Comput Stat
DOI 10.1007/s00180-011-0266-0

ORIGINAL PAPER

Copula analysis of mixture models

M. Vrac · L. Billard · E. Diday · A. Chédin

Received: 10 February 2010 / Accepted: 8 June 2011

Abstract Contemporary computers collect databases that can be too large for classical methods to handle. The present work takes data whose observations are distribution functions (rather than the single numerical point value of classical data) and presents a computational statistical approach of a new methodology to group the distributions into classes. The clustering method links the searched partition to the decomposition of mixture densities, through the notions of a function of distributions and of multi-dimensional copulas. The new clustering technique is illustrated by ascertaining distinct temperature and humidity regions for a global climate dataset and shows that the results compare favorably with those obtained from the standard EM algorithm method.

Keywords Classification of distributions · Copulas · Dynamical clustering · Data distributions · Estimation · Mixture model

M. Vrac
Laboratoire de Sciences du Climat et de l'Environment, IPSL-CNRS/CEA/UVSQ, Centre d'Etudes de Saclay, Orme des Merisiers, 91191 Gif-sur-Yvette, France

L. Billard (✉)
Department of Statistics, University of Georgia, Athens, GA 30602, USA
e-mail: lynne@stat.uga.edu

E. Diday
CEREMADE, University of Paris Dauphine, Place du Maréchal de Lattre-de-Tassigny, 75775 Paris, France

A. Chédin
Laboratoire de Meteorologie, Dynamique/IPSL, Ecole Polytechnique, 91128 Palaiseau, France

Published online: 28 June 2011

1 Introduction

Contemporary computers with increasing frequency make possible the collection of massive datasets whose size (e.g., number of observations and number of variables) can be too large for those same computers to analyse. Thus, some form of data aggregation must first occur in order to reduce the dataset to a more manageable size in order for appropriate analyses to proceed. The nature of the aggregation used will depend on the scientific question(s) being asked. For example, the meteorological data (considered in Sect. 5) arose from aggregating a dataset that contained millions of values for each variable (such as temperature, humidity, etc.) clearly too large to analyse by standard methods. In our case, frequency distributions were generated by aggregating values from the same latitude $\times$ longitude grid point.

This work focuses on data for which each observation is a distribution function. The distribution function can be the original observation per se; or, as is illustrated in the world climatology example, it may result from aggregation of (r, say, classical) data points over some suitable domain.

The goal is to develop a methodology for grouping a sample of N ($N = 16200$ in the example of Sect. 5) distributions in the p-dimensional Cartesian product of distributions space, into a finite number K of classes. Let us assume there is an underlying probability density function $f_k(\cdot)$ for each class, $k = 1, \ldots, K$. Then we can write the mixture density

$$f(x_1, \ldots, x_p; \boldsymbol{\alpha}) = \sum_{k=1}^{K} p_k f_k(x_1, \ldots, x_p; \boldsymbol{\alpha}_k) \tag{1}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K, p_1, \ldots, p_K)$ is the parameter with values in $\mathbb{R}^d$ associated with $f(\cdot)$, $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kd_k})$ is the parameter with values in $\mathbb{R}^{d_k}$ associated with $f_k(\cdot; \boldsymbol{\alpha}_k)$, and p_k is the a priori probability that an element from the sample has the density $f_k(\cdot; \boldsymbol{\alpha}_k)$ with $0 < p_k < 1$, $\sum_{k=1}^{K} p_k = 1$, for all $k = 1, \ldots, K$.

For classical data, (1) represents the mixture based on a sample of observations $\boldsymbol{x} = (x_1, \ldots, x_p)$ in $\mathbb{R}^p$. Parametric mixture models for classical data are reviewed in, e.g., Fraley and Raftery (2002). In this setting, this problem of mixture decomposition has been addressed by many authors adopting either of two different approaches. The most widespread approach consists of treating the decomposition problem as an estimation problem, targeted at estimating the parameters $(p_k, \boldsymbol{\alpha}_k, k = 1, \ldots, K)$, usually using maximum likelihood estimation techniques. In general, optimization algorithms are based on the EM algorithm of Dempster et al. (1977). Variations of the EM algorithm and/or adaptations to special situations include the stochastic EM (SEM) algorithm (e.g., Celeux and Diebolt 1986; Meng and Rubin 1991), the classification EM (CEM) algorithm (e.g., Celeux and Govaert 1992), the Monte Carlo EM (MCEM) algorithm (e.g., Tanner and Wong 1987; Wei and Tanner 1990) and those developed by Redner and Walker (1984), with more details in McLachlan and Peel (2000).

Another approach builds on clustering ideas within the framework of classification methodology. These methods consider a set of N observations to be grouped into

K classes $(P_1, \ldots, P_K) = P$ where each class P_k is assimilated to a sample with probability law $f_k(\cdot; \boldsymbol{\alpha}_k)$; see, e.g., the dynamical clustering algorithms of Diday et al. (1974), Schroeder (1976), Scott and Symons (1971) and Symons (1981). These methods were combined with EM concepts to produce a classification EM algorithm by Celeux and Govaert (1992, 1993). Celeux et al. (1989) consider dynamical clustering on mixture distributions. Other classical clustering approaches include iterative relocation algorithms (e.g., Hartigan and Wong 1979; Diday et al. 1974), hierarchical classification (e.g., Brossier 1990), neural networks (e.g., Bishop 1995; Bock 1998), overlapping classification such as additive clustering (e.g., Arabie and Carroll 1980), pyramids (e.g., Diday 1984), and the functional clustering model (e.g., Winsberg and DeSoete 1999; James and Sugar 2003), among others. An excellent review of most of these algorithms can be found in Gordon (1999).

Our purpose is to present details of a new dynamical clustering method for mixture distributions in the context of data analysis where the observed distribution function replaces the single point numerical value of classical data. Further, ideas behind the concept of copulas (see Nelsen 1999) are introduced as part of the methodology. Copulas provide a means of describing dependence relations between a joint distribution function and the corresponding marginal distributions. An important family of copulas is the Archimedean family. The methodology developed leads to estimation questions within copula theory. Genest and MacKay (1986) describe the relationship between 2-dimensional Archimedean copulas and Kendall's tau. Genest and Rivest (1993) considered inference questions for a Frank family copula for classical data through Kendall's tau relationship. Our methodology includes the possibility of using Kendall's tau and also Spearman's rho relationship with copulas. While our approach is new, it could be viewed as a form of hierarchical modeling (using cumulative distribution functions instead of density functions) and with cumulative functions as the functions of functional data analysis.

Some useful formula and definitions relating to functions of distributions along with some basic results in copula theory are presented in Sect. 2. The algorithm of the suggested dynamical clustering method is described in Sect. 4 with the associated estimation issues addressed in Sect. 3. The theory is applied to a bivariate (temperature and humidity) climatological data set in Sect. 5.1, and compared with results obtained from the EM algorithm method in Sect. 5.2.

2 Mixture decomposition for probability distributions

We start with a description of the data and output sought, in Sect. 2.1; this includes the concept of a (joint) distribution function of distribution values. Our approach is to model the data as a mixture of distributions utilizing the concept of copulas; see Sect. 2.2. An important class of copulas, the Archimedean family, is presented briefly in Sect. 2.3.

2.1 Input and output

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ be a p-dimensional random vector taking values in $\mathcal{R}^p$; and let F^j be the distribution function associated with Y_j, $j = 1, \ldots, p$. Here, and throughout

this work, a distribution function, or simply distribution, is taken to be a cumulative distribution function (cdf). Then, we have a sample $\mathfrak{F} = (F_1, \ldots, F_N)$ of N p-dimensional distributions where $F_u = (F_u^1, \ldots, F_u^p)$, $u = 1, \ldots, N$, are realizations of a random vector with F_u^j being the realization of the distribution F^j for observation u, $u = 1, \ldots, N$. While each F_u may be a well-defined known distribution, more typically it will be an empirical distribution $\tilde{F}_u^{(r)}$ estimated in part or entirely from r (say) observations. For example, $\tilde{F}_u^{(r)}$ may be known to follow a normal distribution but its parameters are estimated from the data. In our climatology application (Sect. 5), the $\tilde{F}_u^{(r)}$ are estimated as kernel density functions. Except where necessary to distinguish these cases (such as in "Appendix A.2"), we denote $\tilde{F}_u^{(r)} \equiv F_u$. Each F_u belongs to $\Omega_F = \Omega_F^1 \times \cdots \times \Omega_F^p$, with Ω_F^j being the set of possible distributions to describe the individuals from Ω_F for the jth variable and "$\times$" is the product of spaces operator.

Our aim is to find a partition of this sample of N distributions into K classes; and thence to obtain estimates of the underlying distribution corresponding to the outcome classes, and the proportions of the observations of $\mathfrak{F}$ in each class.

We need the concept of "distribution function of distribution values" and "joint distribution function of distribution values". For clarity of notational presentation, the methodology is described for $p = 1$. In this case, $F_u = F_u^1$ is the distribution function of the observation unit u for this variable, and $\Omega_F = \Omega_F^1$. Key formulae for the general p case are presented in (5) and (12).

Let $\mathfrak{F} = (F_1, \ldots, F_N)$ be a sample of N distributions from the population Ω_F. A *distribution function of distribution values* at the point Z is the function defined by $G_Z : [0, 1] \longrightarrow [0, 1], x \mapsto G_Z(x)$ with

$$G_Z(x) = \mathbb{P}(F(Z) \leq x), \quad \text{for all } x \in \mathbb{R}. \tag{2}$$

In (2), $F(Z)$ is a distribution function, and the domain of Z corresponds to the domain of F. In the climatology application of Sect. 5, the Z refers to values of temperature (and/or humidity).

If the function $G_Z(x)$ is empirically modeled from $\mathfrak{F}$, the distribution function is

$$\begin{aligned} G_Z^e(x) &= \mathbb{P}(F_u \in \mathfrak{F}; F_u(Z) \leq x, u = 1, \ldots, N) \\ &= \frac{\text{card}(F_u \in \mathfrak{F}; F_u(Z) \leq x, u = 1, \ldots, N)}{\text{card}(\mathfrak{F})}. \end{aligned} \tag{3}$$

For instance, Fig. 1 shows $N = 5$ distributions $\{F_u, u = 1, \ldots, 5\}$. Suppose we want to calculate the empirical distribution $G_Z^e(x) \equiv G_Z(x)$. If $x = 0.4$, $G_{Z_i}(x)$ is the percentage of distributions taking a value smaller than or equal to 0.4 at the point Z_i. In this example, $G_{Z_1}(0.4) = 3/5$ and $G_{Z_2}(0.4) = 1/5$.

A *joint distribution function of distribution values* at the point $\mathbf{Z} = (Z_1, \ldots, Z_n)$ is the function defined by $H_{\mathbf{Z}} : [0, 1]^n \longrightarrow [0, 1], \mathbf{x} = (x_1, \ldots, x_n) \mapsto H_{\mathbf{Z}}(\mathbf{x})$ with

$$H_{\mathbf{Z}}(x_1, \ldots, x_n) = \mathbb{P}\left(F_u \in \mathfrak{F}; F_u(Z_1) \leq x_1, \ldots, F_u(Z_n) \leq x_n, u = 1, \ldots, N\right). \tag{4}$$

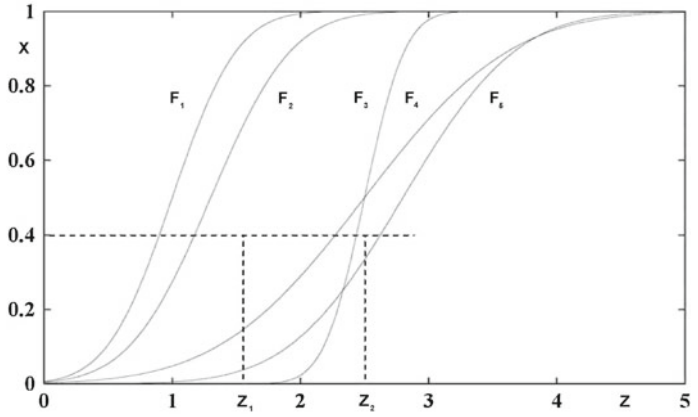

Fig. 1 Data: Observed frequency distributions $F_u, u = 1, \ldots, 5$. *Dotted lines* illustrate calculation of $G_{Z_1}(x)$ and $G_{Z_2}(x)$ from (3)

Notice that the function $G_{Z_i}(x_i), i = 1, \ldots, n$, is just a distribution function of the random variable $F(Z_i)$ which takes values in [0,1]; and $H_{Z_1,\ldots,Z_n}(x_1, \ldots, x_n)$ is an n-dimensional joint distribution function of the random vector $(F(Z_1), \ldots, F(Z_n))$, which takes values in [0, 1] with marginal distributions $G_{Z_i}(x_i), i = 1, \ldots, n$. Therefore, well known properties of univariate and multivariate distribution functions pertain for $G(\cdot)$ and $H(\cdot)$, respectively. For example, for each Z_i, x_i, (i) $G_{Z_i}(x_i)$ is a non-decreasing function of x_i, (ii) $\lim_{x_i \to -\infty} G_{Z_i}(x_i) = 0$, (iii) $\lim_{x_i \to +\infty} G_{Z_i}(x_i) = 1$, (iv) $G_{Z_i}(x_i)$ is continuous from the right; likewise for $H_{\boldsymbol{Z}}(\boldsymbol{x})$. Proofs of (i)–(iii) are found in Diday and Vrac (2005) and of (iv) in Vrac (2002).

