Computational Statistics and Data Analysis 56 (2012) 915–927

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Fuzzy and possibilistic clustering for fuzzy data

Renato Coppi [a,*], Pierpaolo D'Urso [b], Paolo Giordani [a]

[a] *Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy*

[b] *Dipartimento di Analisi Economiche e Sociali, Sapienza Università di Roma, Rome, Italy*

ARTICLE INFO

Article history:
Received 4 January 2010
Received in revised form 16 September 2010
Accepted 17 September 2010
Available online 23 September 2010

Keywords:
Possibilistic models
Cluster analysis
LR fuzzy data
Fuzzy *k*-means
Possibilistic *k*-means

ABSTRACT

The Fuzzy *k*-Means clustering model (F*k*M) is a powerful tool for classifying objects into a set of *k* homogeneous clusters by means of the membership degrees of an object in a cluster. In F*k*M, for each object, the sum of the membership degrees in the clusters must be equal to one. Such a constraint may cause meaningless results, especially when noise is present. To avoid this drawback, it is possible to relax the constraint, leading to the so-called Possibilistic *k*-Means clustering model (P*k*M). In particular, attention is paid to the case in which the empirical information is affected by imprecision or vagueness. This is handled by means of *LR* fuzzy numbers. An F*k*M model for *LR* fuzzy data is firstly developed and a P*k*M model for the same type of data is then proposed. The results of a simulation experiment and of two applications to real world fuzzy data confirm the validity of both models, while providing indications as to some advantages connected with the use of the possibilistic approach.

1. Introduction

Clustering or "unsupervised classification" consists of shaping classes from a set of objects, based on knowing some of their properties. From a statistical viewpoint these properties are represented by a set of features, $X_1, \ldots, X_j, \ldots, X_p$, observed on n objects. The clustering task is subjected to several sources of uncertainty concerning: (i) sampling variability, (ii) number and shape of clusters, (iii) assignment of objects to clusters, (iv) imprecision/vagueness of observed features. In the following sources (iii) and (iv) will be specifically dealt with. A remarkable advance in coping with source (iii) has been achieved by the Fuzzy *k*-Means (F*k*M) procedure (Bezdek, 1981), whose objective function is

$$J_{\mathrm{F}k\mathrm{M}} = \sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^{m} d^2 \left(\mathbf{x}_i, \mathbf{h}_g \right), \tag{1}$$

where u_{ig} represents the degree of membership (sharing) of object i in cluster g (ranging from 0 to 1), $m > 1$ is a "fuzziness coefficient" tuning the amount of fuzziness in the classification structure (for $m \to 1$ no fuzziness is allowed; fuzziness increases with m), $\mathbf{h}_g$ is the prototype of cluster g and $d^2(\cdot, \cdot)$ is an appropriate metric on the features' space. In this context, a cluster is viewed as a fuzzy set (Zadeh, 1965), defined on the "universe" of the n objects. In this approach the constraint $\sum_{g=1}^{k} u_{ig} = 1, \forall i$ is considered when estimating the degrees of membership of an object in the various clusters. It has been observed (Krishnapuram and Keller, 1993, 1996) that this involves the interpretation of the results of F*k*M in the light of a (fuzzy) *partition* of the set of objects, rather than of a *typology* underlying the observed data.

* Corresponding author. Tel.: +39 06 49910731; fax: +39 06 4959241.
E-mail addresses: renato.coppi@uniroma1.it (R. Coppi), pierpaolo.durso@uniroma1.it (P. D'Urso), paolo.giordani@uniroma1.it (P. Giordani).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.09.013

An alternative way of managing the uncertainty associated with source (iii) is through Possibility Theory (see, e.g. Dubois and Prade, 1988). Its use in the clustering problem allows us to get rid of the probabilistic constraint on the u_{ig}'s. In fact, in the possibilistic perspective, u_{ig} represents the degree of membership (possibility) of object i belonging to cluster g or, in other terms, the degree of "compatibility" of the profile $\mathbf{x}_i$ with the characteristics of cluster g embodied by its prototype $\mathbf{h}_g$. The FkM objective function is consequently modified by introducing an additive "penalization" ($P_{\mathrm{P}k\mathrm{M}}$) term which takes care of the balance between the fuzziness of the classification structure and the "compactness" of the clusters:

$$J_{\mathrm{P}k\mathrm{M}} = J_{\mathrm{F}k\mathrm{M}} + P_{\mathrm{P}k\mathrm{M}} = J_{\mathrm{F}k\mathrm{M}} + G[\varphi\left(\mathbf{U}\right); \psi\left(\boldsymbol{\vartheta}\right)], \tag{2}$$

where $\mathbf{U}$ is the $n \times k$ matrix of the compatibility degrees, G, φ and ψ appropriate functions and $\boldsymbol{\vartheta}$ a vector of tuning parameters.

Krishnapuram and Keller (1996) argue that the possibilistic approach provides a "mode-seeking" clustering procedure, to be compared with the "partition-seeking" property of FkM. Thus, PkM clustering methods tend to be more robust with respect to noise as compared to FkM techniques. Nonetheless, there are some limitations in the use of PkM algorithms, in that these may lead to trivial solutions consisting of "coincident clusters" (Barni et al., 1996). This tendency may be dealt with by means of an appropriate initialization of the computational procedure (see Section 4).

A specification of the general function (2) has been given by Yang and Wu (2006) in the following way:

$$J_{\mathrm{P}k\mathrm{M}} = J_{\mathrm{F}k\mathrm{M}} + \frac{\beta}{m^2\sqrt{k}} \sum_{i=1}^{n} \sum_{g=1}^{k} \left(u_{ig}^m \log u_{ig}^m - u_{ig}^m\right), \tag{3}$$

where $\varphi\left(\mathbf{U}\right) = \sum_{i=1}^{n} \sum_{g=1}^{k} \left(u_{ig}^m \log u_{ig}^m - u_{ig}^m\right)$ and $\psi\left(\boldsymbol{\vartheta}\right) = \psi\left(\beta, m, k\right) = \frac{\beta}{m^2\sqrt{k}}$ is a tuning parametric function allowing the control of cluster validity along with the separation of the data set (measured by β). Moreover, it tunes the fuzziness of the clusters represented by the degrees of compatibility whose values are affected by m (for m large, the data points tend to have small degrees of compatibility even with clusters whose prototypes are very close to them). In the following we will adopt the formulation of the possibilistic objective function suggested by Yang and Wu (2006).

However, the role of source (iv) of uncertainty (imprecision/vagueness of empirical data) was not considered until the paper by Sato and Sato (1995), in which a fuzzy clustering technique for fuzzy data is proposed. The crucial points, in this extension of the domain of uncertainty directly managed by the clustering methodology, are as follows. In the first place, it is required to appropriately define the uncertainty associated with the measurement of the X_j's, when these features are imprecisely or vaguely observed. The theory of fuzzy sets provides us with powerful tools for modelling this uncertainty (see Section 2). In the second place, a suitable distance (or dissimilarity measure) on the space of (imprecise/vague) features has to be introduced, along with a coherent way of characterizing the fuzzy clusters (e.g. by means of fuzzy prototypes).

Several authors have recently contributed to this generalization (see, e.g. Hathaway et al., 1996; Yang and Ko, 1996; Yang and Liu, 1999; Auephanwiriyakul and Keller, 2002; D'Urso and Giordani, 2006). In many of these proposals a parametric class of membership functions, namely the LR family (see Section 2), is utilized, within an appropriate extension of the FkM objective function to fuzzy data. In this context, D'Urso and Giordani (2006) suggest a fuzzy clustering method for symmetric LR_1 fuzzy data, based on an FkM criterion using a weighted dissimilarity measure which takes into account the different importance of the centers and spreads of the observed fuzzy variables. For a survey on fuzzy clustering for fuzzy data see D'Urso (2007).

In the present paper, both a fuzzy and a possibilistic approach to clustering fuzzy data are proposed, with reference to the general family of LR_2 fuzzy variables. In the possibilistic case, the penalization term in (3) is adopted in this connection. The paper is organized as follows. In Section 2 the various types of fuzzy data are described. Section 3 is devoted to a generalization of the fuzzy clustering technique for fuzzy data introduced by D'Urso and Giordani (2006). In Section 4 a possibilistic model for clustering LR fuzzy data is introduced, together with the solution equations for estimating its parameters and the respective algorithm. A simulation experiment is carried out in Section 5 and two applications to real world data are illustrated in Section 6. Finally, some concluding remarks, along with a hint to future research in this domain, are made in Section 7.

