Available online at www.sciencedirect.com
SCIENCE DIRECT®
Discrete Applied Mathematics 147 (2005) 27–41
DISCRETE APPLIED MATHEMATICS
www.elsevier.com/locate/dam

Mixture decomposition of distributions by copulas in the symbolic data analysis framework

E. Diday[a], M. Vrac[a, b]

[a] *Centre de Recherche de Mathématiques de la Décision (CEREMADE), Université Paris IX Dauphine, Place du Maréchal de Lattre-de-Tassigny, 75775 Paris, France*
[b] *Laboratoire de Météorologie Dynamique (LMD)/Institut Pierre Simon Laplace, Ecole Polytechnique, 91128 Palaiseau, France*

Received 9 December 2001; received in revised form 16 June 2003; accepted 21 June 2004

Abstract

This work investigates the situation in which each unit from a given set is described by some vector of p probability distributions. Our aim is to find simultaneously a "good" partition of these units and a probabilistic description of the clusters with a model using "copula functions" associated with each class of this partition. Different copula models are presented. The mixture decomposition problem is resolved in this general case. This result extends the standard mixture decomposition problem to the case where each unit is described by a vector of distributions instead of the traditional classical case where each unit is described by a vector of single (categorical or numerical) values. Several generalizations of some standard algorithms are proposed. All these results are first considered in the case of a single variable and then extended to the case of a vector of p variables by using a top-down binary tree approach.

Keywords: Clustering; Copulas; Data mining; Mixture decomposition; Partitioning; Symbolic data analysis

E-mail addresses: diday@ceremade.dauphine.fr (E. Diday), vrac@lmd.polytechnique.fr (M. Vrac).

0166-218X/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2004.06.018

1. Introduction

In a symbolic data table, a cell can contain a distribution (Schweizer, in 1984, says that "distributions are the numbers of the future"!), or several values linked by a taxonomy and logical rules, etc. The need to extend standard data analysis methods (exploratory analyses, clustering, factorial analyses, discrimination methods, etc.) to symbolic data tables is increasing with the need to obtain more accurate information from and to summarise large data sets. For more details on symbolic data analysis see for example Diday [4], Bock and Diday [1]. The idea of operating with distribution functions as data values has been applied in clustering by Janowitz and Schweizer [10]. We are interested here in extending the mixture decomposition problem (as defined for instance in Dempster et al. [3]) to the case where the units are described by distributions. Here, a "variable" Y_j is considered to be a random variable from a set of units Ω to an infinite set of distributions $\mathbf{F}_j$. We consider a sample of N units, summarised in a data table of N rows and p columns, where the ith row is associated with the unit (or "individual") $w_i \in \Omega$. The set of units of this data table is denoted by $\mathbf{E}$ and each column is defined by a variable $Y_j \in \{Y_1, \ldots, Y_p\}$. Each cell (i, j) of this data table contains a distribution $Y_j(w_i) \in \mathbf{F}_j$. The sample of the $N \times p$ distributions of this data table is called the "distribution base".

We first consider the case of a single variable Y (considered later as one of the variables in $\{Y_1, \ldots, Y_p\}$), such that $Y(w) \in \mathbf{F}$, with $\mathbf{F}$ some infinite set of distributions. The case of several variables will be considered in Section 6. The cell associated with a given unit w_i in the column for the variable Y, contains a distribution denoted by $F_i = Y(w_i)$. We denote by X_i the random variable associated with F_i such that $F_i(t) = Pr(X_i \leqslant t)$ for $t \in \overline{\mathbb{R}}$ where $\overline{\mathbb{R}} = [-\infty, +\infty]$. The distribution base is here reduced to the sample set $\mathfrak{F} = \{F_i | i = 1, \ldots, N\}$ of all the distributions contained in the column associated with the variable Y in the data table.

We define the notion of "distribution of distributions", empirically introduced by Diday [5], and developed in a more general probabilistic context by Vrac et al. [15]. Let $\mathbb{R}$ be the set of the possible values for all the continuous random variables X_i, $i = 1, \ldots, N$. Then its σ-algebra is the borelian σ-algebra, denoted by ν. The set $\mathbf{F}$ can be written as

$$\mathbf{F} = \{F | F \text{ is a 1-dimensional distribution function on } (\mathbb{R}, \nu)\}.$$

We define $\mathscr{A}$ to be the σ-algebra on $\mathbf{F}$, as the σ-algebra generated by the sets $A_t^x = \{F \in \mathbf{F} / F(t) \leqslant x\}$, for all $x \in [0, 1]$ and $t \in \mathbb{R}$. Then, in our study, Y is a random variable, where, for each w in Ω, $Y(w)$ is a one-dimensional distribution function $Y(w) = F_w \in \mathbf{F}$:

$$Y : (\Omega, \mathscr{M}, \mathbb{P}) \longrightarrow (\mathbf{F}, \mathscr{A}),$$

$$w \mapsto F_w \in \mathbf{F},$$

with $\mathscr{M}$ the σ-algebra generated on Ω by the sets of units w from Ω, and $\mathbb{P}$ a probability measure on $(\Omega, \mathscr{M})$. We can easily remark that for all $a \in \mathscr{A}$, we have $Y^{-1}(a) \in \mathscr{M}$. The set $\{w \in \Omega | F_w(t) \leqslant x\}$ is measurable by $\mathbb{P}$:

$$Y^{-1}(\{F \in \mathbf{F} | F(t) \leqslant x\}) = \{w \in \Omega | F_w(t) \leqslant x\} \in \mathscr{M}.$$

Definition 1. A "distribution function of distributions" (DFD), associated with Y at a given point $t \in \mathbb{R}$, is defined by

$$G_t(x) = \mathbb{P}(\{w \in \Omega \mid F_w(t) \leqslant x\}),$$

where $x \in \overline{\mathbb{R}}$. Notice that $G_t(x) = 1$ if $x \geqslant 1$ and $G_t(x) = 0$ if $x < 0$.

A DFD can also be interpreted as the distribution function of the random variable y_t from Ω to $\overline{\mathbb{R}}$ such that $y_t(w) = F(t)$ with $Y(w) = F$: we have $\mathbb{P}([y_t(w) \leqslant x]) = \mathbb{P}(\{w \in \Omega \mid F_w(t) \leqslant x\})$. We can compute the empirical frequency function $\overline{G_t}$ by

$$\overline{G}_t(x) = \frac{card(\{F_i \in \mathfrak{F} | F_i(t) \leqslant x\})}{N}.$$

Hence, $\overline{G}_t(x)$ is the frequency of units whose probability of taking a value less than t is equal or less than x. For instance, if the units are pupils and the variable Y is their results in mathematics, then $\overline{G}_{10}(\frac{1}{2})$ is the frequency of pupils whose probability of having a mark less than 10 in mathematics is equal or less than $\frac{1}{2}$.