The functions in (2) and (4) readily generalize when $p > 1$. For example, (4) becomes, for $\boldsymbol{Z} = ((Z_1^1, \ldots, Z_{n_1}^1), \ldots, (Z_1^p, \ldots, Z_{n_p}^p))$, $H_{\boldsymbol{Z}} : [0,1]^n \longrightarrow [0,1]$ where $n = \sum_{j=1}^p n_j, \boldsymbol{x} = ((x_1^1, \ldots, x_{n_1}^1), \ldots, (x_1^p, \ldots, x_{n_p}^p)) \mapsto H_{\boldsymbol{Z}}(\boldsymbol{x})$, with

$$H_{\boldsymbol{Z}}(\boldsymbol{x}) = \mathbb{P}\left(F_u \in \mathfrak{F};\, F_u^1(Z_1^1) \leq x_1^1, \ldots, F_u^p(Z_{n_p}^p) \leq x_{n_p}^p, u = 1, \ldots, N\right). \tag{5}$$

We note that in our application in which the data F_u are cumulative distributions, it follows that for a given variable the $Z_1, \ldots, Z_n$ and the $F_u(Z_1), \ldots, F_u(Z_n)$ have the same order. However, this is not the case for all applications. For example, our methodology can be applied to functional data, not necessarily cumulative distributions, where now the $F_u(Z_i)$'s would not necessarily be ordered even if the Z_i's were. For some applications, it may be necessary to characterize the dependencies between the $F_u(Z_i)$'s in a specific but non-ordered way.

2.2 Modeling dependent distributions with copulas

Schweizer and Sklar (1983) show how copulas link multidimensional joint distribution functions to the one dimensional marginal distributions of the associated

random variables. We give first the definition of a copula and the important Sklar's Theorem which underpins basic copula theory.

From Nelsen (1999), a function $C \equiv C(\boldsymbol{v}), \boldsymbol{v} = (v_1, \ldots, v_n)$ is defined as an n-dimensional *copula* (or n-copula) C from $[0,1]^n \to [0,1]$ if: (i) For all $\boldsymbol{v}$ in $[0,1]^n$, if at least one coordinate of $\boldsymbol{v}$ is 0, $C(\boldsymbol{v}) = 0$, and if all coordinates of $\boldsymbol{v}$ are 1 except v_m, then $C(\boldsymbol{v}) = v_m$; (ii) For all $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ in $[0,1]^n$ such that $\boldsymbol{a} \leq \boldsymbol{b}$, then $V_C([\boldsymbol{a}, \boldsymbol{b}]) \geq 0$, with $V_C([\boldsymbol{a}, \boldsymbol{b}]) = \Delta_a^b C(v) = \Delta_{a_n}^{b_n} \Delta_{a_{n-1}}^{b_{n-1}} \ldots \Delta_{a_1}^{b_1} C(v)$ where the first order difference of C for the mth component of C is $\Delta_{a_m}^{b_m} C(v) = C(v_1, \ldots, v_{m-1}, b_m, v_{m+1}, \ldots, v_n) - C(v_1, \ldots, v_{m-1}, a_m, v_{m+1}, \ldots, v_n)$.

Let H be an n-dimensional distribution function with unidimensional marginal distribution functions $F_1, \ldots, F_n$. Then, from Sklar (1959) Theorem, there exists a copula C such that, for all $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$,

$$H(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{6}$$

If $F_1, \ldots, F_n$ are continuous, then C is unique; otherwise, C is uniquely determined on $RanF_1 \times \ldots \times RanF_n$, where $RanF_u = [0,1]$ is the range of F_u. Conversely, if $F_1, \ldots, F_n$ are distribution functions and C is a copula, the function H defined by (6) is an n-dimensional distribution function with marginal distribution functions $F_1, \ldots, F_n$.

Note that the functions H, $F_1, \ldots, F_n$, and C in Sklar's Theorem can be parametric or non-parametric functions. The modeling of dependencies between marginal distribution functions from our sample $\mathfrak{F}$ can be obtained by extending Sklar's theorem. Let $G_{Z_1}, \ldots, G_{Z_n}$ denote the distribution functions at the points $Z_1, \ldots, Z_n$, and let $H_{Z_1,\ldots,Z_n}$ be the joint distribution function of these distributions. Then, there exists an n-copula C such that, for all $(x_1, \ldots, x_n)$ belonging to $\mathbb{R}^n$,

$$H_{Z_1,\ldots,Z_n}(x_1, \ldots, x_n) = C(G_{Z_1}(x_1), \ldots, G_{Z_n}(x_n)). \tag{7}$$

Moreover, C is uniquely determined on $RanG_{Z_1} \times \cdots \times RanG_{Z_n}$ for continuous $G_{Z_i}, i = 1, \ldots, n$.

From (7), we see that the copula C is a way to model the dependencies between the $(G_{Z_1}, \ldots, G_{Z_n})$. Thus, e.g., in the climatology example in Sect. 5, the G_{Z_i}'s correspond to different temperatures and/or humidities. If there is no dependence between the G_{Z_i}'s, the product copula Π emerges (where a copula $C \equiv C(v_1, \ldots, v_n)$ is a product copula if $C = \prod_{i=1}^n v_i$; see Nelsen 1999). When $p > 1$, the same notions apply with dependencies between variables j_1 and j_2 (say) modelled by the sets $(Z_1^{j_1}, \ldots, Z_{n_{j_1}}^{j_1})$ and $(Z_1^{j_2}, \ldots, Z_{n_{j_2}}^{j_2})$; see (5).

Analogously with (1), we can write $H_{Z_1,\ldots,Z_n}$ as a mixture of parametric distributions,

$$H(x_1, \ldots, x_n; \boldsymbol{\gamma}) = \sum_{k=1}^{K} p_k H_k(x_1, \ldots, x_n; \boldsymbol{\gamma}_k) \tag{8}$$

with, for all $k = 1, \ldots, K, 0 < p_k < 1, \sum_{k=1}^{K} p_k = 1$, where $H_k(\cdot; \boldsymbol{\gamma}_k)$ is the parametric distribution for the mixture component (class) k with parameter $\boldsymbol{\gamma}_k$ belonging to $\mathbb{R}^{d_k}$ (where d_k is the dimension of the parameter $\boldsymbol{\gamma}_k$) and p_k is the a priori probability that the vector $(x_1, \ldots, x_n)$ is in the kth class. The function H_k is the joint distribution at the point $\mathbf{Z} = (Z_1, \ldots, Z_n)$ for the kth component, with marginal distributions $G^k_{Z_1}, \ldots, G^k_{Z_n}$. Therefore, from (8) and Sklar's Theorem (7), there exist copulas $C_k, k = 1, \ldots, K$, such that

$$H(x_1, \ldots, x_n; \boldsymbol{\gamma}) = \sum_{k=1}^{K} p_k C_k\big(G^k_{Z_1}(x_1; \boldsymbol{b}^k_1), \ldots, G^k_{Z_n}(x_n; \boldsymbol{b}^k_n); \boldsymbol{\beta}_k\big). \tag{9}$$

where $\boldsymbol{\gamma} = \{\boldsymbol{b}^k_i, i = 1, \ldots, n; \boldsymbol{\beta}_k, p_k, k = 1, \ldots, K\}$, $\boldsymbol{\beta}_k$ is the parameter of the copula corresponding to the kth class, and $G^k_{Z_i}(\cdot; \boldsymbol{b}^k_i)$ is the distribution function with parameter $\boldsymbol{b}^k_i$ at the point Z_i in the class k. In this formulation, the parameters $\boldsymbol{\gamma}_k$ in (8) become the parameters $\{\boldsymbol{b}^k_i, i = 1, \ldots, n, \boldsymbol{\beta}_k\}$ of (9). Note that while G and C are written in (9) as parametric functions, they can be non-parametric functions. We can easily prove the following results by applying the chain rule to (8) and (9).

Let $h_k(\cdot) \equiv h_k(x_1, \ldots, x_n; \boldsymbol{\gamma}_k) = \partial^n H_k / \partial x_1 \ldots \partial x_n$ denote the probability density function associated with the distribution function $H_k(\cdot)$. Then, $h_k(\cdot)$ can be written as

$$\begin{aligned} h_k(x_1, \ldots, x_n; \boldsymbol{\gamma}_k) = &\left\{ \prod_{i=1}^{n} \frac{dG^k_{Z_i}}{dx_i}(x_i; \boldsymbol{b}^k_i) \right\} \frac{\partial^n}{\partial x_1 \ldots \partial x_n} \\ &\times C_k\big(G^k_{Z_1}(x_1; \boldsymbol{b}^k_1), \ldots, G^k_{Z_n}(x_n; \boldsymbol{b}^k_n); \boldsymbol{\beta}_k\big). \end{aligned} \tag{10}$$

Hence, substituting from (10) into (9), we have that the probability density function $h(\cdot) \equiv h(x_1, \ldots, x_n; \boldsymbol{\gamma}) = \partial^n H / \partial x_1 \ldots \partial x_n$ associated with $H(\cdot)$ can be written as

$$\begin{aligned} h(x_1, \ldots, x_n; \boldsymbol{\gamma}) = &\sum_{k=1}^{K} p_k \left\{ \prod_{i=1}^{n} \frac{dG^k_{Z_i}}{dx_i}(x_i; \boldsymbol{b}^k_i) \right\} \frac{\partial^n}{\partial x_1 \ldots \partial x_n} \\ &\times C_k\big(G^k_{Z_1}(x_1; \boldsymbol{b}^k_1), \ldots, G^k_{Z_n}(x_n; \boldsymbol{b}^k_n); \boldsymbol{\beta}_k\big). \end{aligned} \tag{11}$$

These equations readily generalize to $p > 1$. In this case, (11) becomes

$$\begin{aligned} h\left(x^1_1, \ldots, x^p_{n_p}; \boldsymbol{\gamma}^p\right) = &\sum_{k=1}^{K} p_k \left\{ \prod_{j=1}^{p} \prod_{i=1}^{n_j} \frac{dG^k_{Z^j_i}}{dx^j_i}(x^j_i; \boldsymbol{b}^{jk}_i) \right\} \\ &\cdot \frac{\partial^n}{\partial x^1_1 \ldots \partial x^p_{n_p}} C_k\big(G^k_{Z^1_1}(x^1_1; \boldsymbol{b}^{1k}_1), \ldots, G^k_{Z^p_{n_p}}(x^p_{n_p}; \boldsymbol{b}^{pk}_{n_p}); \boldsymbol{\beta}_k\big) \end{aligned} \tag{12}$$

where $\boldsymbol{\gamma}^p = (\boldsymbol{b}^{jk}_i, \boldsymbol{\beta}_k, p_k, i = 1, \ldots, n_j, j = 1, \ldots, p, k = 1, \ldots, K)$ is the set of parameters, and where $\mathbf{Z} = ((Z^1_1, \ldots, Z^1_{n_1}), \ldots, (Z^p_1, \ldots, Z^p_{n_p}))$ with $n = \sum_{j=1}^{p} n_j$.

Note that while for a given j, the $Z_1^j, \ldots, Z_{n_j}^j$ may be ordered, it is not necessarily the case that, for $j_1 \neq j_2$, the values of $Z_1^{j_1}, \ldots, Z_{n_{j_1}}^{j_1}, Z_1^{j_2}, \ldots, Z_{n_{j_2}}^{j_2}$ are ordered.