2. Fuzzy data

In statistics, we usually process crisp data, typically exact results of measurements and/or of observations. However, in many real-life situations, measurements may be imprecise and the observations vaguely defined. Moreover, in several fields of knowledge (such as social sciences or bio-medicine), both scientific propositions and empirical data are often formulated in terms of natural language. These formulations may be appropriately represented by fuzzy numbers. In a matrix form, a general class of fuzzy data, called *LR fuzzy data*, can be defined as follows:

$$\tilde{\mathbf{X}} \equiv \left\{\tilde{x}_{ij} = (c_{1ij}, c_{2ij}, l_{ij}, r_{ij})_{LR} : i = 1, \ldots, n; j = 1, \ldots, p\right\} \quad (\textit{fuzzy data matrix}), \tag{4}$$

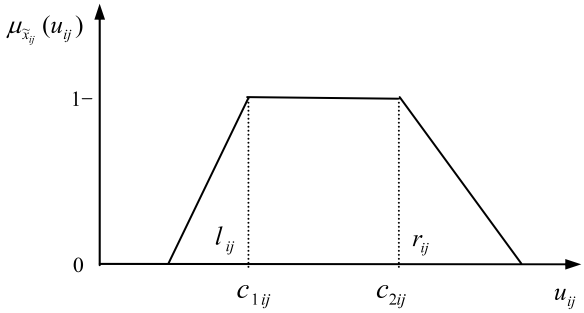

Fig. 1. Trapezoidal membership function.

where $\tilde{x}_{ij} = (c_{1ij}, c_{2ij}, l_{ij}, r_{ij})_{LR}$ represents the *LR* fuzzy variable *j* observed on the *i*-th object, c_{1ij} and c_{2ij} denote the left and right center, respectively, and l_{ij} and r_{ij} the left and right spread, respectively, with the following membership function:

$$\mu_{\tilde{x}_{ij}}(u_{ij}) = \begin{cases} L\left(\dfrac{c_{1ij} - u_{ij}}{l_{ij}}\right) & u_{ij} \leq c_{1ij} \ (l_{ij} > 0), \\ 1 & c_{1ij} \leq u_{ij} \leq c_{2ij}, \\ R\left(\dfrac{u_{ij} - c_{2ij}}{r_{ij}}\right) & u_{ij} \geq c_{2ij} \ (r_{ij} > 0), \end{cases} \tag{5}$$

where $L(z_{ij})$ (and $R(z_{ij})$) is a decreasing 'shape' function from $\Re^+$ to [0, 1] with $L(0) = 1$; $L(z_{ij}) < 1$ for all $z_{ij} > 0, \forall i, j$; $L(z_{ij}) > 0$ for all $z_{ij} < 1, \forall i, j$; $L(1) = 0$ (or $L(z_{ij}) > 0$ for all $z_{ij}, \forall i, j$, and $L(+\infty) = 0$) (Zimmermann, 2001). A particular case of *LR* fuzzy data is the trapezoidal one where $L\left(\frac{c_{1ij}-u_{ij}}{l_{ij}}\right) = 1 - \frac{c_{1ij}-u_{ij}}{l_{ij}}$ and $R\left(\frac{u_{ij}-c_{2ij}}{r_{ij}}\right) = 1 - \frac{u_{ij}-c_{2ij}}{r_{ij}}$ (see Fig. 1).

In the literature, several metrics have been generalized for fuzzy data. For instance, in Trutschnig et al. (2009) a discussion about the expression of L_2-metrics for fuzzy sets in terms of centers and spreads is provided. In the following, we suggest a weighted dissimilarity measure for fuzzy data to be utilized in the clustering models illustrated in Sections 3 and 4. The dissimilarity between each pair of objects is measured by comparing the fuzzy data observed on each object, i.e. by considering, separately, the distances for the centers and the spreads of the fuzzy data and using a suitable weighting system for such distance components. Thus, by considering the *i*-th and *i*′-th objects, we have

$$d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'}) = w_C^2[d^2(\mathbf{c}_{1i}, \mathbf{c}_{1i'}) + d^2(\mathbf{c}_{2i}, \mathbf{c}_{2i'})] + w_S^2[d^2(\mathbf{l}_i, \mathbf{l}_{i'}) + d^2(\mathbf{r}_i, \mathbf{r}_{i'})], \tag{6}$$

where:

$d(\mathbf{c}_{1i}, \mathbf{c}_{1i'}) = \|\mathbf{c}_{1i} - \mathbf{c}_{1i'}\|$ = Euclidean distance between the left centers $\mathbf{c}_{1i}$ and $\mathbf{c}_{1i'}$;
$d(\mathbf{c}_{2i}, \mathbf{c}_{2i'}) = \|\mathbf{c}_{2i} - \mathbf{c}_{2i'}\|$ = Euclidean distance between the right centers $\mathbf{c}_{2i}$ and $\mathbf{c}_{2i'}$;
$d(\mathbf{l}_i, \mathbf{l}_{i'}) = \|\mathbf{l}_i - \mathbf{l}_{i'}\|$ = Euclidean distance between the left spreads $\mathbf{l}_i$ and $\mathbf{l}_{i'}$;
$d(\mathbf{r}_i, \mathbf{r}_{i'}) = \|\mathbf{r}_i - \mathbf{r}_{i'}\|$ = Euclidean distance between the right spreads $\mathbf{r}_i$ and $\mathbf{r}_{i'}$;
$\mathbf{c}_{1i} \equiv (c_{1i1}, \ldots, c_{1ij}, \ldots, c_{1ip})'$, $\quad \mathbf{c}_{1i'} \equiv (c_{1i'1}, \ldots, c_{1i'j}, \ldots, c_{1i'p})'$,
$\mathbf{c}_{2i} \equiv (c_{2i1}, \ldots, c_{2ij} \ldots, c_{2ip})'$, $\quad \mathbf{c}_{2i'} \equiv (c_{2i'1}, \ldots, c_{2i'j}, \ldots, c_{2i'p})'$,
$\mathbf{l}_i \equiv (l_{i1}, \ldots, l_{ij}, \ldots, l_{ip})'$, $\quad \mathbf{l}_{i'} \equiv (l_{i'1}, \ldots, l_{i'j}, \ldots, l_{i'p})'$,
$\mathbf{r}_i \equiv (r_{i1}, \ldots, r_{ij}, \ldots, r_{ip})'$, $\quad \mathbf{r}_{i'} \equiv (r_{i'1}, \ldots, r_{i'j}, \ldots, r_{i'p})'$;

$w_C, w_S \geq 0$ are suitable weights for the center component and the spread component of $d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'})$, where $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_{i'}$ denote the fuzzy data vectors, respectively, for the *i*-th and *i*′-th objects, i.e. $\tilde{\mathbf{x}}_i \equiv \left\{\tilde{x}_{ij} = (c_{1ij}, c_{2ij}, l_{ij}, r_{ij})_{LR} : j = 1, \ldots, p\right\}$ and $\tilde{\mathbf{x}}_{i'} \equiv \left\{\tilde{x}_{i'j} = (c_{1i'j}, c_{2i'j}, l_{i'j}, r_{i'j})_{LR} : j = 1, \ldots, p\right\}$. The weights $w_C, w_S \geq 0$ can be stored in the two-dimensional vector $\mathbf{w} \equiv (w_C \ w_S)'$. Since the membership function value of the centers is maximum, we propose to assume that the (left and right) center distances' weight is higher than (or at least equal to) the (left and right) spread distances' one. Then, we assume the following conditions: $w_C + w_S = 1$ (*normalization condition*) and $w_C \geq w_S \geq 0$ (*coherence condition*). As we can see from (6), we assume that the weights for the left and right center distances and the left and right spreads distances are, respectively, the same.

3. Fuzzy clustering for fuzzy data

Let us consider the fuzzy data matrix (4) and the dissimilarity measure for fuzzy data (6). We assume that the weights in the dissimilarity (6) are objectively computed during the clustering process. In order to classify objects with fuzzy information, in a fuzzy framework, D'Urso and Giordani (2006) propose a fuzzy clustering procedure limiting their attention to the special class of the so-called symmetric LR_1 fuzzy data (i.e. $\tilde{x}_{ij} = (c_{ij}, s_{ij})_{LR}$ where $c_{ij} = c_{1ij} = c_{2ij}$ and $s_{ij} = l_{ij} = r_{ij}$) by using a dissimilarity measure which resembles (6). On the basis of their findings, we are going to introduce a new fuzzy clustering technique capable of managing the general class of LR_2 fuzzy data given in (4) and (5) by means of (6). This technique, called *Fuzzy k-means clustering model for fuzzy data* (FkM-F), can be formalized as follows:

$$\begin{aligned}
\min_{u_{ig},\mathbf{h}_g,\mathbf{w}} : J_{\text{FkM-F}} &\equiv \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m d_F^2(\tilde{\mathbf{x}}_i, \mathbf{h}_g) \\
&= \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m \left[w_C^2 \left[d^2(\mathbf{c}_{1i}, \mathbf{h}_g^{C_1}) + d^2(\mathbf{c}_{2i}, \mathbf{h}_g^{C_2}) \right] + w_S^2 \left[d^2(\mathbf{l}_i, \mathbf{h}_g^{L}) + d^2(\mathbf{r}_i, \mathbf{h}_g^{R}) \right] \right], \qquad (7)\\
\text{s.t.} \quad & u_{ig} \in [0,1], \qquad \sum_{g=1} u_{ig} = 1, \\
\mathbf{w} \equiv (w_C\ w_S)' \geq \mathbf{0}_2, &\quad w_C \geq w_S,\ w_C + w_S = 1,
\end{aligned}$$

where:

