence between $X(s)$ and $X(t)$.

Then

$$\mathbb{F}_{\underline{X},\mathcal{D}}(u) = \psi\left(\frac{|\mathcal{D}_p|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_p}(u)\right] + \frac{|\mathcal{D}_n|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_n}(u)\right]\right) \tag{15}$$

where ϕ is the generator of an bidimensional archimedean copulas.
Of course, using the chain rule, the *Gâteaux density of* $\mathbb{F}_{\underline{X},\mathcal{D}}$ is given by

$$\mathbf{f}_{\underline{X},\mathcal{D},\sigma}(u) = \psi'\left(\frac{|\mathcal{D}_p|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_p}(u)\right] + \frac{|\mathcal{D}_n|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_n}(u)\right]\right)$$
$$\left\{\frac{|\mathcal{D}_p|}{|\mathcal{D}|}\phi'\left[F_{\underline{X},\mathcal{D}_p}(u)\right] f_{\underline{X},\mathcal{D}_p,\sigma}(u) + \frac{|\mathcal{D}_n|}{|\mathcal{D}|}\phi'\left[F_{\underline{X},\mathcal{D}_n}(u)\right] f_{\underline{X},\mathcal{D}_n,\sigma}(u)\right\} \tag{16}$$

4. CQAMML distributions

In functional data analysis, we know that, some times, when we treat smooth data, there is a lot of information in the derivatives of the data. Of course we can apply the *GQAMML* distributions to the concerned derivative, but we can also consider jointly the distribution of the different derivatives. Then we define the *Complete Quasi-Arithmetic Mean of Margins Limit (CQAMML)* $\mathbb{F}^j_{i\,\underline{X},\mathcal{D}}(u)$ (with $i < j$) as follows. Let $\underline{X}$ be a *frv* defined on $\mathcal{D}$ with j successive derivatives, $u \in L^2(\mathcal{D})$ with j successive derivatives:

$$\mathbb{F}^j_{i\,\underline{X},\mathcal{D}}(u) = C\left(\mathbb{F}_{\underline{X}^{[i]},\mathcal{D}}\left(u^{[i]}\right), \ldots, \mathbb{F}_{\underline{X}^{[j]},\mathcal{D}}\left(u^{[j]}\right)\right) \tag{17}$$

where :

- $\underline{X}^{[i]}$ and $u^{[i]}$ are the ith derivatives for $\underline{X}$ and u,
- C is a n-dimensional copula.

Table 1: Results of the 10-fold cross validations

Distributions	**misclassifications**
$F_{\underline{X},\mathcal{D}}$	31.4%
$F_{\underline{X}',\mathcal{D}}$	9.4%
$F_{\underline{X}'',\mathcal{D}}$	5.5%
$\mathbb{F}^1_{0\underline{X},\mathcal{D}}$	16.5%
$\mathbb{F}^2_{1\underline{X},\mathcal{D}}$	4%
$\mathbb{F}^2_{0\underline{X},\mathcal{D}}$	9.4%

Note that the copula C is not necessarily an archimedean copula. The density of the *CQAMML* distribution is a classical joint density used with the *Gâteaux densities* of the different *GQAMML* distributions.

5. Supervised classification

To illustrate the interest of the *QAMML* families of distribution we propose to use it in a supervised classification application. To perform the classification we use the *Gâteaux density of a QAMML distribution* to build a bayesian classifier:

$$P(\omega_i|u) = \frac{\mathbf{f}_{\omega_i,\mathcal{D},h}(u) \cdot P(\omega_i)}{P(u)} \tag{18}$$

where $P(\omega_i|u)$ is the probability that u belong to the ith group, $\mathbf{f}_{\omega_i,\mathcal{D},h}(u)$ the adequate *Gâteaux density*, and $P(u)$ the probability of u (but this latter is constant for all cluster, so it is not necessary to compute it).

We compute the parameters of each cluster using the classical maximum likelihood, and the cluster of u is the cluster with the highest probability $P(\omega|u)$.

The chosen dataset is the well known spectrometric data from Tecator. The data consist in 100 channels of spectrum absorbance (wavelength from 850 nm to 1050 nm). The goal is to distinguish the data with more than 20% of fat content, from the data with less than 20% of fat content. We have performed a 10-fold cross validation on the data, the first derivative, the second derivative using the GQAMML distributions, and jointly on the different derivatives using the CQAMLL distributions, and this with the following parametrization :

- Surface of distributions G : Normal distribution,
- QAMML and GQAMML generators : Clayton generator,
- CQAMML copula : Normal copula.

The table 1 shows the results, and we can see that the best results are given using the distribution of the second derivative, and also considering jointly the distribution of the first and second derivative, but it is well known that the second derivative of these data contains the more interesting information to distinguish the clusters. We can also remark that when we use directly the functionnal data jointly with the derivatives, the quality of the classification decrease, but we know that original functions contain only slight differences between the two groups.

6. Conclusions

The good results of the supervised classification example show that our new families of parametric distributions for functional data can be used in classifications task in FDA. These distributions can be used also in unsupervised classification with existing algorithms. And a lot of parametrization can be chosen using existing copulas in the different level of the QAMML families, and other choices for the distributions of the surface of distributions can be done. So a great field of experimentation is open with the QAMML families of distributions for functional data.

References

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

[2] Cuvelier E. and Noirhomme-Fraiture M., Classification de fonctions continues à l'aide d'une distribution et d'une définies dans un espace de dimension infinie, Conférence EGC07- Namur, Belgique, (2007)

[3] Cuvelier E. and Noirhomme-Fraiture M., A probability distribution of functional random variable with a functional data analysis application, ICDM 06 Conference - MCD 06 Workshop, Hong-Kong, (2006)

[4] Cuvelier E. and Noirhomme-Fraiture M., An approach to stochastic process using quasi-arithmetic means, International Symposium on Applied Stochastic Models and Data Analysis (ASMDA), Crete - Chania, (2007)

[5] Joe, H., Multivariate models and dependence concepts, Chapman and Hall, London,(1997)

[6] Kolmogorov A., Sur la notion de moyenne, Rendiconti Accademia dei Lincei, vol. 12, pages 388-391, number 6, (1930)

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

[8] Nelsen R.B., An introduction to copulas, Springer, London, (1999)