An alternative approach to using (5) and (12) when $p > 1$ is to use Sklar's Theorem twice to obtain a copula of copulas. Thus, from (9) and (11), for each variable Y_j, the distribution $H^{Y_j}(\cdot)$ is found, $j = 1, \ldots, p$. We can then calculate the set of p-dimensional values $(H^{Y_1}, \ldots, H^{Y_p})$ for each of the N observations in Ω; this gives the set H_u, $u = 1, \ldots, N$. Then we can repeat the methodology of (9) and (11) (originally based on the nF_u's) to one based now on these N $\{H_u^{Y_j}, j = 1, \ldots, p\}$ values. For example, to calculate the $H^{Y_j}(\cdot)$, (9) becomes, for $j = 1, \ldots, p$,

$$H^{Y_j}\left(\mathbf{x}_1^j, \ldots, \mathbf{x}_{n_j}^j; \boldsymbol{\gamma}_j\right) = \sum_{k=1}^{K_j} p_k^j C_k\left(G^k_{Z_1^j}\left(\mathbf{x}_1^j; \boldsymbol{b}_1^{jk}\right), \ldots, G^k_{Z_{n_j}^j}\left(\mathbf{x}_p^j; \boldsymbol{b}_p^{jk}\right); \boldsymbol{\beta}_k^j\right),$$

where $\boldsymbol{\gamma}_j = (\boldsymbol{b}_1^{jk}, \ldots, \boldsymbol{b}_p^{jk}, \ p_k^j, \ \boldsymbol{\beta}_k^j, k = 1, \ldots, K_j)$, $j = 1, \ldots, p$, and where the $x_1^j, \ldots, x_p^j$ are the values of the cumulative distribution function of the variable Y_j estimated in $Z_1^j, \ldots, Z_p^j$. Then, when based on the $N\{H_u^{Y_j}, j = 1, \ldots, p\}$ values, (9) becomes

$$H(x_1, \ldots, x_p; \boldsymbol{\gamma}) = \sum_{k=1}^{K'} p_k' C_k\left(G'^k_{Y_1}\left(x_1; \boldsymbol{b}_1'^k\right), \ldots, G'^k_{Y_p}\left(x_p; \boldsymbol{b}_p'^k\right); \boldsymbol{\beta}_k'\right) \tag{13}$$

where now $\boldsymbol{\gamma} = (\boldsymbol{b}_j'^k, \ j = 1, \ldots, p, \ p_k', \ \boldsymbol{\beta}_k', \ k = 1, \ldots, K')$, and where $G'^k_{Y_j}$ is the distribution function of the $H^{Y_j}(\cdot)$ values of the kth component. Thus the dependencies between the variables Y_j, $j = 1, \ldots, p$, are modeled through the copula in (13). There are still n values of Z_i as in the use of (12) directly; but by using (9) and (11), n_j are used for each application of (9) and (11) by j. The dependencies within each set of Z_i^j for each j are modeled first through the copulas of (9), and then the dependencies between the variables Y_j are modeled through the copulas of (13).

2.3 Archimedean copulas

Our focus is on Archimedean copulas, a large parametric class with several attractive features. Archimedean copulas are characterized by the following relationship.

Let ϕ be a continuous strictly decreasing function from [0, 1] to $[0, \infty]$ such that $\phi(1) = 0$ and let $\phi^{[-1]}$ be its pseudo-inverse function. Let $C(v_1, \ldots, v_n)$ be a function from $[0, 1]^n$ to [0, 1] which satisfies

$$C(v_1, \ldots, v_n) = \phi^{[-1]}(\phi(v_1) + \cdots + \phi(v_n)). \tag{14}$$

Then, $C(v_1, \ldots, v_n)$ is an n-dimensional Archimedean copula. See Nelsen (1999).

From Diday and Vrac (2005), an n-dimensional Archimedean copula $C_n(v_1, \ldots, v_n)$ satisfies

$$C_n(v_1, \ldots, v_n) = \phi_n^{[-1]}(\phi_n(C_{n-1}(v_1, \ldots, v_{n-1})) + \phi_n(v_n)) \tag{15}$$

where

$$C_{n-1}(v_1, \ldots, v_{n-1}) = \phi_{n-1}^{[-1]}(\phi_{n-1}(C_{n-2}(v_1, \ldots, v_{n-2})) + \phi_{n-1}(v_{n-1})) \tag{16}$$

and so on, with $0 \leq v_1, \ldots, v_n \leq 1$ and where ϕ_i is a continuous strictly decreasing convex function, $i = 1, \ldots, n$. For a parametric copula, we note that in (15) and (16), $\phi_n(\cdot)$ and $\phi_{n-1}(\cdot)$ would contain parameters which can take different values as n changes.

The Frank (1979) family of copulas is given by, for $n = 2$, for $(v_1, v_2) \in [0, 1]^n$,

$$C(v_1, v_2; \beta) = (\ln \beta)^{-1} \ln\{1 + [(\beta^{v_1} - 1)(\beta^{v_2} - 1)]/(\beta - 1)\} \tag{17}$$

for $\beta > 0$ and $\beta \neq 1$; and is generated by $\phi_\beta(y) = -\ln[(1 - \beta^y)/(1 - \beta)]$. It follows that $\phi^{[-1]}(y) = [\ln(\beta)]^{-1} \ln[1 - (1 - \beta)e^{-y}]$. Hence, from (15), the Frank copula for $n > 2$ can be easily generated. Other important Archimedean copulas are the Clayton family 1978, the Genest–Ghoudi family (1994), and the Ali–Mikhail–Haq family (1978), among others; see Nelsen (1999). Properties of copulas are given in Nelsen (1999).

3 Estimation

The basic algorithms used (see Sect. 4) involve estimation of parameters. Behind these is the question of the choice of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$. These are covered in turn.

3.1 Estimation of the parameters

Optimizing any of the clustering criterion (such as (29) or (30) in Sect. 4) involves first estimating the n univariate distributions $G_Z(x; \boldsymbol{b})$ and the parameters $\boldsymbol{b}$ if a parametric $G_Z(\cdot)$ is taken, then the copula linking these functions $C(\cdot; \boldsymbol{\beta})$ which implies also estimating the copula parameters $\boldsymbol{\beta}$ when a parametric copula is used, and finally the mixture ratios p_k.

3.1.1 Estimation of $G_Z(x)$

Estimating a distribution function and/or the related probability density function has received considerable attention in the literature. For example, Silverman (1986) provides an excellent introduction to empirical density estimation techniques; and Prakasa Rao (1983) studies theoretical aspects of the subject. One approach would be to adapt these methods to the notion of copulas and mixture distributions. Thus, estimation

of the distributions, $G_Z(x)$, can be achieved by extending the classical histogram approach to give the empirical frequency as given in (3).

A second approach is to use a kernel density function. There are many possibilities. One such choice is an adaptation of the Parzen (1962) truncated window approach. Hence, the distribution function $G_Z(x)$ can be estimated through

$$\hat{f}(x) = \frac{1}{c_N} \frac{1}{Nh} \sum_{u=1}^{N} Ke\left(\frac{x - x_u}{h}\right) \tag{18}$$

where c_N is such that $\int_0^1 \hat{f}(x)dx = 1$, Ke is the kernel function and h is the window width. One choice of h is that value automatically estimated by the mean integrated square error (MISE) formula $h = 1.06\sigma N^{-1/5}$ where σ is the standard deviation calculated from the sample, when the kernel function being used is the standard normal density; this choice is typically used when the true probability density function is not known. The constant 1.06 changes for other kernel functions. Details for choices of kernel Ke and the calculation of the window h can be found in Silverman (1986).

Alternatively, parametric approaches could be used. For example, the distribution $G(x)$ could be modeled as a Dirichlet's law. In one dimension, this becomes the beta law

$$f(x; \boldsymbol{b}) = \frac{x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1}}{\int_0^1 y^{\alpha_1 - 1}(1 - y)^{\alpha_2 - 1} dy}, \quad 0 < x < 1, \tag{19}$$

where $\boldsymbol{b} = (\alpha_1, \alpha_2)$ are parameters with $\alpha_i > 0, i = 1, 2$. Hence, we can determine

$$G(x; \boldsymbol{b}) = \int_0^x f(t; \boldsymbol{b})dt.$$

The parameter $\boldsymbol{b}$ can be estimated using classical techniques such as the maximum likelihood method to give $\hat{G}(x; \boldsymbol{b}) = G(x; \hat{\boldsymbol{b}})$. Another approach is to use a Gaussian law for $f(\cdot)$.

3.1.2 Estimation of the copulas

For discussion purposes, let us assume we wish to work with the log-likelihood classification criterion (30), and a parametric copula. The parameters of the copulas to be estimated must maximize the function $L = W_2(P, \boldsymbol{\gamma}')$. If each observation u is described by F_u, let $\{F_u(Z_i), i = 1, \ldots, n\}$ be denoted by $\{x_i, i = 1, \ldots, n\}$. Then, the parameters $\boldsymbol{\beta}_k, k = 1, \ldots, K$, are estimated to be those which maximize

$$L = \sum_{k=1}^{K} \sum_{u \in P_k} \ln \left[\left\{ \prod_{i=1}^{n} \frac{dG_{Z_i}}{dx_i}\left(x_i; \boldsymbol{b}_i^k\right) \right\} \times \frac{\partial^n}{\partial x_1 \ldots \partial x_n} \right.$$
$$\left. \times \; C_k\left(G_{Z_1}^k\left(x_1; \boldsymbol{b}_1^k\right), \ldots, G_{Z_n}^k\left(x_n; \boldsymbol{b}_n^k\right); \boldsymbol{\beta}_k\right) \right] \tag{20}$$

for given specified copula functions $C_k(\cdot)$. For the Frank family of copulas (17), we can show that when $n = 2$, writing $C \equiv C_k(x_1, x_2; \beta)$,

$$\frac{\partial^2 C}{\partial x_1 \partial x_2} = \frac{(\beta - 1)\beta^{x_1 + x_2} \ln \beta}{[(\beta - 1) + (\beta_1^x - 1)(\beta_2^x - 1)]^2}. \tag{21}$$

For the Clayton family of copulas,

$$\frac{\partial^2 C}{\partial u \partial v} = (\beta + 1)(uv)^{-\beta - 1}(u^{-\beta} + v^{-\beta} - 1)^{-2 - 1/\beta};$$

for the Genest–Ghoudi family,

$$\frac{\partial^2 C}{\partial u \partial v} = \frac{1}{\beta}\left(\frac{1}{\beta} - 1\right)\{1 - [(1 + u^\beta)^{1/\beta} + (1 + v^\beta)^{1/\beta}]^\beta\}^{1/\beta - 2}(uv)^{\beta - 1}$$
$$\times [(1 + u^\beta)(1 + v^\beta)]^{1/\beta - 1};$$

and for the Ali–Mikhail–Haq family,

$$\frac{\partial^2 C}{\partial u \partial v} = \frac{(1 - \beta)[1 + \beta(1 - u)(1 - v)]}{[1 + \beta(1 - u)(1 - v)]^3}.$$

The relevant copula derivative is then substituted into the function L in (20) as appropriate.

Numerical iteration The function L of (20) can be maximized using numerical methods. Let us write $W_i^k \equiv G_{Z_i}^k(x_i; \boldsymbol{b}_i^k)$, $i = 1, \ldots, n$. i.e., we can write (20) as

$$L = \sum_{k=1}^{K} \sum_{u \in P_k} \ln \left[\left\{ \prod_{i=1}^{n} \frac{d}{dx_i} G_{Z_i}\left(x_i; \boldsymbol{b}_i^k\right) \right\} \times l_k(\boldsymbol{\beta}_k) \right]$$

with

$$l_k(\boldsymbol{\beta}_k) = \frac{\partial^n}{\partial w_1^k \ldots \partial w_n^k} C_k\left(G_{Z_1}^k\left(x_1, \boldsymbol{b}_1^k\right), \ldots, G_{Z_n}^k\left(x_n; \boldsymbol{b}_n^k\right); \boldsymbol{\beta}_k\right) \tag{22}$$

where $\boldsymbol{\beta}_k$ is the vector of parameters associated with the copula $C_k(\cdot)$. Then, estimating the copula $C_k(\cdot)$ involves finding those $\boldsymbol{\beta}_k$ which maximize $l_k(\boldsymbol{\beta}_k)$ in (22)

for specified $C_k(\cdot;\boldsymbol{\beta}_k)$, for each k. If explicit expressions for $\hat{\boldsymbol{\beta}}_k$ cannot be obtained, numerical methods are employed.

