On possibilistic clustering with repulsion constraints for imprecise data

Maria Brigida Ferraro, Paolo Giordani

Abstract

In possibilistic clustering the objects are assigned to clusters according to the so-called membership degrees taking values in the unit interval. Differently from fuzzy clustering, it is not required that the sum of the membership degrees of an object in all the clusters is equal to one. This is very helpful in the presence of outliers, which are usually assigned to the clusters with membership degrees close to zero. Unfortunately, a drawback of the possibilistic approach is the tendency to produce coincident clusters. A remedy is represented by the use of a repulsion term among prototypes in the loss function forcing the prototypes to be 'enough' far from each other. Here, a possibilistic clustering model with repulsion constraints for imprecise data, managed in term of fuzzy sets, is introduced. Two applications to synthetic and real fuzzy data are considered in order to analyze how the proposed clustering model works in practice.

Keywords: Cluster analysis, possibilistic approach, repulsion term, fuzzy data.

1 Introduction

In cluster analysis we aim at determining a small number k of groups (clusters) from a set of $n \gg k$ objects. Every group is composed by objects homogeneous according to a given dissimilarity measure based on p observed features $X_1, \ldots, X_p$. For this purpose, a well-known tool is the Fuzzy k-

Means algorithm (FkM) [2]:

$$\min_{\mathbf{U},\mathbf{H}} J_{FkM} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2\left(\mathbf{x}_i, \mathbf{h}_g\right), \tag{1a}$$

$$\text{s.t.} \quad u_{ig} \in [0,1], i = 1,\ldots,n, g = 1,\ldots,k, \tag{1b}$$

$$\sum_{i=1}^{n} u_{ig} = 1, g = 1,\ldots,k, \tag{1c}$$

where u_{ig} is the membership degree of object i in cluster g and $\mathbf{h}_g = [h_{g1},\ldots,h_{gp}]$ is the so-called prototype for cluster g, that is, a p-vector identifying each cluster and is the g-th row of the prototype matrix $\mathbf{H}$ of order $(k \times p)$. The membership degrees u_{ig} are stored in the matrix $\mathbf{U}$ of order $(n \times k)$ and express the extent to which an object belongs to a cluster. The closer u_{ig} is to 1, the higher is the membership of object i in cluster g. In fact, the objects can be assigned to more than one cluster (with different membership degrees), but the sum of the membership degrees of an object in all the clusters must be equal to 1 (fuzzy approach to clustering). Finally, $d^2\left(\mathbf{x}_i, \mathbf{h}_g\right)$ is the dissimilarity measure (usually the squared Euclidean distance) between object i and prototype g and $m > 1$ is a fuzziness coefficient tuning the amount of fuzziness in the solution.
The optimal parameter matrices $\mathbf{H}$ and $\mathbf{U}$ can be found iteratively updating the membership degrees keeping fixed the prototypes and viceversa. The update formulas are

$$u_{ig} = \frac{1}{\sum_{g'=1}^{k} \left(\frac{d^2(\mathbf{x}_i, \mathbf{h}_g)}{d^2\left(\mathbf{x}_i, \mathbf{h}_{g'}\right)} \right)^{\frac{1}{m-1}}} \tag{2}$$

and, if $d^2\left(\mathbf{x}_i, \mathbf{h}_g\right) = \sum_{j=1}^{p} \left(x_{ij} - h_{gj}\right)^2$,

$$h_{gj} = \frac{\sum_{i=1}^{n} u_{ig}^{m} x_{ij}}{\sum_{i=1}^{n} u_{ig}^{m}}, g = 1,\ldots,p. \tag{3}$$

Although FkM is a very powerful tool, the constraints in (1c) may lead to anomalous results. In particular, it may occur that an object is assigned to a cluster even if it is far from the corresponding prototype. By inspecting (2) this can be easily explained noting that the generic u_{ig} is obtained comparing

the dissimilarity between object i and prototype g to the sum of the dissimilarities between object i and all the prototypes. This suggests to interpret the membership degrees in FkM as degrees of *sharing* of object i in cluster g.

The constraints of FkM are such that possible outliers, i.e. objects "very far" from all the cluster prototypes, can be assigned to a cluster with a high membership degree when their distance from the corresponding prototype is high, but remarkably lower than those from the remaining $k-1$ prototypes. In the literature, to overcome this drawback several proposals have been introduced. These can be seen as robust versions of FkM. In particular, at least four approaches can be found. These are the metric approach (see, e.g., [9]) in which a suitable metric for handling outliers is considered in the loss function, the noise approach (see, e.g, [5]) in which outliers are forced to belong to an additional cluster, called noise cluster, with high membership degrees (of course, such a cluster is not formed by homogeneous objects), the evidential approach (see, e.g., [13]) in which objects belong to several subsets of classes and the possibilistic approach in which outliers tend to have low membership degrees in all the clusters. The latter approach shall be adopted in this paper.