$m > 1$ is a weighting exponent that controls the fuzziness of the obtained partition;

u_{ig} indicates the membership degree of the i-th object in the g-th cluster;

$d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_g)$ represents the suggested dissimilarity measure (6) between the i-th object and the prototype of the g-th cluster; analogously for its components $d^2(\mathbf{c}_{1i}, \mathbf{h}_g^{C_1})$, $d^2(\mathbf{c}_{2i}, \mathbf{h}_g^{C_2})$, $d^2(\mathbf{l}_i, \mathbf{h}_g^{L})$, $d^2(\mathbf{r}_i, \mathbf{h}_g^{R})$, where the fuzzy vector $\tilde{\mathbf{h}}_g \equiv \left\{ \tilde{h}_{gj} = (h_g^{C_1}, h_g^{C_2}, h_g^{L}, h_g^{R})_{LR} : j = 1, \ldots, p \right\}$ represents the fuzzy prototype of the g-th cluster, $\mathbf{h}_g^{C_1} \equiv (h_{g1}^{C_1}, \ldots, h_{gj}^{C_1}, \ldots, h_{gp}^{C_1})'$, $\mathbf{h}_g^{C_2} \equiv (h_{g1}^{C_2}, \ldots, h_{gj}^{C_2}, \ldots, h_{gp}^{C_2})'$, $\mathbf{h}_g^{L} \equiv (h_{g1}^{L}, \ldots, h_{gj}^{L}, \ldots, h_{gp}^{L})'$, $\mathbf{h}_g^{R} \equiv (h_{g1}^{R}, \ldots, h_{gj}^{R}, \ldots, h_{gp}^{R})'$ are p-vectors, whose j-th element refers to the j-th variable, that denote, respectively, the (left and right) centers and the (left and right) spreads of the g-th fuzzy prototype. Note that, in the symmetric LR_1 fuzzy data case, (7) reduces to the proposal by D'Urso and Giordani (2006).

By solving the constrained quadratic minimization problem (7) via the Lagrangian multiplier method with respect to u_{ig} and by setting the first derivatives of $J_{\text{FkM-F}}$ with respect to $\mathbf{h}_g^{C_1}$, $\mathbf{h}_g^{C_2}$, $\mathbf{h}_g^{L}$, $\mathbf{h}_g^{R}$ and $\mathbf{w}$ equal to zero we obtain the following *iterative solution*:

$$u_{ig} = \frac{\left[w_C^2 \left[d^2(\mathbf{c}_{1i}, \mathbf{h}_g^{C_1}) + d^2(\mathbf{c}_{2i}, \mathbf{h}_g^{C_2}) \right] + w_S^2 \left[d^2(\mathbf{l}_i, \mathbf{h}_g^{L}) + d^2(\mathbf{r}_i, \mathbf{h}_g^{R}) \right] \right]^{-\frac{1}{m-1}}}{\sum_{g'=1}^{k} \left[w_C^2 \left[d^2(\mathbf{c}_{1i}, \mathbf{h}_{g'}^{C_1}) + d^2(\mathbf{c}_{2i}, \mathbf{h}_{g'}^{C_2}) \right] + w_S^2 \left[d^2(\mathbf{l}_i, \mathbf{h}_{g'}^{L}) + d^2(\mathbf{r}_i, \mathbf{h}_{g'}^{R}) \right] \right]^{-\frac{1}{m-1}}}, \qquad (8)$$

$$\mathbf{h}_g^{C_1} = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{c}_{1i}}{\sum_{i=1}^{n} u_{ig}^m}, \qquad \mathbf{h}_g^{C_2} = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{c}_{2i}}{\sum_{i=1}^{n} u_{ig}^m}, \qquad \mathbf{h}_g^{L} = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{l}_i}{\sum_{i=1}^{n} u_{ig}^m}, \qquad \mathbf{h}_g^{R} = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{r}_i}{\sum_{i=1}^{n} u_{ig}^m} \qquad (9)$$

and

$$w_C = \frac{\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m \left[d^2(\mathbf{l}_i, \mathbf{h}_g^{L}) + d^2(\mathbf{r}_i, \mathbf{h}_g^{R}) \right]}{\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m \left[d^2(\mathbf{c}_{1i}, \mathbf{h}_g^{C_1}) + d^2(\mathbf{c}_{2i}, \mathbf{h}_g^{C_2}) + d^2(\mathbf{l}_i, \mathbf{h}_g^{L}) + d^2(\mathbf{r}_i, \mathbf{h}_g^{R}) \right]} \quad (w_S = 1 - w_C). \qquad (10)$$

By observing (10) we can conclude that the normalization condition is implicitly satisfied. To take into account the coherence condition, it can be shown that (7) is a parabola with respect to w_S. In fact, (10) coincides with the abscissa of its vertex. When we get $w_S > 0.5$, the solution in (10) is unfeasible. However, among the feasible solutions, it is easy to see that the minimum of (7) with respect to w_S is attained when $w_S = 0.5$ (and $w_C = 0.5$). See, for further details, D'Urso and Giordani (2006).

Algorithm FkM-F ($\tilde{\mathbf{X}}, m, k$)

Step 0a. Generate randomly the membership degree matrix $\mathbf{U}^{(0)}$ subject to (7).

Step 0b. Compute the prototypes $\tilde{\mathbf{H}}^{(0)}$ according to (9) using $\mathbf{U}^{(0)}$.

Step 1. Update the weights $w_C^{(t)}$ and $w_S^{(t)}$ according to (10) keeping fixed $\mathbf{U}^{(t-1)}$ and $\tilde{\mathbf{H}}^{(t-1)}$, where $t \geq 1$ denotes the iteration number, and set $w_C^{(t)} = w_S^{(t)} = 0.5$ if $w_S^{(t)} > 0.5$.

Step 2. Update the prototypes $\tilde{\mathbf{H}}^{(t)}$ according to (9) keeping fixed $\mathbf{U}^{(t-1)}$.
Step 3. Update the membership degree matrix $\mathbf{U}^{(t)}$ according to (8) keeping fixed $\tilde{\mathbf{H}}^{(t)}$ and $w_C^{(t)}$ and $w_S^{(t)}$.
Step 4. If $\left\|\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)}\right\| < \varepsilon$, where ε is a non-negative small number fixed in advance, the algorithm has converged, otherwise go to *Step* 1.

4. Possibilistic clustering for fuzzy data

For classifying objects with fuzzy information, following a possibilistic approach, we suggest the *Possibilistic k-means clustering model for fuzzy data* (P*k*M-F):

$$
\begin{aligned}
\min_{u_{ig},\mathbf{h}_g,\mathbf{w}} : J_{\text{P}k\text{M-F}} &\equiv J_{\text{F}k\text{M-F}} + P_{\text{P}k\text{M-F}} \\
&= \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_g) + \frac{\beta}{m^2\sqrt{k}} \sum_{i=1}^{n}\sum_{g=1}^{k}\left(u_{ig}^m \log u_{ig}^m - u_{ig}^m\right) \\
&= \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m \left[w_C^2 \left[d^2(\mathbf{c}_{1i}, \mathbf{h}_g^{C_1}) + d^2(\mathbf{c}_{2i}, \mathbf{h}_g^{C_2}) \right] \right. \\
&\quad \left. + \; w_S^2 \left[d^2(\mathbf{l}_i, \mathbf{h}_g^{L}) + d^2(\mathbf{r}_i, \mathbf{h}_g^{R}) \right] \right] + \frac{\beta}{m^2\sqrt{k}} \sum_{i=1}^{n}\sum_{g=1}^{k}\left(u_{ig}^m \log u_{ig}^m - u_{ig}^m\right),
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
&\text{s.t.} \quad u_{ig} \in [0, 1], \\
&\mathbf{w} \equiv (w_C \; w_S)' \geq \mathbf{0}_2, \quad w_C \geq w_S, \; w_C + w_S = 1,
\end{aligned}
$$

where $\beta, m, k > 0$.

The objective function $J_{\text{P}k\text{M-F}}$ in (11) is constituted by two terms.

- $J_{\text{F}k\text{M-F}} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_g)$: this term is equivalent to the objective function of the F*k*M-F model which requires the dissimilarity $d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_g)$ to be as small as possible.
- $P_{\text{P}k\text{M-F}} = \frac{\beta}{m^2\sqrt{k}} \sum_{i=1}^{n}\sum_{g=1}^{k}\left(u_{ig}^m \log u_{ig}^m - u_{ig}^m\right)$: this is the *penalization term*; it is constructed by an analogous of the *partition entropy PE* (Bezdek, 1974), i.e. $\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m \log u_{ig}^m \approx$ partition entropy, and an analogous of the *partition coefficient PC* (Bezdek, 1981), i.e. $\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^m \approx$ partition coefficient. The penalization term should try to force u_{ig} to be as large as possible.