One such method is the Newton-Raphson technique. Thus, for example, when $n = 3$, we use the iterative relationship at each iteration $s, s = 1, 2, \ldots$, writing $l_k(\boldsymbol{\beta}_k) \equiv l(\boldsymbol{\beta})$ for simplicity,

$$\boldsymbol{\beta}^{s+1} = \boldsymbol{\beta}^s + \{I(\hat{\boldsymbol{\beta}})\}^{-1}\text{grad}(\boldsymbol{\beta}^s) \tag{23}$$

where the information matrix is

$$I(\boldsymbol{\beta}) = \left(\frac{-\partial^2 l(\beta_1, \beta_2)}{\partial\beta_i\,\partial\beta_j}\right), \quad i, j = 1, 2,$$

and the gradient vector is

$$\text{grad}(\boldsymbol{\beta}) = (\partial l(\beta_1, \beta_2)/\partial\beta_i), \quad i = 1, 2,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2)$ is two-dimensional for an $n = 3$ dimensional Archimedean copula and where $I(\hat{\boldsymbol{\beta}})$ is estimated by $I(\boldsymbol{\beta}^s)$.

Copula functions for more than two variables can be quite difficult to define. However, when, in (22), $n > 2$, this difficulty can be circumvented by exploiting the relationship (15) which relates an n-dimensional copula to a two-dimensional copula. To illustrate, let $W_i \equiv W_i^k, i = 1, 2, 3$, denote the $n = 3$ random variables in (22). We consider the copula $C_1(w_1, w_2; \beta_1)$ to be the link between the variables W_1 and W_2 and the copula $C_2(\cdot)$ as the link between the random variables $C_1(\cdot;\beta_1)$ and W_3, viz.,

$$C_2(C_1(w_1, w_2; \beta_1), w_3; \beta_2).$$

We first estimate β_1 and hence $C_1(\cdot;\beta_1)$ from realizations $(w_{11}, \ldots, w_{1N})$ and $(w_{21}, \ldots, w_{2N})$ as described above. This allows us to compute realizations of $C_1(\cdot;\beta_1)$ as $\{C_1(w_{11}, w_{21}; \hat{\beta}_1), \ldots, C_1(w_{1N}, w_{2N}; \hat{\beta}_1)\}$. These realizations along with the $(w_{31}, \ldots, w_{3N})$ are used to estimate β_2 and hence we can estimate $C_2(w_1, w_2, w_3; \hat{\beta}_2)$. Continuing in this manner, we can estimate $C_n(w_1, \ldots, w_n; \hat{\boldsymbol{\beta}})$ where now $\hat{\boldsymbol{\beta}} \equiv (\hat{\beta}_1, \ldots, \hat{\beta}_{n-1})$. Note that even so when the number of dimensions is large, care is needed to implement this procedure.

Correlation coefficients An alternative approach is to estimate the underlying Archimedean copulas through correlation coefficients. For notational simplicity, let us assume we have the n random variables $X_1, \ldots, X_n$ with joint distribution function $H(x_1, \ldots, x_n)$ and marginal distribution functions $F_1(x_1), \ldots, F_n(x_n)$, respectively, with the dependencies expressed through the copula $C_n(\cdot;\boldsymbol{\beta})$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n-1})$, as in Sklar's Theorem, i.e.,

$$H(x_1, \ldots, x_n) = C_n(F_1(x_1), \ldots, F_n(x_n); \boldsymbol{\beta}) \equiv C_n(\boldsymbol{\beta}).$$

By extending Nelsen (1999), it is easily shown (see Hillali 1998) that Kendall's coefficient of association τ satisfies, for $n = 2$,

$$\tau = \tau\{C_2(\boldsymbol{\beta})\} = \left\{2^2 \int C_2(v_1, v_2; \boldsymbol{\beta}) dC_2(v_1, v_2; \boldsymbol{\beta}) - 1\right\}.$$

Hence, by estimating τ and exploiting this relationship, the copula C can be found. One approach is to extend the idea of Hillali (1998) as follows. We wish to estimate the copula through (15). For clarity, let us take the case $n = 3$ (therefore, $\boldsymbol{\beta} = (\beta_1, \beta_2)$), and let X_i have realizations $\{x_{i1}, \ldots, x_{iN}), i = 1, 2, 3$. For each (X_i, X_j) pair, we estimate the corresponding $\hat{\tau}$ value and take the average to give

$$T^* = \{\hat{\tau}(X_1, X_2) + \hat{\tau}(X_1, X_3) + \hat{\tau}(X_2, X_3)\}/3. \tag{24}$$

Then, the estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ satisfies the relationship

$$\tau\{C_3(\hat{\boldsymbol{\beta}})\} = T^*. \tag{25}$$

The parameter β_1 can be viewed as the coefficient of association between X_1 and X_2 so that it is estimated by

$$\hat{\beta}_1 = \tau^{-1}\{\hat{\tau}(X_1, X_2)\}. \tag{26}$$

Then, in turn, $\hat{\beta}_2$ is the value of β_2 which satisfies $\tau\{C(\hat{\beta}_1, \beta_2)\} = T^*$.

Notice that $\tau\{C(\beta_2)\}$ only depends on β_2 and not on the distributions (X_1, X_2) and (X_1, X_3); therefore it is not possible to estimate β_2 from $\tau\{C(\beta_2)\}$ and then β_1 from $\tau\{C(\beta_1, \hat{\beta}_2)\} = T^*$. This difficulty is avoided by estimating β_1, as in Hillali's method from (25). We then can estimate $C_1(X_1, X_2; \beta_1)$ and hence determine its realizations

$$\{C_1(X_{11}, X_{21}; \hat{\beta}_1), \ldots, C_1(X_{1N}, X_{2N}; \hat{\beta}_1)\}.$$

The parameter β_2 is now interpreted as the coefficient of association between the random variables $Z = C(X_1, X_2; \beta_1)$ and X_3. Then, the parameter β_2 is estimated as being that value which satisfies

$$\tau\{C_2(\beta_2)\} = \hat{\tau}(Z, X_3). \tag{27}$$

The generalization to $n > 3$ variables flows through. As such, this method is well adapted to n-dimensional copulas in general through (15).

Or, instead of using Kendall's τ, we can use Spearman's ρ, where now

$$\rho = \rho\{C_n(\boldsymbol{\beta})\} = \frac{1}{[(n+1)^{-1} - 2^{-n}]} \left\{\int v_1 \ldots v_n dC_n(v_1, \ldots, v_n; \boldsymbol{\beta}) - 2^{-n}\right\}.$$

The same ideas carry through where now τ is replaced by ρ in the Eqs. 24–27 above. See Vrac (2002) for details; see also Genest and Rivest (1993) for estimation of bivariate Archimedean copulas for classical data using Kendall's τ.

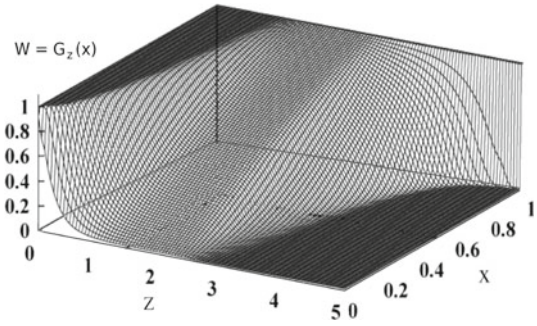

Fig. 2 Surface distribution of distributions in Fig. 1 Data—calculated using Parzen's window for $h(Z) = MISE$

3.1.3 Estimation of $\{p_k\}$

The mixing ratios $\{p_k, k = 1, \ldots, K\}$ are estimated in the usual way with

$$\hat{p}_k = \frac{\text{card}(P_k)}{\text{card}(\mathfrak{F})}. \tag{28}$$

Alternative estimators of p_k are suggested in Celeux and Govaert (1993).

3.2 Choice of Z

The estimation steps of Sect. 4 presuppose values of Z have been chosen. These choices can be induced by the nature of the estimated function of the distributions $G_Z(x)$, and the densities of these distributions $g_Z(x)$.

The *surface* S *of distributions of distribution values* $G_Z(x)$, associated with the population Ω and the random variable in the domain V, is $S = \{(Z, x, w); Z \in V; x \in [0, 1]; w = G_Z(x)\}$. The *surface* S' *of densities of distributions* $g_Z(x)$ associated with the population Ω and the random variable in the domain V, is $S' = \{(Z, x, w); Z \in V; x \in [0, 1]; w = g_Z(x)\}$.

For the data of Fig. 1, the surface S of the distributions $G_Z(x)$ is shown in Fig. 2, where representations of the $G_Z(x)$ each in one dimension are shown for several values of Z. Here, each $G_Z(x)$ was estimated via the kernel density method for a Gaussian kernel using the Parzen truncated window. The window width $h \equiv h(Z)$ was calculated by the mean integrated square error formula with the standard deviation σ estimated from the sample $\{F_1(Z), \ldots, F_N(Z)\}$ for each Z. By taking the derivative of the surfaces $G_Z(x)$, we can obtain the corresponding density functions $g_Z(x)$ of the observed distributions.

Intuitively, natural choices of the Z's correspond to changes in the nature of these surfaces. That this is so follows from recognizing that a given choice of Z is not good if all the observed distributions of F_u in the distributions base $\mathfrak{F}$ have the same value at that Z, as this would inhibit the partitioning process. Rather, good choices of Z are those Z_{i^*} (say) for which there exist distinct classes of values among the set of values $\{F_u(Z_{i^*}), u = 1, \ldots, N\}$. Equally important, a priori knowledge from experts (in the area from which the data were generated) can help identify where such "bumps"

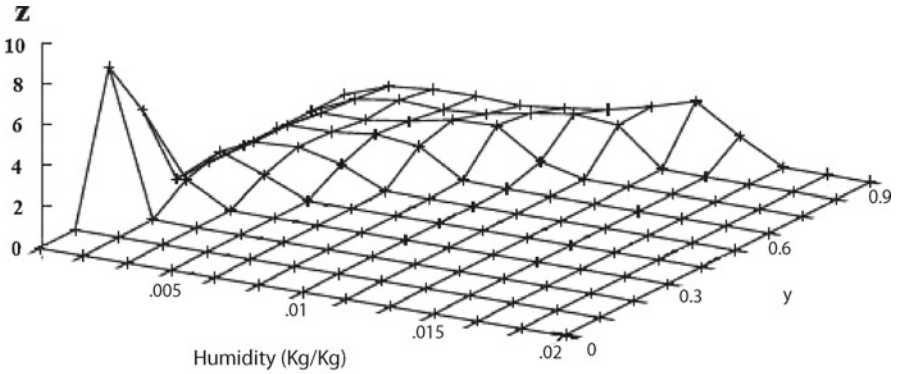

Fig. 3 Surface densities of distributions for humidity data—calculated from 16,200 observed distributions

might occur. For example, Fig. 3 shows the surface S' of the densities $g_Z(x)$ calculated from the 16,200 distributions of the humidities from the climatology data considered in Sect. 5. The clear inflection point at $Z_1 = 0.000003$ identifies this as a suitable Z value; whereas the $Z_2 = 0.006$ value also used in the actual analysis (in Sect. 5) comes from experts.

Although the definitions of the surfaces S and S' are written and illustrated here for the case $p = 1$, they can be extended to the general case $p > 1$. However, the visual representation of choosing the Z values might be complex when more than one dimension is used.

Vrac (2002) and Diday and Vrac (2005) proposed a triangle method to assist in the choices of Z; Jain and Dubes (1988) also proposed methods to help identify clustering tendencies. While for the data considered in Sect. 5, the actual specifics of these Z values were not an issue, the question of what might be in general the best choices and how many Z's remains. It is known, however, that convergence does occur regardless of the number n of Z's used for empirical copulas; see Vrac (2002).

4 Clustering algorithm

The clustering algorithm proposed is one obtained by adapting the dynamical clustering method developed by Diday et al. (1974) and Celeux et al. (1989) for classical observations in pattern recognition and by Symons (1981) for clustering multinormal observations, to the present situation whereby we seek the best grouping of the N distributions (observations) in $\mathfrak{F}$ into K classes $P = (P_1, \ldots, P_K)$. The main idea at each iteration/step, is to estimate the parameters of the densities $h_k(\cdot)$, $k = 1, \ldots, K$, which best describe the classes for the current partition according to a specified given clustering criterion.