Definition 2. A "k-joint distribution function of distributions" (k-JDFD) associated with Y at a given point $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$ is defined by

$$H_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = \mathbb{P}(\{w \in \Omega \mid F_w(t_1) \leqslant x_1; \ldots; F_w(t_k) \leqslant x_k\}),$$

where $x = (x_1, \ldots, x_k) \in \overline{\mathbb{R}}^k$. Notice that $H_t(x) = 0$ if $x_i < 0$ for some i and that $H_t(x) = 1$ if, for all $i = 1, \ldots, k$, $x_i \geqslant 1$.

A k-JDFD can be interpreted as the distribution of the random variable: $y = (y_{t_1}, \ldots, y_{t_k})$. Therefore, we can also write $H_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = \mathbb{P}([y_{t_1} \leqslant x_1] \cap \cdots \cap [y_{t_k} \leqslant x_{t_k}])$.

In the following, we suppose that $t_1 < \cdots < t_k$.

Proposition 1. *For all $t \in \mathbb{R}$, G_t is a distribution function. For all $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$, H_t is a k-dimensional joint distribution function with marginal distributions $G_{t_1}, \ldots, G_{t_k}$.*

Proof. This result comes from the definition of a distribution function and a n-dimensional distribution function (see Diday [5] and Vrac [14]).

From Proposition 1, we can define a "k-joint density of distributions" by

$$h_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = \frac{\partial^k H_{t_1,\ldots,t_k}(x_1, \ldots, x_k)}{\partial x_1 \ldots \partial x_k}.$$

This function is the probability density function associated with the random variable $y = (y_{t_1}, \ldots, y_{t_k})$.

2. Link with copulas

A copula (Schweizer and Sklar [12]) is a function relating the joint distribution function of a vector of k random variables $(X_1, \ldots, X_k)$ and its k one-dimensional distribution functions (i.e. the marginal distributions). The following definition is given by Nelsen [11].

Definition 3 (*k-Copula [11,12]*). A k-dimensional copula (or k-copula) is a function C from $[0,1]^k$ to $[0,1]$ with the following properties:

(1) For every $u \in [0,1]^k$, $C(u) = 0$ if at least one coordinate of u is 0.
(2) If all coordinates of u are 1 except one which is $u*$, then $C(u) = u*$.
(3) For every $\mathbf{a}=(a_1,\ldots,a_k)$ and $\mathbf{b}=(b_1,\ldots,b_k)$ in $[0,1]^k$ such that $\mathbf{a} \leqslant \mathbf{b}$, $V_C([\mathbf{a},\mathbf{b}]) \geqslant 0$, where

$$V_C([\mathbf{a},\mathbf{b}]) = \Delta_a^b C(t) = \Delta_{a_k}^{b_k} \Delta_{a_{k-1}}^{b_{k-1}} \ldots \Delta_{a_1}^{b_1} C(t)$$

and

$$\Delta_{a_j}^{b_j} C(t) = C(t_1,\ldots,t_{j-1},b_j,t_{j+1},\ldots,t_k) - C(t_1,\ldots,t_{j-1},a_j,t_{j+1},\ldots,t_k)$$

is the first order difference of C for the jth coordinate.

For example, in two dimensions ($k=2$), the third condition gives: $C(a_2,b_2)-C(a_2,b_1)-C(a_1,b_2)+C(a_1,b_1) \geqslant 0$. In the following, we denote $Ran\ G$ as the range of the mapping G. For a copula with the properties of Definition 3, Sklar gave the following theorem:

Theorem 1 (*Sklar [13]*). *Let H be a k-dimensional distribution with marginal distributions $G_1,\ldots,G_k$. Then there exists a k-copula C such that for all $(x_1,\ldots,x_k) \in [0,1]^n$,*

$$H(x_1,\ldots,x_k) = C(G_1(x_1),\ldots,G_k(x_k)). \tag{1}$$

Moreover, if $G_1,\ldots,G_k$ are continuous, then C is unique; otherwise C is uniquely determined on $Ran\ G_1 \times \cdots \times Ran\ G_k$. Conversely, if $G_1,\ldots,G_k$ are distribution functions and C is a copula, the function H defined by (1) *is a k-dimensional distribution function with marginal distributions $G_1,\ldots,G_k$.*

From Proposition 1 and Sklar's Theorem 1, we obtain the following proposition:

Proposition 2. *Let $H_{t_1,\ldots,t_k}$ be a k-joint distribution function of distributions. Let $G_{t_1},\ldots,G_{t_k}$ be k distributions functions of distributions. If $G_{t_1},\ldots,G_{t_k}$ are the k marginal distributions of $H_{t_1,\ldots,t_k}$, then there exists a k-copula C such that for all $\mathbf{x}=(x_1,\ldots,x_k) \in [-\infty,+\infty]^k$:*

$$H_{t_1,\ldots,t_k}(x_1,\ldots,x_k) = C(G_{t_1}(x_1),\ldots,G_{t_k}(x_k)). \tag{2}$$

Moreover, if $G_{t_1},\ldots,G_{t_k}$ are continuous, then C is unique; otherwise C is uniquely determined on $Ran\ G_{t_1} \times \cdots \times Ran\ G_{t_k}$. Conversely, if $G_{t_1},\ldots,G_{t_k}$ are distribution functions of distributions and C is a copula, the function $H_{t_1,\ldots,t_k}$ defined by (2) *is a k-joint distribution function of distributions with marginal distributions $G_{t_1},\ldots,G_{t_k}$.*

2.1. Parametric families of copulas

The simplest copulas, denoted by M, Π and W, respectively are

$$M(u,v) = min(u,v), \quad \Pi(u,v) = uv \quad \text{and} \quad W(u,v) = max(u+v-1,0).$$

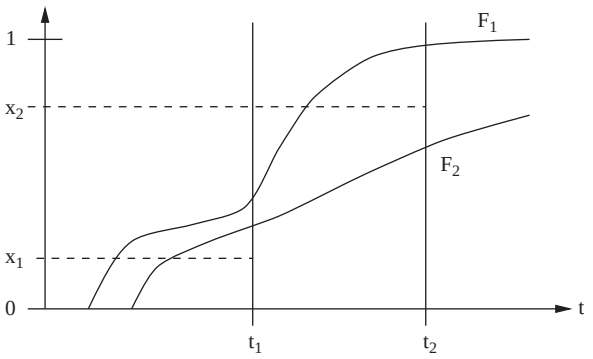

Fig. 1. The distribution base $\mathfrak{F}$ is reduced to F_1 and F_2.