Notice that in the objective function $J_{\text{P}k\text{M-F}}$ there also appears the term $\frac{\beta}{m^2\sqrt{k}}$, where β is a *normalization term* that measures the degree of separation of the dataset, and it is reasonably defined as a sort of variance coefficient, i.e. $\beta = \frac{\sum_{i=1}^{n}\left(\|\mathbf{c}_{1i}-\bar{\mathbf{c}}_1\|^2+\|\mathbf{c}_{2i}-\bar{\mathbf{c}}_2\|^2+\left\|\mathbf{l}_i-\bar{\mathbf{l}}\right\|^2+\|\mathbf{r}_i-\bar{\mathbf{r}}\|^2\right)}{4n}$, in which $\bar{\mathbf{c}}_1 = \frac{\sum_{i=1}^{n}\mathbf{c}_{1i}}{n}$, $\bar{\mathbf{c}}_2 = \frac{\sum_{i=1}^{n}\mathbf{c}_{2i}}{n}$, $\bar{\mathbf{l}} = \frac{\sum_{i=1}^{n}\mathbf{l}_i}{n}$, $\bar{\mathbf{r}} = \frac{\sum_{i=1}^{n}\mathbf{r}_i}{n}$; m is the weighting exponent that controls the fuzziness of the obtained partition (notice that $m > 0$, but, in general, according to Yang and Wu (2006), we suggest to set $m = 2$); $\sqrt{k}$ is a parameter used to control the steepness of the membership functions. For an in depth discussion of the reason why in the penalization term $P_{\text{F}k\text{M-F}}$ the coefficient $\frac{\beta}{m^2\sqrt{k}}$ appears, see Yang and Wu (2006).

It can be shown that the optimal prototypes and weights can be found using (9) and (10), respectively, and the optimal values of $\mathbf{U}$ are given by

$$
u_{ig} = \exp\left(-\frac{m\sqrt{k}\left[w_C^2 \left[d^2(\mathbf{c}_{1i}, \mathbf{h}_g^{C_1}) + d^2(\mathbf{c}_{2i}, \mathbf{h}_g^{C_2}) \right] + w_S^2 \left[d^2(\mathbf{l}_i, \mathbf{h}_g^{L}) + d^2(\mathbf{r}_i, \mathbf{h}_g^{R}) \right] \right]}{\beta} \right). \tag{12}
$$

Notice also that, following the same reasoning of F*k*M-F, if $w_S > 0.5$, then the update of w_C and w_S is $w_C = w_S = 0.5$.

As it is well-known (see, e.g Barni et al., 1996; Timm et al., 2004; Tjhi and Chen, 2007), an undesirable property of the *k*-means algorithms based on the possibilistic approach is the tendency to produce coincident clusters. In other words, the algorithms may determine all the prototypes in the same location. This can be explained by noticing that the possibilistic loss functions can usually be decomposed into the sum of *k* terms (one for every cluster), which can be minimized independently of each other. This depends on the absence of the probabilistic constraint on the membership degrees ($\sum_{g=1}^{k} u_{ig} = 1, \forall i$). In this case, when there exists a single optimal point for a cluster prototype, all the prototypes move over there. A practical hint that is often adopted (see, e.g Krishnapuram and Keller, 1996), is to use the solution from F*k*M as starting point for P*k*M. In a similar way, we recommend using the F*k*M-F solution as the starting point for the P*k*M-F algorithm.

Table 1
Results of the simulation experiment.

Design variable	${}^{FkM\text{-}F}SSQ$	${}^{PkM\text{-}F}SSQ$	${}^{FkM\text{-}F}u_{cluster}$	${}^{PkM\text{-}F}u_{cluster}$	${}^{FkM\text{-}F}u_{outlier}$	${}^{PkM\text{-}F}u_{outlier}$	${}^{FkM\text{-}F}w_S$	${}^{PkM\text{-}F}w_S$
$n = 25$	0.0078	0.0016	50.46	44.44	0.65	0.11	0.44	0.42
$n = 49$	0.0022	0.0010	50.35	43.73	0.64	0.11	0.45	0.44
L	0.0024	0.0007	50.00	49.77	0.67	0.09	0.48	0.48
S	0.0084	0.0557	51.22	32.48	0.59	0.12	0.37	0.33
L&S	0.0025	0.0006	50.00	50.00	0.69	0.13	0.48	0.48
WS	0.0078	0.0002	50.46	49.61	0.65	0.13	0.42	0.43
PS	0.0025	0.0025	50.35	38.56	0.64	0.09	0.46	0.43
OL	0.0028	0.0011	50.02	44.32	0.58	0.01	0.50	0.47
OS	0.0064	0.0018	51.20	43.51	0.83	0.32	0.33	0.35
OL&S	0.0027	0.0011	50.00	44.43	0.53	0.00	0.50	0.48
Overall	0.0030	0.0012	50.41	44.09	0.65	0.11	0.44	0.43

5. Simulation experiment

In this section we report the results of a simulation experiment carried out in order to assess the capabilities and the weaknesses of F*k*M-F and P*k*M-F. The most relevant question to be answered by means of the simulation study is whether the clusters resulting from P*k*M-F are less sensitive to outliers than those from F*k*M-F. It is worth mentioning that the robust property of P*k*M-F can be proved following the same strategy adopted by Yang and Wu (2006). Specifically, it is straightforward to show that the cluster prototypes in (9) with the membership degrees in (11) are particular M-estimators (see Huber, 1991). Therefore, the primary aim of the simulation experiment is to assess whether such a theoretical property of P*k*M-F holds in practice. Furthermore, we aim at studying whether the use of the F*k*M-F solution as starting point for P*k*M-F limits the risk of obtaining coincident clusters and whether F*k*M-F and P*k*M-F are able to tune the influence of the center and spread distances by means of w_C and w_S.

The set-up of the simulation study can be summarized as follows. We randomly generate LR_2 fuzzy data sets such that a structure with $k = 2$ clusters is known and the presence of an outlier is visible. The randomly generated data sets have two different data sizes with the number of objects equal to $n = 25$ or $n = 49$ and the number of variables equal to $p = 2$. First, the left and right centers of all the objects are generated randomly from $U[-\theta_{loc}, 0.5 - \theta_{loc}]$ and $U[0.5 - \theta_{loc}, 1 - \theta_{loc}]$, respectively, whereas the left and right spreads from $U[0, 0.5]$. The parameter θ_{loc} helps us to distinguish the clusters with respect to the location information, as it will be clarified below. To distinguish the cluster memberships, we operate as follows. In the *L* case, we add the value $2\theta_{loc}$ to the left and right centers of all the fuzzy variables for the objects from number $(n-1)/2 + 1$ to $n - 1$. Therefore, in this case, we construct two clusters of size $(n-1)/2$ distinguished with respect to the left and right centers, i.e. the *locations* of the objects. In the *S* case, we add the value θ_{size} to the left and right spreads of all the fuzzy variables for the objects from number $(n-1)/2 + 1$ to $n - 1$ (also the role of θ_{size} will be discussed in the following). This allows us to distinguish the clusters by considering the spreads of the fuzzy data, i.e. the *sizes* of the objects. Finally, in the *L&S* case, the clusters are distinguished by both the *locations* and *sizes* of the objects. This is done by adding $2\theta_{loc}$ to the left and right centers and θ_{size} to the left and the right spreads of the objects from number $(n-1)/2 + 1$to $n - 1$. The parameters θ_{loc} and θ_{size} are chosen in such a way that the clusters are *well* or *partially* separated (cases *WS* and *PS*, respectively). In the former case we set $\theta_{loc} = \theta_{size} = 1$, in the latter one $\theta_{loc} = \theta_{size} = 0.5$. Moreover, we randomly generate an outlier (object n) according to three different procedures. In the *OL* (*outlier by location*) case, the outlier is characterized by an anomalous location. This is done by adding $2\theta_{loc}$ to the left and right centers of variable 2 for object n. In the *OS* (*outlier by size*) case, we add θ_{size} to the left and right spreads of variable 2 for object n. Finally, in the *OL&S* (*outlier by location & size*) case, object n is an outlier due to the left and right centers of variable 2 (added $2\theta_{loc}$) and the left and right spreads (added θ_{size}). For every level of every design variable, ten data sets are randomly generated. Therefore, 2 (data sizes) × 3 (cluster structures) × (levels of separation between clusters) × 3 (outlier types) × 10 (replications) = 360 fuzzy data sets are randomly generated during the simulation experiment. To limit the risk of hitting a local optimum, F*k*M-F is run using five random starts and then considering the solution pertaining to the run of the algorithm that leads to the smallest function value. P*k*M-F is run using the F*k*M-F output (corresponding to the smallest function value) as the (rational) starting point.