For each partition, this involves determining the distributions $G^k_{Z_i}(\cdot)$, $i = 1, \ldots, n$, $k = 1, \ldots, K$, and thence estimating its parameters $\boldsymbol{b}^k_i$ whenever a parametric G is taken. It also involves fixing the copula models $C_k(\cdot)$, $k = 1, \ldots, K$, and includes

estimating the associated parameters $\boldsymbol{\beta}_k$ when parametric copulas are chosen. Details of these estimations were given in Sect. 3.

There are many possible clustering criteria, $W(P, \boldsymbol{\gamma})$, with associated parameters $\boldsymbol{\gamma}$, that can be used to determine the best partition $P = (P_1, \ldots, P_K)$, such as the log-likelihood criterion (see, e.g., Symons 1981)

$$W_1(P, \boldsymbol{\gamma}) = \sum_{u=1}^{N} \ln \left[\sum_{k=1}^{K} p_k h_k(F_u(Z_1), \ldots, F_u(Z_n); \boldsymbol{\gamma}_k) \right]; \tag{29}$$

or, a classification criterion such as the widely used log-likelihood classification criterion (see, e.g., Celeux et al. 1989)

$$W_2(P, \boldsymbol{\gamma}') = \sum_{k=1}^{K} \sum_{u \in P_k} \ln[h_k(F_u(Z_1), \ldots, F_u(Z_n); \boldsymbol{\gamma}_k)] \tag{30}$$

where now $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}_k,\ k = 1, \ldots, K)$. Notice that this criterion does not use the mixing probabilities $p_k, k = 1, \ldots, K$; it uses the distribution functions $h_k, k = 1, \ldots, K$, directly. This allows for more robust clusters to be formed.

Although the log-likelihood criterion (29) is widely used within a clustering context, it is more generally employed when there is a greater interest in modeling/estimating the global distribution $h(\cdot)$. In contrast, the log-likelihood classification criterion (30) gives more importance to the conditional distributions $h_k(\cdot), k = 1, \ldots, K$, and thus is more useful when the focus is put more on the classes found from the partitioning process than on the modeling/estimation of the whole density. Both types of criteria are possible in the proposed method. Indeed, other types of clustering criteria can be used. What is important is that **a** criterion be selected, against which the optimal set of classes $(P_1, \ldots, P_K)$ can be ascertained.

Suppose we take the log-likelihood criterion (29). Let the initialization of the partition be $P^0 = (P_1^0, \ldots, P_K^0)$, and let the partition after the sth iteration be $P^s = (P_1^s, \ldots, P_K^s)$. Then, the algorithm consists of two successive and iterative steps, viz.,

Step 1: Estimation of the parameters of the mixture distribution (11) (or (12), as appropriate) by maximizing the selected criterion (e.g., (29)), based on P^s, to give p_k^{s+1} and $\boldsymbol{\gamma}_k^{s+1}$; and

Step 2: Definition of the new partition $\{P_k^{s+1}, k = 1, \ldots, K\}$ where P_k^{s+1} is defined as

$$P_k^{s+1} = \left\{ F_u;\ p_k^{s+1} h_k\left(F_u; \boldsymbol{\gamma}_k^{s+1}\right) \geq p_m^{s+1} h_m\left(F_u; \boldsymbol{\gamma}_m^{s+1}\right) \right. \\ \left. \text{for all } m \neq k, m = 1, \ldots, K \right\}. \tag{31}$$

When $|W(P^{s+1}, \boldsymbol{\gamma}^{s+1}) - W(P^s, \boldsymbol{\gamma}^s)| < \epsilon$, for some preassigned small value of ϵ, the process stops.

The allocation step (31) is written for a criterion such as (29); when a clustering classification criterion such as (30) is used, the mixing parameters p_k^{s+1} and p_m^{s+1}

terms in (31) are omitted. The basic idea is that at the $(s+1)$th iteration, units F_u are moved into (i.e., allocated to) the P_k^{s+1} which optimizes the partition at this iteration for the given partitioning criterion. Note that this 'move' can keep the unit F_u in the same class it occupied after the preceding iteration.

Since the algorithm is a K-means type algorithm, the initial partition can affect the final estimates of the parameters and the resulting clustering. This starting partition must be chosen as close as possible to the final partition. Hence, expert knowledge is certainly valuable. For example, in our climate application in Sect. 5, latitudinal strip-clusters were used as an initial partition. However, it is still possible that the algorithm converges toward a local minimum. Consequently, the algorithm is run several times for each desired number K of clusters, and the approximate weight of evidence (AWE) criterion (see Eq. 32 below) is used to obtain the final partition.

There are as many as three sets of parameters involved in Step 1, corresponding respectively to the mixture ratios $p_k, k = 1, \ldots, K$, the copula parameters $\boldsymbol{\beta}_k$, $k = 1, \ldots, K$, in $C(\cdot; \boldsymbol{\beta}_k)$, and the marginal distribution parameters $\boldsymbol{b}_i^k$, $k = 1, \ldots, K$, $i = 1, \ldots, n$, in $G(\cdot; \boldsymbol{b}_i^k)$, as detailed in Sect. 3. If a non-parametric marginal distribution $G(\cdot)$ is chosen, the algorithm can be applied in a similar manner; likewise, for a non-parametric copula. That this algorithm converges, and in a finite number $S^* \in \mathbb{N}$ of iterations, is proven (along with some other asymptotic properties) in the "Appendix".

This adaptation of the (Diday et al. 1974; Schroeder 1976; Scott and Symons 1971; Symons 1981) classical dynamical clustering method to distributions works well, as demonstrated by its application to some climatology data described in Sect. 5, and substantiated by the convergence properties. There are other classical optimization algorithms which could be considered for adaptation to the present situation; see Sect. 1.

Finally, in clustering analyses the number of classes K is usually prespecified as, to date, the literature does not provide a completely satisfactory method to assess K. There are many criteria that have been suggested in the literature. While it is not the goal of this paper to evaluate these criteria, one such criterion (used in the application of Sect. 5) is the approximate weight of evidence (AWE) criterion suggested by Banfield and Raftery (1993), viz., for given K,

$$AWE(K) = -2\log(L_C) + 2d(3/2 + \log N) \tag{32}$$

where L_C is the classification maximum likelihood (e.g., the maximized value of (30)), d is the number of parameters to be estimated, and N is the sample size. Then, the clustering algorithm is run for many specific values of K; that K which maximizes $AWE(K)$ is selected.

5 An application

5.1 Copula methodology

The foregoing theory is illustrated by an analysis of an atmospheric dataset covering the globe from the European Center for Medium-range Weather Forecasts (ECMWF)

located in Reading U. K. Data points are realized as grid points over the earth at each latitude and longitude degree, and extended in altitude to 37 temperature and 24 humidity data point levels. Clearly, technically, these are not random points. However, for simplification of the problem and illustration of the approach, those assumptions are made on the dataset. The temperatures used are those forecast six hours earlier for midnight on December 15, 1999 at 3-dimensional latitude $\times$ longitude $\times$ altitude grid points. The objective then is to partition the weather world into well-defined temperature and humidity ($p = 2$) regions by latitude and longitude based on these data including estimation of the underlying probability distribution function for each identified region. There are essentially two discretization steps involved, viz., discretize the globe by grids in three dimensions, and then discretize the surfaces S (or S') at these grids according to Z (see Sect. 3.2).

The first discretization step develops the temperature-humidity patterns for every other (i.e., 2° apart) latitude-longitude grid point. Hence, $N = 16,200$. At each of these N grid-points, the temperature distributions $F_u^1(\cdot)$ are calculated from the 37 temperature altitude level values; and likewise the humidity distributions $F_u^2(\cdot)$ are calculated from its 24 altitude level values. Hence, the main idea is that the distributions characterize the variability of the temperature and humidity all along the vertical of the grid-point (Note that the temperature usually does not simply decrease for higher altitudes, because of the phenomenon of inversion that occurs after the tropopause). This representation of the data is more informative than many of the classical representations typically used, such as the average, since the variation within each observation is retained for our method whereas it is lost when an average (say) is used. The temperature and humidity profiles $F_u = (F_u^1(\cdot), F_u^2(\cdot))$, $u = 1, \ldots, N$, are estimated (through (18)) by the Parzen method where we take the window h to be the mean integrated square error (MISE) values, and where in this case $c_N = 1$. The aim is to group these N distributions covering both temperature and humidity into K classes.

We give the results for the coupling approach (13) where $\boldsymbol{Y} = (Y_1, Y_2)$ where Y_1 = temperature and Y_2 = humidity. The $p = 2$ (or, equivalently, $n = n_1 + n_2 = 2 + 2 = 4$) values of $\boldsymbol{Y} \equiv \boldsymbol{Z}$ selected (at the second discretization step) were $\boldsymbol{Z} = ((Z_1^1, Z_2^1), (Z_1^2, Z_2^2)) = (225, 265, 0.00003, 0.006)$. The analysis was run on several choices of $\{Z_i^j,\ i = 1, \ldots, n\}$ and different numbers n for each $j = 1, 2$. For these data, the same results were obtained showing insensitivity to the actual number and choice of Z_i^j, due to the fact that the cumulative distribution functions were quite smooth. Also, the choice of the two temperature thresholds $Z_1^1 = 225\,\mathrm{K}$ and $Z_2^1 = 265\,\mathrm{K}$ (K $\equiv$ Kelvin degrees) corresponding to the 25th percentiles were determined (in consultation with a meteorological expert and by observing where the inflection points occurred in the surface of the distributions $G_Z(x)$ or the densities $g_Z(x)$) to be used in the estimation of the distributions. Additional analyses run by adding two additional Z_i values in each tail along with the selected 225 K and 265 K values also gave the same results. The humidity threshold values were determined as inflection points in $g_Z(x)$ and from experts (as illustrated in Sect. 3.2, where now $Z \equiv Y_2$).

A Frank family copula $C(\cdot; \boldsymbol{\beta}_k)$ of (17) was fitted and distributions $G_{Z_i}(\cdot; \alpha_1^k, \alpha_2^k)$ corresponding to the beta law of (19), $i = 1, 2$, were adopted for the kth class

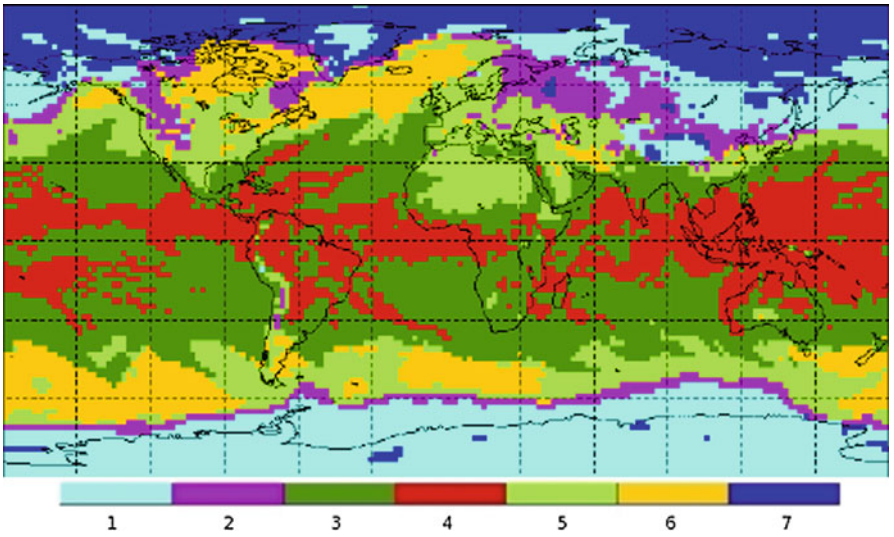

Fig. 4 Classification into 7 temperature and humidity regions—based on Frank copulas, beta $G(\cdot)$'s. Regions: *1* South Polar (Cold and Dry), *2* Sub-Polar (Relatively Cold and Dry), *3* SubTropical (Relatively Hot and Wet), *4* Tropical (Hot and Wet), *5* Temperate, *6* Sub-Temperate (Relatively Warm and Dry), *7*-North Polar (Frigid and Dry)

Table 1 Parameters of the classification in 7 clusters for temperature and humidity

Region	β'_k	α_1^k in Y_1	α_2^k in Y_1	α_1^k in Y_2	α_2^k in Y_2	p'_k
1	0.000001	6.71	2.14	5.70	5.22	0.20
2	0.100001	70.00	70.00	10.42	14.54	0.06
3	0.200001	18.97	88.13	8.06	145.22	0.25
4	0.050867	19.53	112.07	6.49	357.52	0.14
5	0.362295	12.32	31.49	5.03	18.55	0.14
6	0.126157	0.87	7.18	3.32	7.18	0.09
7	0.003896	23.22	4.77	13.37	3.11	0.12

(i.e., region), $k = 1, \ldots, K$. Also, the clustering criterion used was the log-likelihood criterion of (29). The initial partition was constructed according to latitudes by defining K strips of latitudes to give a kind of prior tropical class and two (or more, etc.) non-tropical classes. In our case, we wanted an odd number of classes to keep the geographical symmetry (in latitude) of the earth's atmosphere with respect to a central tropical cluster.