These copulas are special cases of some parametric families of copulas such as the following:

(1) Clayton [2] proposed the copula $C_b(u, v) = max([u^{-b} + v^{-b} - 1]^{-1/b}, 0)$ where $b \in [-1, \infty)\backslash\{0\}$ with the following special cases: $C_{-1} = W$, $C_0 = \Pi$, $C_\infty = M$.
(2) Frank [7,8], has defined $F_b(u, v) = -1/b\, \ell n(1 + (e^{-bu} - 1)(e^{-bv} - 1)/(e^{-b} - 1))$ where $b \in \overline{\mathbb{R}}\backslash\{0\}$ and with the following special cases: $F_{-\infty} = W$, $F_0 = \Pi$, $F_\infty = M$.

Many other families are given by Nelsen [11] and some methods for the estimation of the parameters of parametric copulas are given by Vrac [14].

Notice that the Clayton and Frank parametric families of copulas are triangular norms (called t-norms). Indeed, copulas and t-norms are related: a copula with the associativity condition is a t-norm, and a t-norm with the Lipschitz condition (given in two dimensions in Lemma 1) is a copula.

Lemma 1 (*Lipschitz condition*). *Let H be an increasing 2-dimensional function, defined on $\overline{\mathbb{R}}^2$ with marginal distributions F and G. Then for all (x_1, y_1) and (x_2, y_2) in $\overline{\mathbb{R}}^2$*

$$|H(x_2, y_2) - H(x_1, y_1)| \leqslant |F(x_2) - F(x_1)| + |G(y_2) - G(y_1)|.$$

Example. The distribution base $\mathfrak{F}$ (see Fig. 1) consists of two distributions F_1 and F_2 and the copula C is defined by $C(u, v) = \overline{H}_{t_1,t_2}(x_1, x_2)$ with

(1) $u = \overline{G}_{t_1}(x_1) = card(\{F_i \in \mathscr{F} / F_i(t_1) \leqslant x_1\})/N$,
(2) $v = \overline{G}_{t_2}(x_2) = card(\{F_i \in \mathscr{F} / F_i(t_2) \leqslant x_2\})/N$,
(3) $\overline{H}_{t_1,t_2}(x_1, x_2) = card(\{F_i \in \mathscr{F} / F_i(t_1) \leqslant x_1\} \cap \{F_i \in \mathfrak{F} / F_i(t_2) \leqslant x_2\})/N$.

The possible values of $\overline{G}_{t_1}(x_i)$ and $\overline{G}_{t_2}(x_j)$ are only 0, $\frac{1}{2}$, 1. For instance, from $u = \overline{G}_{t_1}(x_0) = 0$, $v = \overline{G}_{t_2}(x_2) = \frac{1}{2}$ and $\overline{H}_{t_1,t_2}(x_1, x_2) = 0$, it follows that $C(0, \frac{1}{2}) = 0$. In the same way, we obtain $C(0, 0) = C(0, \frac{1}{2}) = C(0, 1) = 0$, $C(1, 0) = 0$, $C(1, \frac{1}{2}) = \frac{1}{2}$, $C(1, 1) = 1$.

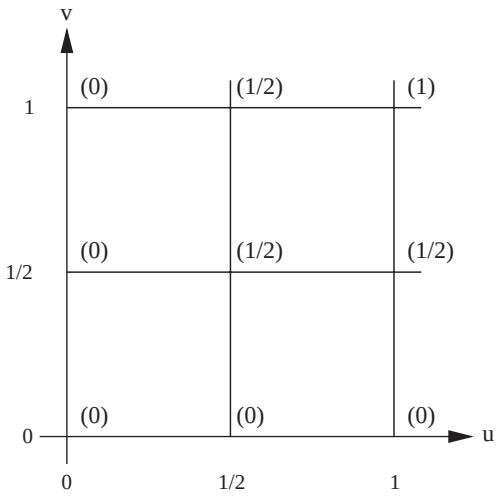

Fig. 2. The copula values $C(u, v)$ are given between brackets () and associated with each $(u, v) = (G_{t_1}(x_1), G_{t_2}(x_2))$.

For each calculation of $C(u, v)$ (see Fig. 2), the model $C \equiv Min$ is satisfied. This can be proved more generally for any pair of similar distribution functions without crossing (i.e. as long as $F_1(t) \geqslant F_2(t)$ for all t).

3. Fit between a unit and a JDFD by using an approximation of the density

We define a mapping a which measures the "fit" between a distribution $Y(w) = F_w$ and a given 2-JDFD $H_{t_1,t_2} = C(G_{t_1}, G_{t_2})$, by setting $a : \Omega \times [0, 1]^2 \longrightarrow \mathbb{R}^+$ with $a(w, \varepsilon) = [Y(w)R(\varepsilon)C(G_{t_1}, G_{t_2})] \in \mathbb{R}^+$, where $\varepsilon = (\varepsilon_1, \varepsilon_2)$ is a threshold. Here $x_i = F_w(t_i)$ and $R(\varepsilon)$ is defined by $F_w R(\varepsilon) H_{t_1,t_2} = V_{H_{t_1,t_2}}([\mathbf{a}, \mathbf{b}])$, with $\mathbf{a} = (x_1 - \varepsilon_1, x_1 + \varepsilon_1)$ and $\mathbf{b} = (x_2 - \varepsilon_2, x_2 + \varepsilon_2)$. That is,

$$\begin{aligned} F_w R(\varepsilon) C(G_{t_1}, G_{t_2}) = {} & C(G_{t_1}(x_1 + \varepsilon_1), G_{t_2}(x_2 + \varepsilon_2)) \\ & - C(G_{t_1}(x_1 + \varepsilon_1), G_{t_2}(x_2 - \varepsilon_2)) \\ & - C(G_{t_1}(x_1 - \varepsilon_1), G_{t_2}(x_2 + \varepsilon_2)) \\ & + C(G_{t_1}(x_1 - \varepsilon_1), G_{t_2}(x_2 - \varepsilon_2)). \end{aligned} \tag{3}$$

Notice that due to Proposition 1, G_{t_1} and G_{t_2} are increasing, so we have $G_{t_i}(x_i + \varepsilon_i) \geqslant G_{t_i}(x_i - \varepsilon_i)$.