The results of the simulation experiment are given in Table 1, which contains some indices reporting the performances of F*k*M-F and P*k*M-F distinguished with respect to every level of every design variable (first column). Columns 2 and 3 contain the median values of the sum of squares between the prototypes known in advance and the ones estimated by F*k*M-F and P*k*M-F, respectively. More specifically, apart from object n, we know in advance that the first $(n-1)/2$ objects belong to a cluster and the latter $(n-1)/2$ to a different cluster. This allows us to compute the prototypes of these two clusters as the mean of the assigned objects, say ${}^K\tilde{\mathbf{h}}_g \equiv \left({}^K\mathbf{h}_g^{C_1}, {}^K\mathbf{h}_g^{C_2}, {}^K\mathbf{h}_g^{L}, {}^K\mathbf{h}_g^{R}\right)$, $g = 1, 2$, with, for instance,

$$ {}^K\mathbf{h}_1^{C_1} = \left[\sum_{i=1}^{(n-1)/2} c_{1i1} \Big/ (n-1)/2 \quad \sum_{i=1}^{(n-1)/2} c_{1i2} \Big/ (n-1)/2 \right] \quad \text{and} $$

$$ {}^{K}\mathbf{h}_2^{C_1} = \left[\sum_{i=(n-1)/2+1}^{n-1} c_{1i1} \Big/ (n-1)/2 \quad \sum_{i=(n-1)/2+1}^{n-1} c_{1i2} \Big/ (n-1)/2 \right], $$

where c_{1i1} and c_{1i2} denote the randomly generated left centers for object i. Then, we can evaluate how much the estimated prototypes (say ${}^{\text{F}k\text{M-F}}\tilde{\mathbf{h}}_g$ and ${}^{\text{P}k\text{M-F}}\tilde{\mathbf{h}}_g$) differ from ${}^{K}\tilde{\mathbf{h}}_g$ by means of ${}^{\text{F}k\text{M-F}}SSQ$ and ${}^{\text{P}k\text{M-F}}SSQ$. With respect to F$k$M-F, assuming without loss of generality that ${}^{K}\tilde{\mathbf{h}}_g$ is estimated by ${}^{\text{F}k\text{M-F}}\tilde{\mathbf{h}}_g$, $g = 1, 2$, it is

$$ {}^{\text{F}k\text{M-F}}SSQ = \sum_{g=1}^{2} \left(\left\| {}^{K}\mathbf{h}_g^{C_1} - {}^{\text{F}k\text{M-F}}\mathbf{h}_g^{C_1} \right\|^2 + \left\| {}^{K}\mathbf{h}_g^{C_1} - {}^{\text{F}k\text{M-F}}\mathbf{h}_g^{C_1} \right\|^2 + \left\| {}^{K}\mathbf{h}_g^{L} - {}^{\text{F}k\text{M-F}}\mathbf{h}_g^{L} \right\|^2 + \left\| {}^{K}\mathbf{h}_g^{R} - {}^{\text{F}k\text{M-F}}\mathbf{h}_g^{R} \right\|^2 \right). $$

Similarly, we can derive ${}^{\text{P}k\text{M-F}}SSQ$. Note that we decide to compute the median values instead of the mean values because the distributions of the SSQ's are skew and a few anomalous values occur. The next two columns of Table 1 report the average percentage of non-outlier objects assigned to the cluster with the biggest size. We assume that the method assigns an object to a cluster if the membership degree is higher than 0.50 (hard clustering). Apart from the outlier, every simulated data set can be split up into two halves by assigning the first $(n-1)/2$ objects to one cluster and the latter $(n-1)/2$ to the other cluster. Thus, we can conclude that FkM-F and PkM-F work well if ${}^{\text{F}k\text{M-F}}u_{cluster}$ and ${}^{\text{P}k\text{M-F}}u_{cluster}$ are equal to 50. The farther from 50 ${}^{\text{F}k\text{M-F}}u_{cluster}$ (or ${}^{\text{P}k\text{M-F}}u_{cluster}$) is, the worse F$k$M-F (or P$k$M-F) detects the clusters in the hard clustering sense. The next two columns of Table 1 contain the average maximum membership degree of the outlier and the last two the average weight w_S.

By observing the columns of Table 1 concerning the median SSQ scores we can see that ${}^{\text{F}k\text{M-F}}SSQ > {}^{\text{P}k\text{M-F}}SSQ$. It holds for all the levels of the design variables except for the S case. Also this occurs considering the overall median values (${}^{\text{F}k\text{M-F}}SSQ = 0.0030$ and ${}^{\text{P}k\text{M-F}}SSQ = 0.0012$). It follows that P$k$M-F provided more accurate estimates of the prototypes if compared with those resulting from FkM-F. This can be explained by taking into account the robustness property of PkM-F, which tends to give lower membership degrees to the objects far away from the prototypes.

The analysis of columns 4 and 5 highlights that FkM-F works well in detecting the two clusters. In fact, the overall mean ${}^{\text{F}k\text{M-F}}u_{cluster}$ value is very close to 50 (50.41). It is interesting to see that when the clusters are distinguished with respect to the locations and the sizes of the objects (L&S case), FkM-F and PkM-F worked perfectly in the sense that they always were able to detect the cluster structure (mean ${}^{\text{F}k\text{M-F}}u_{cluster}$ = mean${}^{\text{P}k\text{M-F}}u_{cluster} = 50.00$). Furthermore, it is worth mentioning that we never observed the coincident cluster problem.

To investigate how FkM-F and PkM-F managed outliers, we can inspect the average ${}^{\text{F}k\text{M-F}}u_{outlier}$ and ${}^{\text{P}k\text{M-F}}u_{outlier}$ values given in columns 6 and 7 of Table 1. We see that the FkM-F wrongly tended to assign the outlier to a cluster with an average membership degree equal to 0.65. Instead, we found out that PkM-F was able to detect the outliers by not assigning them to any cluster in the hard clustering sense (the overall average ${}^{\text{P}k\text{M-F}}u_{outlier}$ value was equal to 0.11).

All in all, these results show the usefulness of FkM-F and PkM-F for clustering fuzzy data. Both methods seem to suffer to a certain extent from (unusual) situations in which the clusters are distinguished mainly on the basis of the spread information. The use of the FkM-F output as starting point for the PkM-F algorithm appears to be very helpful since it prevents the risk of obtaining coincident clusters. Furthermore, PkM-F seems to work remarkably well in the presence of outliers.

6. Applications

In order to show how FkM-F and PkM-F work in practice, we are going to consider two applications. The former involves an existing data set known in the literature (see Yang et al., 2004). The latter concerns data we collected concerning the perceptions and opinions of a group of students about the current global economic and financial crisis.

6.1. Cars data

This data set pertains to the measurements on $p = 3$ features with respect to $n = 10$ cars. Two of them (Comfort and Safety) are of LR_2 fuzzy type, whereas crisp scores are observed for the variable Price. The data are given in Table 2.

Since the variables have different units of measurement, we decide to preprocess the data. To do it, we operate as follows. The left and right centers are standardized using the mean and the standard deviation of the (left and right) centers' values of each variable. After that, each left and right spread is divided by the standard deviation of the corresponding centers. This way of preprocessing the data helps us to eliminate unwanted differences among the variables, without losing relevant information concerning the widths of the fuzzy data. Prior analyses showed that the cars can be well distinguished into $k = 2$ clusters. Tables 3 and 4 contain, respectively, the prototypes (after applying the inverse preprocessing procedure) and membership degrees from FkM-F setting $m = 2$ (we get $w_C = w_S = 0.5$).

By observing Tables 3 and 4 we can see that six cars are assigned to Cluster 1, which can be labeled as 'high quality cars' and four to Cluster 2 ('low quality cars'). The labels of the clusters can be explained by noting that the features of the prototype of Cluster 1 are better than those of the prototype of Cluster 2. Specifically, cars belonging to Cluster 1 are better in terms of Comfort and Safety and more expensive than those belonging to Cluster 2. It is important to stress the flexibility of FkM-F, which is capable of managing jointly precise (Price) and imprecise (Comfort and Safety) information. In fact, the

Table 2
Cars data (data take the form (left center, right center, left spread, right spread)).

Object	Price (NT$10000)	Comfort	Safety
Virage (China-Motor)	(63.9, 63.9, 0.0, 0.0)	(10.0, 10.0, 2.0, 2.0)	(9.0, 9.0, 3.0, 3.0)
New Lancer (China-Motor)	(51.9, 51.9, 0.0, 0.0)	(6.0, 6.0, 2.0, 2.0)	(6.0, 6.0, 3.0, 3.0)
Galant (China-Motor)	(71.8, 71.8, 0.0, 0.0)	(10.0, 14.0, 2.0, 0.0)	(12.5, 17.5, 3.0, 0.0)
Tierra Activa (Ford)	(46.9, 46.9, 0.0, 0.0)	(6.0, 6.0, 2.0, 2.0)	(6.0, 6.0, 3.0, 3.0)
M2000 (Ford)	(64.6, 64.6, 0.0, 0.0)	(8.0, 8.0, 2.0, 2.0)	(9.0, 9.0, 3.0, 3.0)
Tercel (Toyota)	(45.8, 45.8, 0.0, 0.0)	(2.0, 6.0, 0.0, 2.0)	(6.0, 6.0, 3.0, 3.0)
Corolla (Toyota)	(74.3, 74.3, 0.0, 0.0)	(10.0, 14.0, 2.0, 0.0)	(12.0, 12.0, 3.0, 3.0)
Premio G2.0 (Toyota)	(72.9, 72.9, 0.0, 0.0)	(10.0, 10.0, 2.0, 2.0)	(12.5, 17.5, 3.0, 0.0)
Cerfiro (Yulon-Motor)	(69.9, 69.9, 0.0, 0.0)	(8.0, 8.0, 2.0, 2.0)	(12.0, 12.0, 3.0, 3.0)
March (Yulon-Motor)	(39.9, 39.9, 0.0, 0.0)	(2.0, 6.0, 0.0, 2.0)	(0.5, 5.5, 0.0, 3.0)

Table 3
Prototypes from F*k*M-F (data take the form (left center, right center, left spread, right spread)).