We ran our copula methodology for $K = 5, \ldots, 18$ classes and calculated the approximate weight of evidence (AWE) criterion (32) where in our case L_C is the maximized value of (29), $d = 2 \times K$, and $N = 16{,}200$ is the number of atmospheric profiles. For these data, this AWE criterion was maximized at $K = 7$.

The resulting classes and parameter estimates are shown in Fig. 4 and Table 1, respectively. Notice that for these classes, the estimated beta law parameters $(\hat{\alpha}_1^k, \hat{\alpha}_2^k)$

vary substantially across regions reflecting the highly variable weather patterns from one region to another (as should be expected). The results are good and consistent with those found for each variable analysed alone (not shown). First, note the tropical class 4 which describes particularly well a region of high meteorological significance, namely the Inter Tropical Convergence Zone (ITCZ) and more acute transitions to colder and drier classes further away from the equator. Secondly, the northern winter and the southern summer are identified; see, e.g., how the winter's north polar regions are quite colder than the summer's south polar region which is more like the sub-north-polar region. The distinctiveness of the Himalayas and Andes (both colder) and of the southern Australian desert (drier) from the surrounding geography is easily identified by this analysis. Also, the humidity spiral centered over $60°N$ and $60°E$ is observed. Also, estimates of the mixing parameters (in Table 1) are consistent with the surface coverage for these clusters (in Fig. 4); e.g., the Sub-Tropical region (class 3) at $p'_3 = 0.25$ reflects the fact this region covers more of the globe than does, say, the relatively cold and dry region (class 2) for which $p'_3 = 0.06$.

This seven class partition, described above, adds the desirable feature (that was missing in previous meteorological studies using $K = 5$ classes) that differences between winter in the northern hemisphere and summer in the southern hemisphere are clearly identified. Furthermore, prior climatological classes with $K = 5$ found the classes to be too large. For example, Chédin et al. (1985) and Achard (1991) used $K = 5$ classes corresponding to two polar, two temperate and one tropical classes. Because of the small number of classes, this partition in effect assumes equivalent behavior (i.e., similar thermodynamic profiles) in the winter in the northern hemisphere and in the summer in the southern hemisphere (and conversely), and does not properly describe the transitions between polar and temperate zones or between temperate and tropical zones.

As a complementary experiment, the temperature and humidity profiles of February 1 2000 (6 weeks ahead) have been classified onto the seven previously determined classes. This classification (i.e., determination of the best associated cluster) has been realized based on Eq. 31. The resulting map of distribution of the clusters is presented in Fig. 5. The agreement between the "forecasted" clusters and the map of observed mean temperature between 500 and 700 hectopascal (hPa), as well as with the map of total water vapor content (maps not shown) has great precision. Most of the water vapor and temperature structures are correctly retrieved with high accuracy. Notice also that the South Polar region is by February starting to cool down as it heads to its winter season and so becomes more like the former North Polar region (dark blue); and likewise, north subpolar regions (light blue) in the north are warming up with a larger portion becoming more like the December South Polar region. The faster change of the South Polar region compared to the North Polar region is due to the fact that only insolation plays a role in the south, whereas in the north the inertia of the temperature of the oceans can be predominant. Again, this feature is entirely consistent with known weather changes from December to February in these regions. Consequently, this clustering method allows the researcher not only to define precise and useful structures, but also to coherently infer the classes (or, clusters) associated with new statistical entities (here, the atmospheric profiles of a future day).

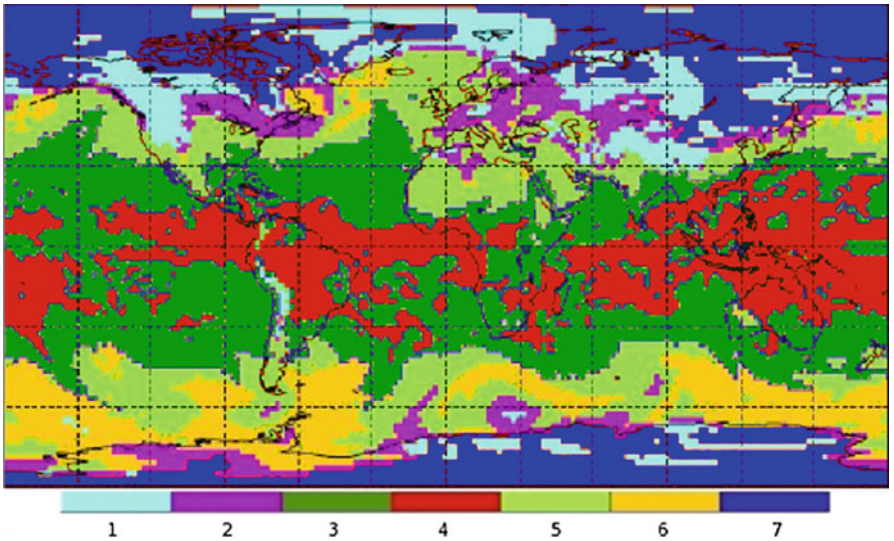

Fig. 5 Forecast (February 1 2000) Temperature and Humidity Regions—based on Frank copulas, beta $G(\cdot)$'s. Regions: *1* South Polar (Cold and Dry), *2* Sub-Polar (Relatively Cold and Dry), *3* SubTropical (Relatively Hot and Wet), *4* Tropical (Hot and Wet), *5* Temperate, *6* Sub-Temperate (Relatively Warm and Dry), *7* North Polar (Frigid and Dry)

Finally, the class distributions $h_k(\cdot), k = 1, \ldots, K$ and the mixture distributions $h(\cdot)$ can be calculated, if desired. The details are omitted; see Vrac (2002).

5.2 Comparison with EM algorithm

It is interesting to compare our approach with the EM method since both are based on statistical models to define clusters. Therefore, the data were also analysed by two EM clustering methods (Dempster et al. 1977), using the form of this algorithm as given in McLachlan and Peel (2000). In each case, relevant classes were found, based on both temperature and humidity variables; all gave results less consistent with climatological classifications of the globe as defined by experts compared with those for the copula method proposed herein. We describe these briefly. Complete details, including plots of the corresponding class regions, plots of the distribution functions for temperature and for humidity, by class, and detailed descriptions of similarities and differences with those given herein for the copula method are in Vrac (2002).

The first EM method was based on raw numerical data taking the 16,200 grid points and fixing values for 5 specific temperature and 5 specific humidity variables from the most reliable raw data values (37 temperatures and 24 humidity values) available. These specific values were those obtained by first running a standard classification and regression tree (CART) analysis on these (37 temperatures and 24 humidity) values with the referent classification being the seven clusters obtained from a hierarchical ascending clustering applied to these 61 variables. (For these data, the most

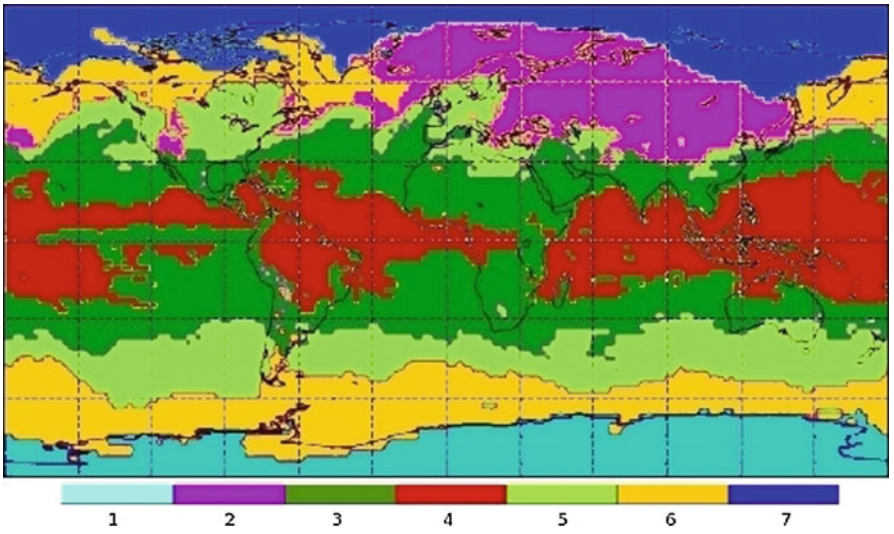

Fig. 6 Classification of 7 Temperature and Humidity Regions—EM Algorithm on Raw Data Regions: *1* South Polar (Cold and Dry), *2* Sub-Polar (Relatively Cold and Dry), *3* SubTropical (Relatively Hot and Wet), *4* Tropical (Hot and Wet), *5* Temperate, *6* Sub-Temperate (Relatively Warm and Dry), *7* North Polar (Frigid and Dry)

discriminant temperature (T) values were those at T_1, $(T_9 + T_{10})/2$, $(T_{16} + T_{17})/2$, $(T_{24} + T_{25})/2$ and $(T_{32} + T_{33})/2$; and the most discriminant humidity (H) values were those at H_1, $(H_4 + H_5)/2$, $(H_8 + H_9)/2$, $(H_{16} + H_{17})/2$ and $(H_{20} + H_{21})/2$, where the subscript refers to the altitude level measured at each grid point starting with the lowest altitude; e.g, T_1 is the temperature at the lowest altitude.) The seven classifications obtained from the EM algorithm applied to the resulting raw data values are as shown in Fig. 6. Comparing Fig. 6 with the classifications of Fig. 4, we see that classes are very poorly defined lacking, e.g., the dynamic nature of class boundaries with relatively 'smooth' edges. There is however a coherency in that the differences between northern winters and southern summers are identified. On the other hand, while there are some air incursions (albeit badly defined) in the Northern Hemisphere such as the Gulf Stream, there are none at all in the Southern Hemisphere. Furthermore, regions known to be tropical are identified as a mixture of regions. It is added that when, instead of using the CART approach, a principal component analysis was run retaining those which accounted for 90% of the variance to run the EM algorithm almost identical results (to those in Fig. 6) were obtained.

The second EM algorithm was based on the 16,200 probability distributions of temperature and humidity profiles for functional data, which estimates the parameters of a ($p = 2$, and $n_1 = n_2 = 2$; hence, $n_1 + n_2 = 4$) multivariate normal distribution without restrictions on the covariance matrix. This produced the classes of Fig. 7. This classification is an improvement over that of Fig. 6 in that class boundaries are more dynamic than for the first EM method. However, classes in general are not well defined. For example, class 7 (red) encompasses completely different atmospheric profiles, grouping together mountain areas such as the Himalayas and the Alps with

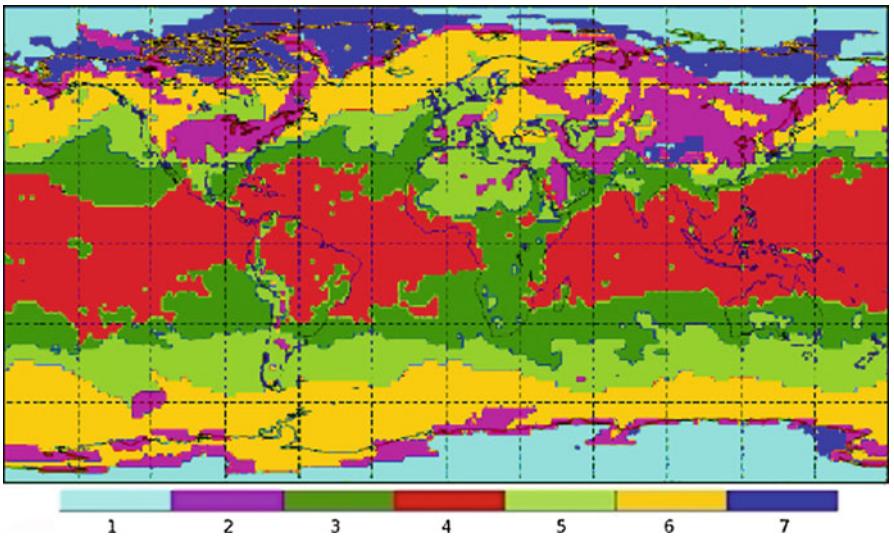

Fig. 7 Classification of 7 Temperature and Humidity Regions—EM Algorithm on Distributions Regions: *1* South Polar (Cold and Dry), *2* Sub-Polar (Relatively Cold and Dry), *3* SubTropical (Relatively Hot and Wet), *4* Tropical (Hot and Wet), *5* Temperate, *6* Sub-Temperate (Relatively Warm and Dry), *7* North Polar (Frigid and Dry)

polar oceanic and the American plains areas. The air incursion corresponding to the Gulf Stream is missing, as is the dry Southern Australian desert.