Proposition 3. *Let h be the density function of the random variable $y = (y_{t_1}, y_{t_2})$ defined in Section 1. Let $a(w, \varepsilon) \in [0, 1]$. If $\partial\varepsilon = 4\varepsilon_1\varepsilon_2$, then $h_\varepsilon(x) = a(w, \varepsilon)/\partial\varepsilon$ is an approximation of $h(x_1, x_2)$ and $h_\varepsilon(x) \to h(x)$ a.s. when $\partial\varepsilon \to 0$.*

Proof. This can be proved easily from the fact that we can show $a(w,\varepsilon) = H_{t_1,t_2}(x_1 + \varepsilon_1, x_2 + \varepsilon_2) - H_{t_1,t_2}(x_1 + \varepsilon_1, x_2 - \varepsilon_2) - H_{t_1,t_2}(x_1 - \varepsilon_1, x_2 + \varepsilon_2) + H_{t_1,t_2}(x_1 - \varepsilon_1, x_2 - \varepsilon_2)$. There follows indeed:

$a(w,\varepsilon) = \mathbb{P}(\{w \in \Omega / x_1 - \varepsilon_1 \leqslant F_w(t_1) \leqslant x_1 + \varepsilon_1;\ x_2 - \varepsilon_2 \leqslant F_w(t_2) \leqslant x_2 + \varepsilon_2\}) \in [0,1]$. By the definition of a density, for each x, $\lim_{\varepsilon \to 0} h_\varepsilon(x) = h(x)$ a.s.

4. The mixture decomposition problem of distributions and two algorithms for solving it

The distributions base $\mathfrak{F} = \{F_1, \ldots, F_N\}$ is a description of the set of units $\mathbf{E} = \{w_1, \ldots, w_N\} \subset \Omega$. The set $\mathfrak{F}$ is considered to be a sample of N observations from a random variable Y such that, for $w \in \Omega$, $Y(w) \in \mathbf{F}$, with $\mathbf{F}$ an infinite set of distributions (see Section 1). The k-joint distribution function of distributions $H_{t_1,\ldots,t_k}(x_1, \ldots, x_k)$ associated with $\mathfrak{F}$ is denoted by $H(x)$. We denote $P = (P_1, \ldots, P_\ell)$ to be a partition of $\mathfrak{F}$ such that each class P_i can be considered as a sample of observations from a random variable $\mathbb{Y}_i : \Omega \to \mathbf{F}$ such that $\mathbb{Y}_i(w) \in \mathbf{F}$. The k-JDFD $H^i_{t_1,\ldots,t_k}(x_1, \ldots, x_k; \alpha_i)$ (denoted by $H_i(x; \alpha_i)$) associated with P_i depends on a parameter α_i. The mixture decomposition problem of distribution functions of distributions can be resolved in the following way: find a partition $P = (P_1, \ldots, P_\ell)$ of $\mathfrak{F}$, the mixture ratios $(p_1, \ldots, p_\ell)$ and the parameters $(\alpha_1, \ldots, \alpha_\ell)$ such that

$$H(x) = \sum_{i=1}^{\ell} p_i H_i(x; \alpha_i). \tag{4}$$

It can be shown (see Section 5) that the standard mixture decomposition problem with standard (classical) data is a special case of this general problem. In order to solve this latter problem, we reformulate it in terms of a mixture decomposition of density functions, by setting in the case of H:

$$h(x) = \frac{\partial H(x)}{\partial x_1 \ldots \partial x_k} \quad \text{and} \quad h_i(x, \alpha_i) = \frac{\partial H_i(x; \alpha_i)}{\partial x_1 \ldots \partial x_k}.$$

Then, the mixture decomposition equation (4) becomes

$$h(x) = \sum_{i=1}^{\ell} p_i h_i(x; \alpha_i). \tag{5}$$

Notice that in the case of $k = 2$, the link between h and H and the copula model is given by $h(x) = \partial^2 C(G_{t_1}(x_1).G_{t_2}(x_2))/\partial x_1 \partial x_2$. Then,

$$h(x) = \frac{\partial G_{t_1}(x_1)}{\partial x_1} \times \frac{\partial G_{t_2}(x_2)}{\partial x_2} \times \frac{\partial^2 C(G_t(x_1), G_t(x_2))}{\partial u_1 \partial u_2}.$$

The parameters $\alpha_i = (d_i, b_i)$ depend on the parameters of the chosen copula family model (for instance, the Frank family, see Section 2) denoted b_i and on the parameters of the chosen

distribution family model denoted d_i. In order to approximate the density function $h_i(x, \alpha_i)$ or to calculate it, we can use the method presented in Section 3:

$$h_i(x; \alpha_i, \varepsilon_i)\partial\varepsilon_i = a_i(w; \alpha_i, \varepsilon_i) = [Y(w)R(\varepsilon_i)C_i(G^i_{t_1}(x_1; d_{i_1}), G^i_{t_2}(x_2; d_{i_2}), b_i)], \quad (6)$$

where $x = (x_1, x_2) = (F_w(t_1), F_w(t_2))$, $F_w = Y(w)$, $\alpha_i = (d_i, b_i)$ with $d_i = (d_{i_1}, d_{i_2})$, $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i})$ and $\partial\varepsilon_i = 4\varepsilon_{1i}\varepsilon_{2i}$. Hence, the mixture decomposition model can be resolved in the following way: $h(x) = \sum_{i=1}^{\ell} p_i h_i(x, \alpha_i, \varepsilon_i)$.

Given the models associated with G and C, the decomposition can be obtained by maximizing a criterion based on the likelihood (see Diday et al. [6]). With $F_w = Y(w)$, $x_i = F_w(t_i)$, $x = (x_1, \ldots, x_k)$ and $\alpha = (\alpha_1, \ldots, \alpha_\ell)$ the parameters of the densities h_i, this criterion can be the likelihood function:

$$L(x, \alpha) = \prod_{i=1}^{N} \sum_{k=1}^{\ell} p_k k_k(x_i, \alpha_k),$$

the log-likelihood function:

$$\log L(x, \alpha) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{\ell} p_k h_k(x_i, \alpha_k)\right),$$

the classifier likelihood function:

$$CL(x, \alpha) = \prod_{i=1}^{\ell} p_i \prod_{w \in P_i} h_i(x, \alpha_i)$$

or, the classifier log-likelihood function:

$$C \log L(x, \alpha) = \sum_{i=1}^{\ell} \sum_{w \in P_i} \log(p_i h_i(x, \alpha_i)).$$

We suggest the two following algorithms where

Input: a set **E** of units described by distributions, a given partition $(P_1, \ldots, P_\ell)$, a parametric copula family $\mathbb{C}$, and optionally a parametric distribution family law $\mathbb{G}$.