Cluster	Price (NT$10000)	Comfort	Safety
Cluster 1	(70.1, 70.1, 0.0, 0.0)	(9.4, 10.9, 2.0, 1.3)	(11.4, 13.2, 3.0, 1.9)
Cluster 2	(46.8, 46.8, 0.0, 0.0)	(4.2, 6.1, 1.0, 2.0)	(4.9, 6.0, 2.3, 3.0)

Table 4
Membership degrees and *SSQ*'s from F*k*M-F.

Object	Cluster 1	Cluster 2	$SSQ(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_1)$	$SSQ(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_2)$
Virage (China-Motor)	0.83	0.17	2.15	10.70
New Lancer (China-Motor)	0.06	0.94	13.04	0.90
Galant (China-Motor)	0.92	0.08	2.60	30.63
Tierra Activa (Ford)	0.03	0.97	15.82	0.55
M2000 (Ford)	0.73	0.27	2.98	7.85
Tercel (Toyota)	0.03	0.97	21.44	0.72
Corolla (Toyota)	0.94	0.06	1.63	25.82
Premio G2.0 (Toyota)	0.94	0.06	1.63	26.06
Cerfiro (Yulon-Motor)	0.92	0.08	1.31	14.27
March (Yulon-Motor)	0.08	0.92	31.96	2.69

structure of the prototypes is inherited by the observed data. Therefore, the prototypes are of LR_2 fuzzy type. Nevertheless, since the scores of the variable Price are crisp numbers, the features of the prototypes for such a variable are degenerate LR_2 fuzzy numbers with the same left and right centers and zero spreads, i.e. crisp numbers.

All the membership degrees of the objects clearly assigned (i.e. with membership degrees > 0.50) to Cluster 2 are very high (>0.90) and the same holds for those objects clearly assigned to Cluster 1 except for Virage (China-Motor) and M2000 (Ford) belonging to Cluster 1 with membership degrees equal to 0.83 and 0.73, respectively. The last two columns of Table 4 contain the sum of squares (*SSQ*) between object i ($\tilde{\mathbf{x}}_i \equiv (\mathbf{c}_{1i}, \mathbf{c}_{2i}, \mathbf{l}_i, \mathbf{r}_i)$, $i = 1, n$) and prototype g ($\tilde{\mathbf{h}}_g \equiv (\mathbf{h}_g^{C_1}, \mathbf{h}_g^{C_2}, \mathbf{h}_g^{L}, \mathbf{h}_g^{R})$, $g = 1, k$), namely $SSQ(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_g) = \left\|\mathbf{c}_{1i} - \mathbf{h}_g^{C_1}\right\|^2 + \left\|\mathbf{c}_{2i} - \mathbf{h}_g^{C_2}\right\|^2 + \left\|\mathbf{l}_i - \mathbf{h}_g^{L}\right\|^2 + \left\|\mathbf{r}_i - \mathbf{h}_g^{R}\right\|^2$. It would be advisable to obtain low *SSQ* values between objects clearly assigned to a given cluster and the corresponding prototypes. For instance, with respect to Cluster 2, this is the case for New Lancer (China-Motor), Tierra Activa (Ford) and Tercel (Toyota), the *SSQ* values of which are the lowest ones (ranging from 0.55 to 0.90) and all the membership degrees are higher than 0.94. A possible outlier is March (Yulon-Motor). In fact, despite its high membership degree in Cluster 2 (0.92), the corresponding *SSQ* value is sensibly higher than the previous ones (2.69). By observing its original features, we can see that March (Yulon-Motor) is the worst car since it is characterized by the lowest values of price, comfort and safety. It follows that its features are different with respect to the features of the Cluster 2 prototype and much more different with respect to those of the Cluster 1 prototype (the corresponding *SSQ* value is 31.96). Thus, its high membership degree in Cluster 2 mainly depends on its distance from Cluster 2 rather than on its closeness to Cluster 1. Similar comments could be drawn with respect to some cars belonging to Cluster 1. However, to corroborate our claims and to check whether further outliers can be found, we run P*k*M-F (using the results in Tables 3 and 4 as starting point). Setting $m = 2$ (and $k = 2$) the solution (not reported here) is too fuzzy with very low membership degrees even for cars which are expected to be well assigned to a given cluster. We thus reduce the fuzziness parameter and decide to consider as optimal the solution with $m = 1.5$. Such a solution can be found in Tables 5 and 6.

By comparing Tables 4 and 6, we can see that the clusters' labels remain the same even if the differences between the two clusters are less evident. Among the four objects not clearly assigned to any cluster (namely, the possible outliers), the analysis of Table 6 shows that some of them (Corolla (Toyota), Premio G2.0 (Toyota) and Galant (China-Motor)) are related to some limited extent to Cluster 1 but they are extremely high quality cars and much more expensive than the fictitious one resulting from the Cluster 1 prototype. Conversely, March (Yulon-Motor) slightly belongs to Cluster 2 (with membership

Table 5
Prototypes from P*k*M-F (data take the form (left center, right center, left spread, right spread)).

Cluster	Price (NT\$10000)	Comfort	Safety
Cluster 1	(66.2, 66.2, 0.0, 0.0)	(8.7, 8.9, 2.0, 1.9)	(10.0, 10.2, 3.0, 2.9)
Cluster 2	(48.3, 48.6, 0.0, 0.0)	(5.1, 6.1, 1.0, 1.5)	(5.9, 6.1, 2.9, 3.0)

Table 6
Membership degrees and *SSQ*'s from P*k*M-F.

Object	Cluster 1	Cluster 2	$SSQ(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_1)$	$SSQ(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_2)$
Virage (China-Motor)	0.84	0.05	0.51	8.36
New Lancer (China-Motor)	0.11	0.91	6.34	0.25
Galant (China-Motor)	0.06	0.00	7.07	27.50
Tierra Activa (Ford)	0.04	0.95	8.60	0.15
M2000 (Ford)	0.90	0.13	0.30	5.85
Tercel (Toyota)	0.01	0.62	13.53	1.35
Corolla (Toyota)	0.20	0.00	4.59	22.64
Premio G2.0 (Toyota)	0.18	0.00	4.90	22.83
Cerfiro (Yulon-Motor)	0.77	0.02	0.75	11.65
March (Yulon-Motor)	0.00	0.21	22.36	4.45

degree equal to 0.21) because, as already noticed, it is a low quality car which is remarkably cheaper and characterized by lower levels of Comfort and Safety.

6.2. Students data

This section is devoted to an application to real data characterized by imprecision. In fact, we asked a set of undergraduate students attending the Course of Statistics at the Faculty of Political Sciences of Sapienza University of Rome in the academic year 2009–2010 to fill in a questionnaire. The items of the questionnaire (see Appendix) refer to the opinions and perceptions about the recent global economic and financial crisis. The first question (Q1) concerns whether the crisis is due to financial speculation. The next questions concern the opinion about the utility of a new regularization of financial markets (Q2) and the feeling about the need for a drastic change of the economic system (Q3). Question Q4 refers to the opinion on the adequacy of the recent EU economic measures to face up to the crisis. Finally, the remaining questions deal with the Italian situation. In question Q5, opinions about the Italian economic measures are asked. The last question (Q6) concerns the perception about the trend of the Italian economy during the next three years.

The responses to each of these questions are collected by means of a Visual Analogue Scale (see, e.g. Huskisson, 1983), inviting every responder to fill in a vertical mark along a segment of length one joining the two opposite extreme opinions, according to the position where the responder feels she (he) should locate her- (him-) self. The data are fuzzified assuming that a certain degree of imprecision is associated to every opinion/perception. In particular, we consider *LR* fuzzy numbers with center ($c_1 = c_2$) equal to the distance between the left bound and the filled mark and spread given by

$$l = r = \begin{cases} 0.125 & 0.125 \le c \le 0.875, \\ c & c < 0.125, \\ 1 - c & c > 0.875. \end{cases}$$

This is based upon the assumption that when one looks at the unitary segment and decides to transform her (his) opinion into a specific location along the segment, she (he) intuitively splits the segment into two halves and then each half into two halves in order to better approximate her (his) assessment. This underlying mechanism suggests associating to any given location a fuzzy interval of length 0.25. Hence, the choice of spreads is equal to 0.125, except for extreme locations where the imprecision naturally decreases to the extent to which the sign gets closer to the bounds. See also Coppi et al. (2006). The data (concerning $n = 27$ students and $p = 6$ variables) were collected on the 25th of May 2010. The recorded data are given in Table 7, where only the centers of the fuzzy data are reported.