Thus we see that the copula methodology (i.e., Fig. 4) has produced results that are more consistent with global classifications as developed by climatology experts than have these EM algorithmic methods.

6 Conclusion

Based on the dataset analysed, the proposed methodology which incorporates copulas into clustering techniques, has produced more coherent classes than other known methods against which it was compared. It is known that EM methods are biased in terms of partitioning but unbiased in terms of law. In contrast, our method is unbiased in terms of partitioning but biased in terms of law. Comparisons with yet more known methods can be reasonably expected to reach the same conclusions. The present methodology was based on extending the dynamical clustering classical approach to distribution-valued data. Other approaches such as those based on other clustering algorithms should also be explored.

A number of questions remain for further development including a rigorous study of the number and choices of the $Z_i, i = 1, \ldots, n$, points in the $G_{Z_i}(\cdot)$ distributions and the implications of those choices. While for our data (where this was not an issue), and intuitively, the proposed procedure is robust, a definitive study of this issue also needs to be undertaken.

In a different direction, adding a spatial component to the methodology would be an interesting, and challenging problem left for future research. For the present application, it is preferable not to have a spatial component so as to let the algorithm find the regions/classes by itself without any a priori information. Moreover, we note that the classes obtained are not temporal classes in which a conditional spatial structure could be modeled. Instead, the classes are spatial regions, gathering together locations of the world having the same atmospheric conditions. Thus, it would be difficult to use a spatial structure when two locations in the same class (i.e., with relatively equivalent meteorological features) could be extremely far away from each other and separated by other classes (e.g., Himalayas and Polar regions).

On the other hand, whenever a classification of geographical regions is being explored, methods that take into account contiguity constraints could lead to better results. In our case, it was deemed preferable to let the clustering algorithm be as free as possible from geographical constraints. Indeed, in our application, some important climate phenomena (e.g., extreme precipitation, local low pressure systems) are on such a small spatial scale that incorporating contiguity constraints could prevent the algorithm from capturing such events. However, incorporating constraints into the algorithm is well worthy of future consideration.

Further, our approach can be viewed as a form of hierarchical modeling where the bottom level of the hierarchy—the raw data—is removed. In our case, the first level is the determination of the marginal distributions $G(\cdot)$ which are then used to estimate the parameters of the copula functions $C(\cdot;\beta)$. Then standard hierarchical modeling techniques allow us to parameterize the model with distribution functions instead of density functions. It would be interesting to extend this approach as a Bayesian hierarchical model methodology, using, e.g., the ideas of Richardson and Green (1997). Note also that the distribution functions which formed the "raw data" of our method are special cases of the "functions" of functional data analysis.

Finally, there remain questions of identifiability of our mixture model. Here, by identifiability we mean that two mixtures of the form in (1) with components and mixing probabilities $(K,\ p_k)$ and $(K^*,\ p_k^*)$, respectively, are identifiable if $K = K^*$ and $p_k = p_k^*$, except for permutation of the component identifiers. It is known from Yakowitz and Spragins (1968) that a necessary and sufficient condition that the class of finite mixture of distributions ($f(\cdot)$ on the left side of (1)) be identifiable is that the class of distributions from which the mixture distributions are selected (the $f_k(\cdot)$'s on the right side of (1)) be linearly independent over $\mathbb{R}$ see also, e.g., Titterington et al. (1985), Li and Sedransk (1988), McLachlan and Peel (2000). Titterington et al. (1985) noted that in many cases, the Yakowitz and Spragins conditions can be hard to verify, though they do go on to show that it is often possible to establish their sufficiency conditions. In the present paper, we have a finite mixture of copulas (as in (9)). Intuitively, by analogy with Eqs. 1 and 9 (or Eqs. 8 and 9), it seems not unreasonable to suggest similar suitably adapted conditions would pertain for the copula setting. This may mean that attention would perforce be restricted to finite mixtures of specific copula families. We leave this open problem as a topic for future work. In contrast, for EM based algorithms, questions of identifiability have been studied by, e.g, Bock and Gibbons (1996), Chan and Kuk (1997), and Kuk and Chan (2001), with Kuk and Chan (2001) showing that when an identifiability problem exists, implementing the

unconstrained EM algorithm is valid and that the loss of uniqueness of the estimates is usually not a major issue.

Recalling Schweizer (1984) that "Distributions are the numbers of the future", we have developed a methodology for grouping N "observed" distribution functions into K classes, as but one step along the path pointed out by Schweizer. Perhaps the most important issue is the need to develop adequate analytical methods for different types of complex data, such as distributions, classes of data which will only grow as computers expand their capabilities.

The authors wish to thanks the referees for their careful reading and helpful comments which have improved the manuscript. Partial support from the National Science Foundation is gratefully acknowledged.

Appendix A: Convergence properties

Three convergence properties related to the use of the clustering algorithm (of Sect. 4) can be derived. In "Appendix A.1", its convergence to a locally optimal solution, in a finite number of iterations S^*, is proved; followed in "Appendix A.2" by the derivation of some asymptotic properties. In "Appendix A.3", convergence results for a global distribution function of distribution values are obtained. As these properties are developed, we remind ourselves that the estimation process is hierarchical, with first the marginal (cumulative) distribution functions being estimated and then the copula parameters being estimated from these estimated marginal distributions.

A.1 Convergence of the clustering algorithm

Proposition 1 *The algorithm for the mixture decomposition of copulas by the dynamical clustering algorithm (of Sect. 4) converges to a locally optimal solution in a finite number of iterations.*

Proof We prove this result for the log-likelihood classification criterion (30); the proof is similar for other clustering criteria. Let $P^s = (P_1^s, \ldots, P_K^s)$ denote the partition into K classes at the sth iteration; and let $\boldsymbol{\gamma}^s = (\boldsymbol{\gamma}_1^s, \ldots, \boldsymbol{\gamma}_K^s)$ denote the values of the parameters at the sth iteration.

Let us write (30), at the sth iteration, with $W_2(\cdot) \equiv W(\cdot)$, as

$$W(P^s; \boldsymbol{\gamma}^s) = \sum_{k=1}^{K} W(P_k^s; \boldsymbol{\gamma}_k^s),$$

$$W\left(P_k^s; \boldsymbol{\gamma}_k^s\right) = \sum_{u \in P_k^s} \ln\left[h_k\left(F_u(Z_1), \ldots, F_u(Z_n); \boldsymbol{\gamma}_k^s\right)\right]$$

where $W(P_k^s; \boldsymbol{\gamma}_k^s)$ is the log-likelihood classification criterion for the kth cluster; P^s is the class that results from the allocation process (Step 2) based on the parameters $\boldsymbol{\gamma}^{s-1}$; and $\boldsymbol{\gamma}^s = g(P^s)$ where $g(\cdot)$ is the parameterization function (in our case, the maximum likelihood method) which gives, at the sth iteration, the new estimates $\boldsymbol{\gamma}^s$

of the parameters based on these classes. We want to show that $\{W(P^s, \boldsymbol{\gamma}^s)\}$ converges, is increasing in value and is stationary. Here stationarity is defined to mean that there exists an integer S^* such that for all $s \geq S^*$, $W(P^s; \boldsymbol{\gamma}^s) = W(P^*; \boldsymbol{\gamma}^*)$, where P^* is the partition with parameters $\boldsymbol{\gamma}^*$ at the S^*'th iteration.

First, from (31) by definition it follows that

$$W\left(P^{s+1}; \boldsymbol{\gamma}^{s+1}\right) \geq W\left(P^s; \boldsymbol{\gamma}^{s+1}\right). \tag{33}$$

Next, we can show that $W(P_k^s; \boldsymbol{\gamma}_k^{s+1}) \geq W(P_k^s; \boldsymbol{\gamma}_k^s)$, by construction of the parameterization g. Since the function g is using the maximum likelihood method, we have for all possible $\boldsymbol{\gamma}_k^s$ calculated from P_k^s,

$$\boldsymbol{\gamma}_k^{s+1} = \arg\max_{\boldsymbol{\gamma}_k^s} \sum_{u \in P_k^s} \ln\left[h_k\left(F^u(Z_1), \ldots, F_u(Z_n); \boldsymbol{\gamma}_k^s\right)\right].$$

Therefore, for all $\boldsymbol{\gamma}_k^s$, it follows that

$$\sum_{u \in P_k^s} \ln\left[h_k\left(F_u(Z_1), \ldots, F_u(Z_n); \boldsymbol{\gamma}_k^{s+1}\right)\right] \geq \sum_{u \in P_k^s} \ln\left[h_k\left(F_u(Z_1), \ldots, F_u(Z_n); \boldsymbol{\gamma}_k^s\right)\right]$$

and hence $W(P_k^s; \boldsymbol{\gamma}_k^{s+1}) \geq W(P_k^s; \boldsymbol{\gamma}_k^s)$. Summing over each class, $k = 1, \ldots, K$, we have

$$W\left(P^s; \boldsymbol{\gamma}^{s+1}\right) \geq W\left(P^s; \boldsymbol{\gamma}^s\right). \tag{34}$$

Combining (33) and (34), we have

$$W\left(P^{s+1}; \boldsymbol{\gamma}^{s+1}\right) \geq W\left(P^s; \boldsymbol{\gamma}^{s+1}\right) \geq W\left(P^s; \boldsymbol{\gamma}^s\right). \tag{35}$$

The relation (35) therefore implies that $\{W(P^s, \boldsymbol{\gamma}^s), s \in \mathbb{N}\}$ is increasing and can only take a finite number of values since N is finite. Therefore, it converges in a finite number of iterations and is stationary in the sense that there exists $S^* \in \mathbb{N} | W(P^s, \boldsymbol{\gamma}^s) = W(P^{S^*}, \boldsymbol{\gamma}^{S^*})$ for all $s \geq S^*$.

Remark 1 Estimation of the copula parameters with a maximum likelihood based method requires specifying the function $G_{Z_i}(\cdot)$. In this (parametric) case, it is assumed that the form of these functions and their derivatives are known.

A.2 Asymptotic behavior

Let $\mathfrak{F} = \{\tilde{F}_1^{(r)}, \ldots, \tilde{F}_N^{(r)}\}$ denote a sample of N realizations of a random variable with values that are distribution functions. The function $\tilde{F}_u^{(r)}$ is an estimation of the true distribution function which describes the unit u, and is calculated from r_u numerical realizations $\{x_{ui}, i = 1, \ldots, r_u\}$ for the unit $u = 1, \ldots, N$. Without loss of generality,

we take $r_u = r$ for all $u = 1, \ldots, N$. Let us suppose the true distribution of F_u is Gaussian $\mathcal{N}(\mu_u, \sigma_u^2)$ and the parameters are estimated by

$$m_u = \tilde{\mu}_u = \frac{1}{r}\sum_{i=1}^{r} x_{ui}, \quad s_u^2 = \tilde{\sigma}_u^2 = \frac{1}{r-1}\sum_{i=1}^{r}(x_{ui} - \tilde{\mu}_u)^2. \tag{36}$$

For K classes, consider the following hypotheses:

H^1: There exists a partition into K classes $(P_1, \ldots, P_K)$ of the $\{\tilde{F}_u^{(r)}, u = 1, \ldots, N\}$ such that $P_k = \{\tilde{F}_u^{(r)} | \tilde{F}_u^{(r)}$ is an estimation of the Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k^2)\}$; i.e., P_k consists of those $\tilde{F}_u^{(r)}$ which are estimates of the distribution $\mathcal{N}(\mu_k, \sigma_k^2)$, $k = 1, \ldots, K$; or, equivalently,

H^2: Each distribution function $\tilde{F}_u^{(r)}$, $u = 1, \ldots, N$, is an estimation of one of the K Gaussian distributions $\mathcal{N}(\mu_k, \sigma_k^2)$, $k = 1, \ldots, K$.