Output: a partition and a copula C_i for each class, and optionally a distribution law for each G_i at each t_i.

Algorithm 1. This algorithm is defined in two steps called representation and allocation (see Diday et al. [6]):

- *Step* 1, *representation*: estimation of the parameters $(\alpha_1, \ldots, \alpha_\ell)$ which maximise the chosen criterion (L, $\log L$, etc).
- *Step* 2, *allocation*: creation of a new partition $(P_1, \ldots, P_\ell)$, where P_i is the set of units in the partition such that

$$P_i = \{x | p_i h_i(x, \alpha_i) \geqslant p_m h_m(x, \alpha_m), \text{ with } i < m \text{ in case of equality}\}.$$

When the criterion is bounded (which is the case of CL or $C \log L$), it is easy to show that this algorithm converges. There are several variants for the choice of p_i, for instance $p_i = card(P_i)/N$ at the last step, or at each step.

Algorithm 2. This algorithm is based on the two steps of the EM algorithm (Dempster et al. [3]). We start from an initial solution (p_i^0, α_i^0) at step $n = 0$ for $i = 1, \ldots, \ell$ and then at step n we have:

- *E-step* (*Estimation*): $t_i^n(x) = p_i^n h_i(x, \alpha_i^n)/\sum_{w\in\Omega} p_i^n h_i(x, \alpha_i^n)$ for $i = 1, \ldots, \ell$ and any $w \in \Omega$, which is the posterior conditional probability than an individual w belongs to the class i at the nth iteration.
- *M-step* (*Maximisation*): this step consists of estimating α_i^{n+1} $(i = 1, \ldots, \ell)$, which are the roots of:

$$\sum_{w\in\Omega} t_i^n(Y(w)) \frac{\partial \log h(Y(w), \alpha_i^{n+1})}{\partial \alpha_i} = 0,$$

with $p_i^{n+1} = 1/N \sum_{w\in\Omega} t_i^n(Y(w))$.

From the estimated parameters, a partition can be obtained by the maximum a posteriori (MAP) principle but in that case, the partition is "biased": the densities of the classes of the resultant partitions are not the estimated densities given by the EM algorithm.

Example. Suppose the data table contains five units $\mathbf{E} = \{w_1, \ldots, w_5\}$ with $F_1(t_1) = 0.1$, $F_1(t_2)=0.4$; $F_2(t_1)=0.2$, $F_2(t_2)=0.3$; $F_3(t_1)=0.6$, $F_3(t_2)=0.7$; $F_4(t_1)=0.7$, $F_4(t_2)=0.8$; $F_5(t_1) = 0.8$, $F_5(t_2) = 0.9$. The given copula parametric family is defined by $C(u, v) = b_1 M + b_2 W$, where $b_i \in \{0, 1\}$ with $b_1 = 0$ if $b_2 = 1$, $b_1 = 1$ if $b_2 = 0$ (the copulas W and M are defined in Section 2). By applying Algorithm 1, the process has converged towards the partition: $(P_1^2, P_2^2) = \{\{F_1, F_2\}, \{F_3, F_4, F_5\}\}$ and the mixture decomposition

$$H(x) = \sum_{i=1}^{2} p_i H_i(x; \alpha_i) = \frac{2}{5} W(G_{t_1}^1(x_1), G_{t_2}^1(x_2)) + \frac{3}{5} M(G_{t_1}^2(x_1), G_{t_2}^2(x_2)),$$

where $x = (x_1, x_2)$ and $G_{t_j}^i$ is defined by the empirical distribution of the class P_i^2 at t_j. See Diday [5] for details.

5. The special case of the standard mixture decomposition problem

5.1. Properties of a distribution base of unit mass distributions

Our aim in this section is to imbed the standard mixture decomposition problem into the mixture decomposition of distributions problem, in the case of a single quantitative random variable $Z : \Omega \longrightarrow \overline{\mathbb{R}}$. Each value taken by an individual w can be transformed easily into a distribution which takes the value 0 until $Z(w)$ (not included) and the value 1 otherwise. Such a distribution is called "unit mass". More formally, if $Z(w_i) = z_i$, the distribution associated

with w_i is defined by $F_i(t) = \mathbb{P}(X_i \leqslant t)$, where the random variable X_i associated with w_i is such that its distribution F_i satisfies: $F_i(t) = 0$ if $t < z_i$, and $F_i(t) = 1$ if $t \geqslant z_i$.

Proposition 4. *If the distribution base $\mathfrak{F}$ contains only unit mass distributions F_i as above, if F_z is the distribution associated with the random variable Z, if t_i increase with i, and if G_{t_i} is empirically modeled, we have the following results*:

(i) *If $H(x_1, \ldots, x_p) = C(G_{t_1}(x_1), \ldots, G_{t_p}(x_p))$, then C is the M copula.*
(ii) *If $x_p < 1$, then $Min(G_{t_1}(x_1), \ldots, G_{t_p}(x_p)) = G_{t_p}(x_p)$.*
(iii) *If $x < 1$, then $G_t(x) = Pr(Z > t) = 1 - F_z(t)$.*
(iv) *If $x_p < 1$, then $F_z(t_p) = 1 - H(x_1, \ldots, x_p)$.*

Proof. We need the following Lemma:

Lemma 2. *If $\mathfrak{F}$ is a set of unit mass distributions, $x_i \in [0, 1[$ for $i = 1, \ldots, j$, $A_j = \{F \in \mathfrak{F} \mid F(t_j) = 0\}$ and $B_j = \{F \in \mathfrak{F} \mid F(t_i) \leqslant x_i, 1 \leqslant i \leqslant j\}$, then we have $A_j = B_j$ and $|A_j| = Min_{i=1,\ldots,j} |A_i|$.*

Proof of the Lemma. Suppose we have $F \in B_j$. Then $F \in \mathfrak{F}$ and $F(t_j) < x_j$ by definition of B_j. As $\mathfrak{F}$ is a set of unit mass distributions and $x_j \in [0, 1[$, we have necessarily $F(t_j) = 0$. Therefore, $F \in A_j$ and we have $B_j \subseteq A_j$.