We start by running F*k*M-F on the raw data with $k = 3$ and $m = 2$. We obtain the prototypes and membership degrees reported in Tables 8 and 9, respectively. Also note that we find $w_C = w_S = 0.5$.

The analysis of Table 8 highlights that three well interpretable clusters are obtained. The students belonging to Cluster 2 are worried about the crisis and are characterized by a very pessimistic point of view. In fact, they do not think that reconsidering the economic and financial rules will provide a solution to the crisis and that the only possibility is to fully reorganize the economic system. As a consequence, they do not think that the EU measures as well as the Italian ones play a fruitful role in order to stimulate the economic growth. All in all, we can refer to this cluster as 'the Left', since this type of attitude is more typical of parties and associations located in the left wing area. The students assigned to the remaining clusters share opinions more close to the ideas supported by the Center and the Right parties. In particular, Cluster 3 can be denoted as 'the Right' since there is a general agreement that the crisis is not a relevant problem and the measures taken by the EU and, especially, by the Italian government are helpful. Finally Cluster 1 may be called 'the Center-Right' because

Table 7
Students data (only the centers of the fuzzy data are reported).

Student (n.)	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
1	1.00	0.00	1.00	0.52	0.00	0.00
2	1.00	0.03	0.96	0.17	0.03	0.03
3	0.63	0.60	0.50	0.48	0.75	0.46
4	0.77	0.68	0.69	0.72	0.47	0.38
5	0.92	0.98	0.40	0.47	0.31	0.08
6	0.93	0.30	1.00	0.05	0.00	0.00
7	0.85	0.00	1.00	0.00	0.08	0.00
8	0.98	0.86	0.73	0.46	0.32	0.48
9	0.68	0.28	0.50	0.70	0.82	0.75
10	1.00	0.00	1.00	0.00	0.00	0.00
11	0.51	0.50	0.51	0.52	0.49	0.49
12	0.51	0.65	0.53	0.43	0.38	0.47
13	0.80	0.81	0.93	0.54	0.00	0.47
14	0.95	0.82	0.75	0.63	0.03	0.38
15	0.99	0.67	0.52	0.48	0.28	0.47
16	0.88	0.60	0.90	0.69	0.20	0.58
17	0.85	0.82	0.62	0.84	0.28	0.20
18	1.00	0.74	0.72	0.53	0.40	0.30
19	0.67	0.40	0.54	0.36	0.52	0.38
20	0.58	0.43	0.53	0.27	0.54	0.73
21	0.57	0.39	0.33	0.45	0.68	0.71
22	1.00	0.16	0.99	0.54	0.00	0.00
23	0.03	0.97	0.95	0.52	0.03	0.47
24	0.76	0.06	0.85	0.24	0.01	0.02
25	1.00	0.18	0.99	0.01	0.00	0.18
26	0.99	0.43	0.10	0.57	0.24	0.41
27	1.00	0.00	1.00	0.18	0.22	0.03

Table 8
Prototypes from F*k*M-F (data take the form (center, spread)).

Cluster	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
Cluster 1	(0.88, 0.06)	(0.76, 0.11)	(0.69, 0.11)	(0.58, 0.12)	(0.25, 0.10)	(0.38, 0.12)
Cluster 2	(0.94, 0.04)	(0.09, 0.05)	(0.97, 0.02)	(0.18, 0.07)	(0.04, 0.03)	(0.04, 0.03)
Cluster 3	(0.61, 0.12)	(0.50, 0.12)	(0.50, 0.12)	(0.47, 0.12)	(0.54, 0.12)	(0.54, 0.12)

Table 9
Membership degrees and *SSQ*'s from F*k*M-F.

Student (n.)	Cluster 1	Cluster 2	Cluster 3	$SSQ\,(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_1)$	$SSQ\,(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_2)$	$SSQ\,(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_3)$
1	0.12	0.79	0.09	1.90	0.28	2.61
2	0.01	0.98	0.01	1.97	0.02	2.47
3	0.13	0.04	0.84	0.80	2.78	0.12
4	0.64	0.05	0.31	0.18	2.19	0.38
5	0.62	0.12	0.26	0.51	2.61	1.24
6	0.08	0.87	0.05	1.68	0.15	2.40
7	0.05	0.90	0.05	2.49	0.14	2.68
8	0.84	0.04	0.12	0.11	2.08	0.77
9	0.20	0.09	0.70	1.57	3.51	0.45
10	0.04	0.92	0.03	2.60	0.12	3.08
11	0.05	0.01	0.93	0.63	2.27	0.04
12	0.22	0.04	0.74	0.46	2.21	0.13
13	0.71	0.12	0.17	0.30	1.78	1.28
14	0.85	0.06	0.09	0.13	1.87	1.22
15	0.74	0.06	0.21	0.15	1.83	0.52
16	0.69	0.10	0.21	0.25	1.78	0.82
17	0.75	0.07	0.17	0.23	2.46	1.00
18	0.82	0.05	0.13	0.11	1.70	0.71
19	0.13	0.05	0.81	0.65	1.55	0.10
20	0.13	0.06	0.82	1.06	2.44	0.16
21	0.11	0.05	0.84	1.36	3.23	0.18
22	0.16	0.74	0.10	1.42	0.31	2.30
23	0.43	0.20	0.37	1.82	3.85	2.12
24	0.08	0.86	0.07	1.71	0.15	1.93
25	0.08	0.86	0.06	1.81	0.17	2.36
26	0.41	0.16	0.44	0.95	2.48	0.88
27	0.05	0.90	0.05	2.02	0.12	2.29

Table 10
Prototypes from P*k*M-F (data take the form (center, spread)).

Cluster	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
Cluster 1	(0.90, 0.06)	(0.74, 0.12)	(0.68, 0.12)	(0.57, 0.12)	(0.31, 0.11)	(0.40, 0.12)
Cluster 2	(0.95, 0.04)	(0.07, 0.04)	(0.98, 0.02)	(0.12, 0.07)	(0.05, 0.03)	(0.03, 0.03)
Cluster 3	(0.61, 0.12)	(0.52, 0.12)	(0.52, 0.12)	(0.46, 0.12)	(0.53, 0.12)	(0.50, 0.12)

Table 11
Membership degrees and *SSQ*'s from P*k*M-F.

Student (n.)	Cluster 1	Cluster 2	Cluster 3	$SSQ\,(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_1)$	$SSQ\,(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_2)$	$SSQ\,(\tilde{\mathbf{x}}_i, \tilde{\mathbf{h}}_3)$
1	0.00	0.32	0.00	1.95	0.35	2.53
2	0.00	0.93	0.00	1.99	0.02	2.37
3	0.12	0.00	0.69	0.66	2.93	0.12
4	0.63	0.00	0.35	0.15	2.38	0.33
5	0.18	0.00	0.03	0.53	2.77	1.11
6	0.00	0.62	0.00	1.74	0.15	2.26
7	0.00	0.72	0.00	2.50	0.10	2.57
8	0.75	0.00	0.11	0.09	2.23	0.69
9	0.01	0.00	0.17	1.38	3.68	0.54
10	0.00	0.79	0.00	2.62	0.07	2.97
11	0.16	0.00	0.90	0.56	2.42	0.03
12	0.25	0.00	0.73	0.43	2.35	0.10
13	0.28	0.00	0.02	0.40	1.95	1.18
14	0.51	0.00	0.03	0.21	2.05	1.12
15	0.70	0.00	0.22	0.11	1.97	0.48
16	0.44	0.00	0.08	0.26	1.96	0.78
17	0.43	0.00	0.06	0.26	2.69	0.90
18	0.79	0.00	0.13	0.07	1.84	0.62
19	0.17	0.00	0.76	0.55	1.66	0.09
20	0.05	0.00	0.53	0.94	2.54	0.20
21	0.02	0.00	0.45	1.20	3.36	0.25
22	0.01	0.28	0.00	1.48	0.40	2.19
23	0.00	0.00	0.00	1.97	4.05	2.01
24	0.00	0.56	0.00	1.75	0.18	1.84
25	0.00	0.62	0.00	1.83	0.15	2.25
26	0.06	0.00	0.06	0.89	2.63	0.89
27	0.00	0.70	0.00	1.99	0.11	2.20

students belonging to this cluster are quite worried about the crisis and think that the EU is working quite well and, in any case, better than the Italian government.

In the hard clustering sense, we can see that twenty-five students are well assigned to a cluster (with membership degree higher than 0.50). Specifically, nine students belong to Cluster 1, nine to Cluster 2 and seven to Cluster 3. Nonetheless, in some cases, such membership degrees do not show a strong assignment to a given cluster (see, for instance, students n. 4 and n. 5). Finally, at least two outliers (students n. 23 and n. 26) can be found. These students are assigned to the clusters with membership degrees always lower than 0.50. By inspecting the *SSQ* values, we can conclude that some assignments are questionable in the sense that some students well assigned to a given cluster (for instance student n. 9 belonging to Cluster 3 with a membership degree equal to 0.70) seem to be quite far from the corresponding prototypes (and, indeed, from all the prototypes). In this connection, relevant information can be gathered by considering P*k*M-F. In Tables 10 and 11 we report the P*k*M-F solution (we find $w_C = w_S = 0.5$) obtained starting from the previously described F*k*M-F solution and setting $k = 3$ and $m = 1.5$.