Then, if each of H^1 and H^2 implies the other and if the estimated parameters $\tilde{\mu}_u$ and $\tilde{\sigma}_u$ of the Gaussian distributions for each individual u, are not biased, classical results on the convergence of estimators lead us to the following.

Corollary 1 *In the limit as r tends to infinity, $\tilde{F}_u^{(r)}$ converges uniformly to F_u, where F_u follows one of the K Gaussian distributions $\mathcal{N}(\mu_k, \sigma_k^2)$, $k = 1, \ldots, K$, for all $u = 1, \ldots, N$. That is, when r tends to infinity, the distributions $\tilde{F}_u^{(r)}$ from $\mathfrak{F}$ converge to the true distribution functions F_u describing the individuals, $u = 1, \ldots, N$.*

Remark 2 From $\{F_1, \ldots, F_N\}$, we can define the σ-algebra generated by each single function $\{F_u, u = 1, \ldots, N\}$; and we can define a probability measure $\mathbb{P}$ on $[\{F_1, \ldots, F_N\}, \sigma(\{F_{u,u=1,\ldots,N}\})]$, corresponding to a multinomial law with parameters $(p_1, \ldots, p_K)$, where

$$\mathbb{P}([F \in \{F_1, \ldots, F_N\}; F \in P_k]) = p_k \tag{37}$$

with $\sum_{k=1}^{K} p_k = 1$.

Moreover, if the $\{G_{Z_i}^k, i = 1, \ldots, n\}$ (obtained by $\mathfrak{F}$) from each class $k = 1, \ldots, K$, are modeled in an empirical way, then from classical results of functional analysis we have the following.

Corollary 2 *In each class $k = 1, \ldots, K$ and for each Z, the distribution G_Z^k of class k converges uniformly toward a Dirac distribution G_Z^{k*} at point $F_{\mathcal{N}_k}(Z)$, where the function G_Z^{k*} is defined by*

$$G_Z^{k*}(x) = \begin{cases} 0, & \text{if } x < F_{\mathcal{N}_k}(Z), \\ 1, & \text{if } x \geq F_{\mathcal{N}_k}(Z), \end{cases}$$

with $F_{\mathcal{N}_k}$ being the Gaussian distribution function with parameters (μ_k, σ_k^2).

We will also need the following.

Proposition 2 (Diday 2001) *The n-dimensional copula $C(\cdot)$ associated with the joint distribution $H_{Z_1,\ldots,Z_n}(x_1,\ldots,x_n)$ satisfies the properties (i) the domain of $H_{Z_1,\ldots,Z_n}(\cdot)$ is $[0,1]$, and (ii) $C = \Pi = \prod_{i=1}^{n} v_i$ or $C = Min = Min(v_1,\ldots,v_n)$.*

Then using Proposition 2 and Corollary 2, we obtain the following result.

Proposition 3 *In each class $k = 1,\ldots,K$, whatever the number n of values $Z_1,\ldots,Z_n$, if the copula C_k of the class k is defined in an empirical way, it converges toward the Min and product Π copulas; i.e., C_k converges to C_k^* with copula $C_k^* = Min = \Pi$. Moreover, $(x_1,\ldots,x_n) \in \mathbb{R}^n$, and*

$$C_k^*\left(G_{Z_1}^{k*}(x_1),\ldots,G_{Z_n}^{k*}(x_n)\right) \in \{0,1\}.$$

A.3 Mixture decomposition of *distribution function of distributions*

We have seen in Sect. 2 that the distribution functions of distribution values are themselves distribution functions. Instead of computing these distributions class by class, we can compute an estimation of the global distribution functions at each Z with a mixture decomposition of the distributions.

Proposition 4 *If the true probability laws of the observed individuals $\{F_u, u = 1,\ldots,N\}$ are in the classes $(P_1,\ldots,P_K)$ of a partition into K classes according to a multinomial law with parameters $(p_1,\ldots,p_K)$, and if G_Z^k is the distribution function of distribution values in class k at Z, then the global distribution at point Z (called G_Z), is*

$$G_Z(x) = \sum_{k=1}^{K} p_k G_Z^k(x). \tag{38}$$

The parameter p_k is the probability that the true distribution function F_u is in class P_k.

From Corollary 2 and (38), we have the following.

Proposition 5 *For each value Z, the global distribution G_Z defined in (38) converges uniformly toward a distribution G_Z^* defined by:*

$$G_Z^*(x) = \begin{cases} 0, & \text{if } x < F_{\mathcal{N}_1}(Z), \\ \sum_{k'=1}^{k} p_{k'}, & \text{if } F_{\mathcal{N}_k}(Z) \le x < F_{\mathcal{N}_{k+1}}(Z), \\ 1, & \text{if } x \ge F_{\mathcal{N}_K}(Z), \end{cases}$$

with $F_{\mathcal{N}_k}$ being the Gaussian distribution function $\mathcal{N}(\mu_k,\sigma_k^2)$, $k = 1,\ldots,K$. In this proposition, it is assumed that $F_{\mathcal{N}_1}(Z) < \cdots < F_{\mathcal{N}_K}(Z)$.

Moreover, we have seen in Sect. 2 that the joint distribution function at points $Z_1,\ldots,Z_n$ can be written as given in (9). Hence, from Sklar's Theorem, we have:

Proposition 6 *Let X_i denote the random variable characterized by G_{Z_i} (the global distribution function at point Z_i), $i = 1, \ldots, n$; and let the joint distribution function of $(X_1, \ldots, X_n)$ be denoted by $H_{Z_1,\ldots,Z_n}$. Then there exists a copula C such that for all $(x_1, \ldots, x_n) \in [0, 1]^n$,*

$$\begin{aligned} H_{Z_1,\ldots,Z_n}(x_1, \ldots, x_n; \boldsymbol{\gamma}) &= C(G_{Z_1}(x_1; \boldsymbol{b}_1), \ldots, G_{Z_n}(x_n; \boldsymbol{b}_n); \boldsymbol{\beta}) \\ &= C\left(\sum_{k=1}^{K} p_k G^k_{Z_1}(x_1; \boldsymbol{b}^k_1), \ldots, \sum_{k=1}^{K} p_k G^k_{Z_n}(x_n; \boldsymbol{b}^k_n); \boldsymbol{\beta} \right). \end{aligned} \tag{39}$$

From Eq. 39, we deduce there exists a relationship between the mixture of copulas and the mixture of distributions; this relationship based on the copula C in Proposition 6 is:

$$\begin{aligned} &\sum_{k=1}^{K} p_k C_k \left(G^k_{Z_1}(x_1; \boldsymbol{b}^k_1), \ldots, G^k_{Z_n}(x_n; \boldsymbol{b}^k_n); \boldsymbol{\beta}_k \right) \\ &\quad = C\left(\sum_{k=1}^{K} p_k G^k_{Z_1}(x_1; \boldsymbol{b}^k_1), \ldots, \sum_{k=1}^{K} p_k G^k_{Z_n}(x_n; \boldsymbol{b}^k_n); \boldsymbol{\beta}_k \right). \end{aligned} \tag{40}$$

References

Achard V (1991) Trois Problemes dés de d'Analyse 3D de la Structure Thermodynamique de l'Atmosphére par Satellite: Mesure du Contenu en Ozone; Classification des Masses d'Air; Modélisation Hyper Rapide du Transfert Radiatif. Ph.D. Dissertation, University of Paris

Ali MM, Mikhail NN, Haq MS (1978) A class of bivariate distributions including the bivariate logistic. J Multivar Anal 8:405–412

Arabie P, Carroll JD (1980) MAPCLUS: a mathematical programming approach to fitting the ADCLUS model. Psychometrika 45:211–235

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49: 803–821

Bishop CM (1995) Neural networks for pattern recognition. Clarendon Press, Oxford

Bock HH (1998) Clustering and neural networks. In: Rizzi A, Vichi M, Bock HH (eds) Advances in data science and classification. Springer, Berlin, pp 265–277

Bock RD, Gibbons RD (1996) High-dimensional multivariate probit analysis. Biometrics 52:1183–1194

Brossier G (1990) Piecewise hierarchical clustering. J Classif 7:197–216

Celeux G, Diday E, Govaert G, Lechevallier Y, Ralambondrainy H (1989) Classification automatique des données. Dunod Informatique, Paris

Celeux G, Diebolt J (1986) L'Algorithme SEM: Un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densities. Revue de Statistiques Appliquées 34:35–51

Celeux G, Govaert G (1992) A classification EM algorithm for clustering and two stochastic versions. Comput Stat Data Anal 14:315–332

Celeux G, Govaert G (1993) Comparison of the mixture and the classification maximum likelihood in cluster analysis. J Stat Comput Simul 47:127–146

Chédin A, Scott N, Wahiche C, Moulinier P (1985) The improved initialization inversion method: a high resolution physical method for temperature retrievals from satellites of tiros-n series. J Appl Meteorol 24:128–143

Chan JSK, Kuk AYC (1997) Maximum likelihood estimation for probit-linear mixed models with correlated random effects. Biometrics 53:86–97

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65:141–151
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38
Diday E (1984) Une représentation visuelle des classes empietantes: les pyramides. Rapport de Recherche 291 INRIA
Diday E (2001) A generalization of the mixture decomposition problem in the symbolic data analysis framework. Rapport de Recherche, CEREMADE 112:1–14
Diday E, Schroeder A, Ok Y (1974) The dynamic clusters method in pattern recognition. In: Proceedings of international federation for information processing congress. Elsevier, New York, pp 691–697
Diday E, Vrac M (2005) Mixture decomposition of distributions by copulas in the symbolic data analysis framework. Discrete Appl Math 147:27–41
Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97:611–631
Frank MJ (1979) On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. Aequationes Mathematicae 19:194–226
Genest C, Ghoudi K (1994) Une famille de lois bidimensionelles insolite. Compte Rendus Academy Sciences Paris I 318:351–354
Genest C, MacKay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283
Genest C, Rivest LP (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88:1034–1043
Gordon A (1999) Classification. 2nd edn. Chapman and Hall, Boca Raton
Hartigan JA, Wong MA (1979) Algorithm AS136. A k-means clustering algorithm. Appl Stat 28:100–108
Hillali Y (1998) Analyse et modélisation des données probabilistes: Capacités et lois multidimensionelles. Ph.D. Dissertation, University of Paris
Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice Hall, New Jersey
James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. J Am Stat Assoc 98:397–408
Kuk AYC, Chan JSK (2001) Three ways of implementing the EM algorithm when parameters are not identifiable. Biometric J 43:207–218
Li LA, Sedransk N (1988) Mixtures of distributions: a topological approach. Ann Stat 16:1623–1634
McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York
Meng XL, Rubin DB (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. J Am Stat Assoc 86:899–909
Nelsen RB (1999) An introduction to copulas. Springer, New York
Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076
Prakasa Rao BLS (1983) Nonparametric functional estimation. Academic Press, New York
Redner RA, Walker H (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev 26:195–239
Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components. J R Stat Soc Ser B 59:731–792
Schroeder A (1976) Analyse d'un mélange de distributions de probabilité de même type. Revue de Statistiques Appliquées 24:39–62
Schweizer B, Sklar A (1983) Probabilistic metric spaces. North-Holland, New York
Schweizer B (1984) Distributions are the numbers of the future. In: diNola A, Ventre A (eds) Proceedings of the mathematics of fuzzy systems meeting, Naples, Italy. University of Naples, Naples, pp 137–149
Scott AJ, Symons MJ (1971) Clustering methods based on likelihood ratio criteria. Biometrics 27:387–397
Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London
Sklar A (1959) Fonction de répartition a n dimensions et leurs marges. Institute Statistics Université de Paris 8:229–231
Symons MJ (1981) Clustering criteria and multivariate normal mixtures. Biometrics 37:35–43
Tanner MA, Wong WH (1987) The calculation of posterior distribution by data augmentation (with discussion). J Am Stat Assoc 82:528–550
Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York
Vrac M (2002) Analyse et modélisation de données probabilistes par decomposition de mélange de copules et application á une base de données climatologiques. Ph.D. Dissertation, University of Paris

Vrac M, Ch*é*din A, Diday E (2005) Clustering a global field of atmospheric profiles by mixture decomposition of copulas. J Atmos Ocean Technol 22:1445–1459

Wei GCG, Tanner MA (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J Am Stat Assoc 85:699–704

Winsberg S, DeSoete G (1999) Latent class models for time series analysis. Appl Stoch Models Bus Ind 15:183–194

Yakowitz SJ, Spragins LD (1968) On the identifiability of finite mixtures. Ann Math Stat 39:209–214