Suppose now $F \in A_j$. Then $F(t_j) = 0$, which implies $F(t_i) = 0$ for all $i = 1, \ldots, j$ as F is increasing (it is a distribution) and $t_i \leqslant t_{i+1}$. Thus, $F \in B_j$ and therefore $A_j = B_j$. Since by the definition of B_j, we have $B_j = \bigcap_{i=1}^{j} A_i$ and moreover we have proved that $A_j \subseteq B_j$, it follows that $|A_j| = Min_{i=1,\ldots,j} |A_i|$. We have $A_j \subseteq B_j$.

With this lemma, we can now prove Proposition 4.

Proof of Proposition 4. (i) If $H(x_1, \ldots, x_p) = C(G_{t_1}(x_1), \ldots, G_{t_p}(x_p))$, then C is the M copula.

This can be proved in the following way: If all x_i are equal to 1, since all the elements of a distribution base take a value smaller than 1 everywhere, we have by definition of a distribution function of distributions $G_{t_i}(x_i) = 1$ and also, by definition of a k-joint distribution function of distributions, $H(x_1, \ldots, x_p) = 1$. Thus, in that case (i) is true. Suppose now that some x_i are smaller than 1 and suppose we denote them by $x'_1, \ldots, x'_j$ such that their corresponding t denoted by $t'_1, \ldots, t'_j$ are increasing. Then we have $H^j(x'_1, \ldots, x'_j) = H(x_1, \ldots, x_p)$, with $H^j(x'_1, \ldots, x'_j) = C(G_{t'_1}(x'_1), \ldots, G_{t'_j}(x'_j))$, which is a j-JDFD denoted by $H(x'_1, \ldots, x'_j)$ when no doubt exists. This comes from the fact that the set of distributions included in the distribution base which are lower than $x'_1, \ldots, x'_j$ (in $(t'_1, \ldots, t'_j)$) are the same as the ones which are also lower than $x_1, \ldots, x_p$ (in $(t_1, \ldots, t_p)$). We can now apply Lemma 2 by denoting $A_j = \{F \in \mathfrak{F} \mid F(t'_j) = 0\}$ and $B_j = \{F \in \mathfrak{F} \mid F(t_i) \leqslant x_i, 1 \leqslant i \leqslant j\}$. As (empirical) $G_{t'_j}(x'_j) = |\{F \in \mathfrak{F} | F(t'_j) = 0\}|/|\mathfrak{F}|$ and $H(x'_1, \ldots, x'_j) = |\{F \in \mathfrak{F} | F(t'_i) \leqslant x'_i, 1 \leqslant i \leqslant j\}|/|\mathfrak{F}|$, it follows that $G_{t'_j}(x'_j) = |A_j|/|\mathfrak{F}|$, and $H(x'_1, \ldots, x'_j) = |B_j|/|\mathfrak{F}|$. Since from Lemma 2 we have $A_j = B_j$, it follows that $G_{t'_j}(x'_j) = H(x'_1, \ldots, x'_j)$, and so $G_{t'_j}(x'_j) = H(x_1, \ldots, x_p)$. From Lemma 2,

we have also $|A_j| = Min_{i=1,\ldots,j}\,|A_i|$, which implies $G_{t'_j}(x'_j) = Min_{i=1,\ldots,j}\,G_{t'_i}(x'_i)$. Since $Min_{i=1,\ldots,j}\,G_{t'_i}(x'_i) = Min_{i=1,\ldots,p}\,G_{t_i}(x_i)$ (since for i such that $x_i = 1$, we have $G_{t_i}(x_i) = 1$ and for i such that $x_i < 1$ we have $G_{t_i}(x_i) \leqslant 1$), we obtain $H(x_1, \ldots, x_p) = Min_{i=1,\ldots,p}\,G_{t_i}(x_i)$ which shows that $H(x_1, \ldots, x_p) = C(G_{t_1}(x_1), \ldots, G_{t_p}(x_p))$, where C is the *Min* copula.

(ii) If $x_p < 1$, then $Min(G_{t_1}(x_1), \ldots, G_{t_p}(x_p)) = G_{t_p}(x_p)$.

As in the proof of (i), we denote $x'_1, \ldots, x'_j$ (associated with increasing $t'_1, \ldots, t'_j$) to be the x_i's among $x_1, \ldots, x_p$ which are strictly lower than 1. It follows that $x'_j = x_p$ and so from Lemma 2 that $Min(G_{t'_1}(x'_1), \ldots, G_{t'_j}(x'_j)) = G_{t'_j}(x'_j) = G_{t_p}(x_p)$. We have $Min(G_{t_1}(x_1), \ldots, G_{t_p}(x_p)) = Min(G_{t'_1}(x'_1), \ldots, G_{t'_j}(x'_j))$ as shown in the preceding proof. Therefore we have finally: $Min(G_{t_1}(x_1), \ldots, G_{t_p}(x_p)) = G_{t_p}(x_p)$.

(iii) If $x < 1$, then $G_t(x) = Pr(Z > t) = 1 - F_z(t)$.

By definition $F_z(t) = Pr(\{Z(w) \leqslant t\})$ and the empirical function $G_t(x) = Pr(\{F_i \in \mathfrak{F} | F_i(t) \leqslant x\})$ is exactly the proportion of unit mass distributions F_i with value 1 strictly after t (i.e., $t' > t$), as $F_i(t) = 0$ if $t < Z(w_i)$ and $F_i(t) = 1$ if $t < Z(w_i)$. In other words, this means that $G_t(x)$ is the proportion of individuals w such that $Z(w) > t$, and then $G_t(x) = Pr(Z > t) = 1 - F_z(t)$.

(iv) If $x_p < 1$, then $F_z(t_p) = 1 - H(x_1, \ldots, x_p)$.

Indeed, from (i), we have $H(x_1, \ldots, x_p) = Min(G_{t_1}(x_1), \ldots, G_{t_p}(x_p))$; from (ii), we have $Min(G_{t_1}(x_1), \ldots, G_{t_p}(x_p)) = G_{t_p}(x_p)$; and from (iii), we have $F_z(t_p) = 1 - G_{t_p}(x_p)$.