By comparing Tables 8 and 10 we can see that the clusters' interpretation does not differ. Remarkable differences can be found inspecting the membership degree matrices in Tables 9 and 11. We can see that P*k*M-F tends to clearly assign to the clusters a lower number of students. Several students cannot be clearly assigned to any cluster. Therefore, such students can be considered outliers. In this respect, P*k*M-F detects them giving low membership degrees in the clusters. The most relevant case is given by student n. 23 whose degrees of compatibility are equal to zero. By observing Table 7, we can see that her (his) features are remarkably different if compared with all the three prototypes.

7. Conclusions

In this paper, we have proposed some clustering models for fuzzy LR_2 data following the fuzzy and possibilistic approaches. Such models allow us to detect k homogeneous clusters on the basis of n objects described by p fuzzy variables. To characterize every cluster, a fictitious object, i.e. the prototype, has been computed. A crucial assumption of our clustering models is that the prototypes are of LR_2 fuzzy type, inheriting their typology by the observed data. Generally speaking, the prototypes are obtained as a weighted mean of the observed objects using the membership degree information as system of weights. Although every membership degree can range in the unit interval in both the approaches, their meaning remarkably differs, as outlined in the introduction. In order to extend the existing fuzzy and possibilistic clustering models to the fuzzy data case, two ingredients have played a fundamental role. First of all, the comparison between fuzzy data has been done by introducing a new dissimilarity measure. The second ingredient has concerned how to admit possibilistic degrees of membership. This was done by introducing the penalization term proposed by Yang and Wu (2006) which can be split up into two parts resembling, respectively, the partition entropy and partition coefficient (see, e.g. Bezdek, 1981). The possibilistic clustering model for fuzzy data (P*k*M-F) has been developed by using such a penalization term and by replacing the standard Euclidean distance (for crisp data) by the proposed weighted dissimilarity measure.

A nice property of the possibilistic approach to clustering and, indeed, of P*k*M-F is that the solution appears to be more immune to noise since the membership degrees correspond to the notion of compatibility. However, in the literature, there exist several ways to manage the clustering problem in the presence of anomalous data. For instance, we can adopt the metric approach (see, e.g. Hathaway et al., 2000) in which metrics with robust properties are incorporated in the objective functions of the clustering problem or the noise approach (see, e.g. Rehm et al., 2007) in which the outliers are assigned to the so-called noise cluster. Another interesting line of research is given by the evidential approach proposed by Masson and Denœux (2008). This has led to the development of several new clustering techniques that claim to be robust. Unfortunately, as far as we know, all of these techniques deal with crisp data. Generalizations to the fuzzy data case still remain to be done. The only exception can be found in Hung and Yang (2005). Their clustering algorithm, based on an exponential-type distance function, belongs to the metric approach but is limited to the univariate case ($p = 1$). Therefore, the proposed P*k*M-F clustering model represents the first attempt of robust clustering for fuzzy data. In the near future, it will be interesting to extend some existing techniques carried out according to a different approach (metric, noise, evidential) to the fuzzy data case and compare them with P*k*M-F. At the same time, it is worth mentioning that the possibilistic approach is the only robust clustering procedure allowing us to express the membership degrees as degrees of compatibility. The noise and metric approaches lead to membership degrees interpreted as degrees of sharing (as in the fuzzy approach). On the contrary, the evidential approach is related to the concept of belief functions.

In order to avoid the coincident clusters problem it has been suggested to use the output of F*k*M-F as the starting point of P*k*M-F. To evaluate whether this remedy works, a simulation study has been done. It has been found that this limits the risk of obtaining coincident clusters. Nonetheless, it is not possible to state that coincident clusters can always be avoided using such a trick. To this purpose, it will be interesting to derive new possibilistic models such that this undesired property will never occur. In this respect, a promising tool seems to be the inclusion of clustering repulsion terms in the objective function of the clustering problem as proposed in Timm et al. (2004) and Rehm et al. (2010). Such a term forces the prototypes to be the farthest away possible. A suitable extension of the above cited works to the fuzzy data case may hopefully represent a solution to the coincident clusters problem.

Acknowledgements

We would like to express our gratitude to the editor and referees for their helpful comments, which remarkably improved the quality of the paper.

Appendix. Questionnaire

Please make your assessment on the following statements:

Q1. The current global economic crisis has been caused by financial speculation.

Q2. In order to avoid the repetition of crises of this type it is sufficient to regulate the financial markets.

Q3. The current global economic crisis requires a radical change of the economic system.

Q4. How do you evaluate the recent economic measures taken by the EU for coping with the Greek financial crisis and possible similar crises of other European countries (i.e. loan to Greece, European solidarity fund)?

Q5. What is your degree of satisfaction with respect to the measures taken by the Italian government to cope with the current economic crisis?

Q6. What is, in your opinion, the trend of the Italian economy in the next three years, according to the current economic policies of the Italian government?

References

Auephanwiriyakul, S., Keller, J.M., 2002. Analysis and efficient implementation of a linguistic fuzzy c-means. IEEE Transactions on Fuzzy Systems 10, 563–582.
Barni, M., Cappellini, V., Mecocci, A., 1996. Comments on 'a possibilistic approach to clustering'. IEEE Transactions on Fuzzy Systems 4, 393–396.
Bezdek, J.C., 1974. Cluster validity with fuzzy sets. Journal of Cybernetics 3, 58–73.
Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum Press.
Coppi, R., Giordani, P., D'Urso, P., 2006. Component models for fuzzy data. Psychometrika 71, 733–761.
D'Urso, P., 2007. Clustering of fuzzy data. In: de Oliveira, J.V., Pedrycz, W. (Eds.), Advances in Fuzzy Clustering and Its Applications. J. Wiley and Sons, pp. 155–192.
D'Urso, P., Giordani, P., 2006. A weighted fuzzy c-means clustering model for fuzzy data. Computational Statistics and Data Analysis 50, 1496–1523.
Dubois, D., Prade, H., 1988. Possibility Theory. Plenum Press, New-York.
Hathaway, R.J., Bezdek, J.C., Hu, Y., 2000. Generalized fuzzy c-means clustering strategies using L_p norm distances. IEEE Transactions on Fuzzy Systems 8, 576–582.
Hathaway, R.J., Bezdek, J.C., Pedrycz, W., 1996. A parametric model for fusing heterogeneous fuzzy data. IEEE Transactions on Fuzzy Systems 4, 1277–1282.
Huber, P.J., 1991. Robust Statistics. Wiley, New York.
Hung, W.-L., Yang, M.-S., 2005. Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation. Fuzzy Sets and Systems 150, 561–577.
Huskisson, E.C., 1983. Visual analogue scales. In: Melzack, R. (Ed.), Pain Measurement and Assessment. Raven, New York, pp. 33–37.
Krishnapuram, R., Keller, J., 1993. A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems 1, 98–110.
Krishnapuram, R., Keller, J., 1996. The possibilistic c-means algorithm: insights and recommendations. IEEE Transactions on Fuzzy Systems 4, 385–393.
Masson, M.-H., Denœux, T., 2008. ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recognition 41, 1384–1397.
Rehm, F., Klawonn, F., Kruse, R., 2007. A novel approach to noise clustering for outlier detection. Soft Computing 11, 489–494.
Rehm, F., Winkler, R., Kruse, R., 2010. Fuzzy clustering with repulsive prototypes. In: Laurent, A., Lesot, M.-J. (Eds.), Scalable Fuzzy Algorithms for Data Management and Analysis: Methods and Design. Information Science Reference, New York, pp. 332–345.
Sato, M., Sato, Y., 1995. Fuzzy clustering model for fuzzy data. Proceedings IEEE 2123–2128.
Timm, H., Borgelt, C., Doring, C., Kruse, R., 2004. An extension to possibilistic fuzzy cluster analysis. Fuzzy Sets and Systems 147, 3–16.
Tjhi, W.-C., Chen, L., 2007. Possibilistic co-clustering of large document collections. Pattern Recognition 40, 3452–3466.
Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A., 2009. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences 179, 3964–3972.
Yang, M.-S., Hwang, P.Y., Chen, D.H., 2004. Fuzzy clustering algorithms for mixed feature variables. Fuzzy Sets and Systems 141, 301–317.
Yang, M.S., Ko, C.H., 1996. On a class of fuzzy c-numbers clustering procedures for fuzzy data. Fuzzy Sets and Systems 84, 49–60.
Yang, M.S., Liu, H.H., 1999. Fuzzy clustering procedures for conical fuzzy vector data. Fuzzy Sets and Systems 106, 189–200.
Yang, M.-S., Wu, K.-L., 2006. Unsupervised possibilistic clustering. Pattern Recognition 39, 5–21.
Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.
Zimmermann, H.J., 2001. Fuzzy Set Theory and its Applications. Kluwer, Boston.