Hence, the proposition is proved. □

5.2. The standard mixture decomposition is a special case

Here we need to introduce the following notation: $(P_1, \ldots, P_\ell)$ is a partition into ℓ classes of the set $\{w_1, \ldots, w_N\}$ described by the distribution base $\mathfrak{F}$, F_{z_i} is the distribution associated with a quantitative random variable Z^i defined on Ω, $\mathfrak{F}^i$ $(i = 1, \ldots, \ell)$ is a distribution base whose elements are the units mass distributions associated with each value $Z^i(w)$ with $w \in P_i$ (i.e., they take the value 0 for $t \leqslant Z^i(w)$ and 1 for $t > Z^i(w)$), G^i_t is a distribution function of distributions at value t associated with the distribution base $\mathfrak{F}^i$ and $H^i_{t_1,\ldots,t_k}$ is a k-joint distribution function of distributions associated with the same distribution base.

Proposition 5. *If* $H_{t_1,\ldots,t_k} = \sum_{i=1}^{\ell} p_i H^i_{t_1,\ldots,t_k}$ *with* $\sum_{i=1}^{\ell} p_i = 1$, *then* $F_z = \sum_{i=1}^{\ell} p_i F_{z_i}$.

Proof. From Proposition 2, we have $H^i_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = C^i(G^i_{t_1}(x_1), \ldots, G^i_{t_k}(x_k))$, where C^i is a k-copula. Therefore,

$$H_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = \sum_{i=1}^{\ell} p_i C^i(G^i_{t_1}(x_1), \ldots, G^i_{t_k}(x_k)).$$

We choose $x_p < 1$ and use Proposition 4.

From (i), we obtain: $H_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = \sum_{i=1}^{\ell} p_i Min(G^i_{t_1}(x_1), \ldots, G^i_{t_k}(x_k))$.

From (ii), we obtain: $H_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = \sum_{i=1}^{\ell} p_i G^i_{t_k}(x_k)$.

From (iii), we obtain: $H_{t_1,\dots,t_k}(x_1,\dots,x_k) = \sum_{i=1}^{\ell} p_i(1 - F_{z_i}(t_k)) = 1 - \sum_{i=1}^{\ell} p_i F_{z_i}(t_k)$. From (iv), we obtain $F_z(t_k) = 1 - H_{t_1,\dots,t_k}(x_1,\dots,x_k)$ and therefore $F_z(t_k) = \sum_{i=1}^{\ell} p_i F_{z_i}(t_k)$.

Since the same reasoning can be made for any sequence $t_1,\dots,t_k$, it follows finally $F_z = \sum_{i=1}^{\ell} p_i F_{z_i}$. □

5.3. Links between the generalised mixture decomposition problem and the standard one

It follows from Proposition 5 that, by solving the mixture decomposition of distribution of distributions problem, we have solved the standard mixture decomposition problem. This follows from the fact that it is possible to induce $F_{z_i}(t_1),\dots,F_{z_i}(t_k)$, from $G^i_{t_1}(x_1),\dots,G^i_{t_k}(x_k)$ and therefore, the parameters of the chosen model of the density law associated with each Z^i. Moreover, by choosing the "best model" among a given family of possible models (Gaussian, Gamma, Poisson, etc.) for each Z^i, we can obtain a different model for each law of the mixture. By "best model", we mean the model which best fits the $F_{z_i}(t_1),\dots,F_{z_i}(t_k)$ for each i. It would be interesting to compare the results of both approaches: the mixture decomposition of distributions of distributions algorithms, and the standard mixture distribution algorithms in the standard framework. This comparison could be done when the same model is used for each class, or more generally when each law of the mixture follows a different family model.

6. Mixture decomposition with copula model in the case of more than one variable

We have considered the mixture decomposition problem for the case of a single variable. In order to extend our methodology to the case of several variables, we can use multidimensional copulas, for instance the "generalised" parametric family of copulas proposed by Vrac [14]. Such copulas are complicated to write explicitly, their parameters are complicated to estimate and their meaning is not easy to interpret. Therefore, we propose two methods for several variables.

6.1. Binary tree method

This method proceeds as follows: we look for the variable which gives the best mixture decomposition criterion value in two classes and we repeat the process for each resultant class until the size of the classes becomes small enough (as adjudged by the appropriate criteria). In order to select the best variable, the choice of the t_j is important. Since we are looking for a partition of the set of distributions, it is clear that a given t_j is not good if all the distributions F_i of the base $\mathfrak{F}$ take the same value at that t_j value. Also, a particular t_i is a poor choice if all the $F_i(t_j)$ are uniformly distributed in $[0, 1]$. In fact we can say that a t_j is good if distinct clusters of values exist among the set of values: $\{F_i(t_j) \mid i = 1,\dots,N\}$. For instance, Jain and Dubes [9] proposed several methods in order to reveal any clustering tendency. Here, we are dealing with the special case where we look for such a tendency among a set of points in the interval $[0, 1]$.

We suggest a method based on the number of triangles whose vertices are points of [0, 1] and we take the two sides that are nearest in length and larger (respectively smaller) than the remaining third side. These sets of triangles are denoted A (respectively B). For instance, let $(a_1, a_2, a_3) \in [0, 1]^3$ be the vertices of a triangle a. The lengths of the sides of this triangle are: $|a_1 - a_2|$, $|a_1 - a_3|$, $|a_2 - a_3|$. If the two closest are larger than the third one, then $a \in A$, if not $a \in B$. Let X^0 be a random variable which associates $u = \{u_1, \ldots, u_N\}$ for N points randomly distributed in the interval [0, 1], with the value $X^0(u) = (|A| - |B|)/C_N^3 = 6(|A| - |B|)/(n(n-1)(n-2))$ which belongs to $[-1, 1]$. By the distribution of this random variable X^0, we define the hypothesis H^0 that there is no clustering tendency. The greater $X^0(u)$ is, the higher is the clustering tendency of the N points. We calculate the number of triangles whose vertices are points of $U = \{F_i(t_j) \mid i = 1, \ldots, N\}$ for which the two closest sides are larger (respectively smaller) than the remaining side. We denote this number by $A(U)$ (respectively $B(U)$). Given the distribution of X^0, the value $(A(U) - B(U))/C_N^3 = 6(A(U) - B(U))/(n(n-1)(n-2))$ can reject or accept the null hypothesis at a given threshold.

Proposition 6. *The expectation of the random variable X^0 is $\frac{1}{3}$ and its variance is $8/9C_N^3$.*

Proof. The proof consists of realizing that we are considering only flat triangles. A flat triangle belongs to A or B according to the position of the medium point. A triangle has a probability $\frac{2}{3}$ to be in A and $\frac{1}{3}$ in B. We define a random variable T with Bernoulli distribution $B(p)$ with $p=\frac{2}{3}$: set $T(tr)=1$ with probability p (if the triangle tr belongs to A) and $T(tr)=0$ with a probability $1 - p$ (if the triangle tr belongs to B). With N points, we have C_N^3 (with $C_n^p = n!/(n - p)!p!$) and then the random variable $|A| = \sum_{i=1}^{C_N^3} T(tr_i)$ has a binomial distribution $B(C_N^3, p)$ with expectation $\mathbb{E}(|A|) = C_N^3 p$ and variance $\sigma^2(|A|) = C_N^3 p(1-p)$. Then,

$$\mathbb{E}(X^0) = \mathbb{E}\left(\frac{|A| - |B|}{C_N^3}\right) = \mathbb{E}\left(\frac{2|A|}{C_N^3}\right) - 1 = \frac{1}{3} \tag{7}$$

and

$$\sigma^2(X^0) = \sigma^2\left(\frac{|A| - |B|}{C_N^3}\right) = \sigma^2\left(\frac{2|A|}{C_N^3}\right) = \frac{4}{(C_N^3)^2} C_N^3 p(1-p) = \frac{8}{9C_N^3}. \tag{8}$$

When t_1 and t_2 have been found, the mapping a defined in Section 3 can be extended in the following way: $a^* : \Omega \times [0, 1]^2 \longrightarrow \mathbb{R}^+$:

$$a^*(w, \varepsilon) = \int_{t_1}^{t_2} [Y(w)R(\varepsilon)C(G_{t_1}, G_t)]\,\mathrm{d}t \in \mathbb{R}^+.$$

6.2. Coupling method

This method proposed by Vrac [14] and applied by Vrac et al. [15], proceeds as follows. We consider a mixture decomposition on a first variable Y_1 with two given values t_1^1 and t_2^1,

and another mixture decomposition on a second variable Y_2 with two given values t_1^2 and t_2^2. For each $w \in \Omega$, and for each $F_w = (F_w^{Y_1}, F_w^{Y_2}) \in \mathfrak{F}$, we can write:

$$H^{Y_1}_{t_1^1,t_2^1}(x^1) = \sum_{i=1}^{\ell} p_i^{Y_1} C^i_{Y_1}(G^i_{t_1^1}(x_1^1), G^i_{t_2^1}(x_2^1))$$

and

$$H^{Y_2}_{t_1^2,t_2^2}(x^2) = \sum_{i=1}^{\ell} p_i^{Y_2} C^i_{Y_2}(G^i_{t_1^2}(x_1^2), G^i_{t_2^2}(x_2^2)),$$

where

- $H^{Y_j}_{t_1^j,t_2^j}$ is a 2-JDFD at point (t_1^j, t_2^j) from variable Y_j.
- $x^j = (x_1^j, x_2^j) = (F_w^{Y_j}(t_1^j), F_w^{Y_j}(t_2^j))$.
- $p_i^{Y_j}$ is the ith mixture ratio from variable Y_j.
- $C^i_{Y_j}$ is the copula of the component i from variable Y_j.

From each unit w_i and each distribution $F_i \in \mathfrak{F}$, we obtain a new pair $(H^{Y_1}_{t_1^1,t_2^1}(x^1), H^{Y_2}_{t_1^2,t_2^2}(x^2))$ of values of the distributions. Then from the N units of the sample, and the N distributions in $\mathfrak{F}$, we obtain N pairs. A mixture decomposition of distributions by copulas can be realised on this new database. This method has been applied to climatological data and has given very encouraging results (see Vrac [14] and Vrac et al. [15]). Some comparisons have been done on this kind of data: for example, the comparison between the results of our mixture decomposition by copulas and the results of the EM algorithm on probabilistic data and standard numerical data. Moreover, the extensions to copulas of other algorithms are on the way.

7. Conclusion

Many things remain to be done, for instance, studying the case in which each class may be modeled by a different copula family, or comparing the results obtained by the general methods and the standard methods of mixture decomposition on standard data (as they are a special case of distributions). Indeed, the proposed mixture decomposition method can deal with standard data and not just on probabilistic data. Also, the copulas can be modeled by "generalised Archimedian copulas" and the G_t can be modeled at each t by a different distribution family and even a mixture decomposition of distributions. We can also add other criteria taking into account a class variable and a learning set. Notice that the same kind of approach can be used in the case where, instead of having distributions, we have any kind of mapping. But in that case, the interpretation will be less rich than when the mappings are distributions.

References

[1] H.H. Bock, E. Diday (Eds.), Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, Springer, Heidelberg, 2000.
[2] D.G. Clayton, A model for association in bivariate life tables, Biometrika 65 (1978) 141–151.
[3] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. Roy. Statist. Soc. 39 (1977) 1–38.
[4] E. Diday, Extracting information from multivalued surveys or from very extensive data sets by symbolic data analysis, in: A. Ferligoj (Ed.), Advances in Methodology, Data Analysis and Statistics, Metodoloski zveski, vol. 14, Ljubljana, FDV, 1998.
[5] E. Diday, A generalisation of the mixture decomposition problem in the symbolic data analysis framework, CEREMADE Report, vol. 112, May 2001, pp. 1–14.
[6] E. Diday, A. Schroeder, Y. Ok, The dynamic clusters method in pattern recognition, in: Proceeding of IFIP Congress: Information Processing, Elsevier, Amsterdam, North-Holland, J.L. Rosenfeld Publisher, Stockholm, 1974, pp. 691–697.
[7] M.J. Frank, On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$, Aequationes Math. 19 (1979) 53–77.
[8] C. Genest, Frank's family of bivariate distributions, Biometrika 74 (1987) 549–555.
[9] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Advanced Reference Series, vol. 07632, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[10] M. Janowitz, B. Schweizer, Ordinal and percentile clustering, Math. Social Sci. (18) (1989) 135–186.
[11] R.B. Nelsen, An Introduction to Copulas, in: Lecture Notes in Statistics, Springer, New York, 1998.
[12] B. Schweizer, A. Sklar, Probabilistic Metric Spaces, Elsevier, North-Holland, New York, 1983.
[13] A. Sklar, Fonction de répartition à n dimensions et leurs marges, Inst. Statist. Univ. Paris Publ. 8 (1959) 229–231.
[14] M. Vrac, Analyse et modélisation de données probabilistes par Décomposition de Mélange de Copules et Application à une base de données climatologiques, Thèse de doctorat, Université Paris IX Dauphine, 2002.
[15] M. Vrac, E. Diday, A. Chédin, Décomposition de mélange de distributions et application à des données climatiques, Rev. Statist. Appl. LII (1) (2004) 67–96.