The Possibilistic k-Means algorithm (PkM) in the version of Krishnapuram and Keller [11] can be formulated as

$$\min_{\mathbf{U},\mathbf{H}} J_{PkM} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2(\mathbf{x}_i, \mathbf{h}_g) + \sum_{g=1}^{k} \eta_g \sum_{i=1}^{n} (1-u_{ig})^m, \tag{4a}$$

$$\text{s.t.} \quad u_{ig} \in [0,1], i = 1,\ldots,n, \forall g = 1,\ldots,k, \tag{4b}$$

where η_g is a tuning parameter associated with cluster g, weighting its contribution to the penalization function.

By comparing (1a)-(1c) and (4a)-(4b) we can see that the loss in (4a) contains an additional term, which plays the role of avoiding the trivial solution with $u_{ig} = 0, i = 1,\ldots,n, g = 1,\ldots,k$. Furthermore, the row-wise sum of $\mathbf{U}$ is no longer required to be equal to 1.

As for FkM, the optimal solution of PkM can be found iteratively. The update of the prototypes can be done using (3) and that of the membership degrees is given by

$$u_{ig} = \frac{1}{1 + \left(\frac{d_{ig}^2}{\eta_g}\right)^{\frac{1}{m-1}}}. \tag{5}$$

In contrast with (2) the membership degree u_{ig} in (5) is based only on the distance between object i and prototype g allowing us to interpret such a degree as a measure of the *compatibility* of an object with respect to a cluster. Several extensions of PkM have been proposed in the literature. Refer, for instance, to [14, 16, 19]. Unfortunately, as pointed out by Barni et al. [1], the possibilistic approach to clustering suffers from the so-called coincident cluster problem, that is, the obtained solution is formed by one cluster. This is so because the minimization problem in (4a) boils down to the sum of k minimization problems (one for each cluster) that can be minimized independently of each other. A heuristic remedy to this problem suggested by Krishnapuram and Keller [12] is to consider the FkM solution as (rational) starting point of the iterative algorithm. A more powerful remedy is represented by the use of repulsion terms [15, 16, 17] which force the prototypes to be the farthest away possible. In this work we suggest a possibilistic clustering algorithm with repulsion constraints for imprecise data exploiting the proposal by Timm et al. [16]. We chose this proposal because, differently from the other ones, it involves a loss function to be minimized incorporating a repulsion term and the optimal prototypes are found by suitably taking it into account. Instead, in [15, 17], the repulsion term is inserted in the update of the prototypes, but it is not clear what is the loss to be minimized.
The paper is organized as follows. In the next section the PkM proposed by Timm et al. [16] is recalled. In doing so, we suggest an improvement for updating the prototypes based on the Newton algorithm since a non-linear problem must be solved. In Section 3 fuzzy data and metrics for such a kind of data are introduced. In fact, we handle imprecision in terms of fuzzy data. Section 4 deals with PkM with repulsion constraints for fuzzy data. Finally some applications and concluding remarks are given in Sections 5 and 6, respectively.

2 Possibilistic clustering with repulsion constraints

The PkM with Repulsion constraints (PkM-R) [16] can be formalized as

$$\begin{aligned}\min_{\mathbf{U},\mathbf{H}} J_{PkM-R} &= \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2\left(\mathbf{x}_i, \mathbf{h}_g\right) + \sum_{g=1}^{k}\eta_g \sum_{i=1}^{n}\left(1-u_{ig}\right)^m \\ &\quad + \sum_{g=1}^{k}\gamma_g \sum_{g'=1, g'\neq g}^{k} \frac{1}{\xi d^2\left(\mathbf{h}_g, \mathbf{h}_{g'}\right)} \qquad (6a)\\ \text{s.t.} \quad & u_{ig} \in [0,1], i = 1,\ldots,n, \forall g = 1,\ldots,k, \qquad (6b)\end{aligned}$$

where η_g, ξ and γ_g are non-negative parameters to be chosen in advance. In particular, η_g is a cluster-specific parameter tuning the importance of the clusters and, following [16], can be defined as

$$\eta_g = \eta \frac{\sum_{i=1}^{n} u_{ig}^m d_{ig}^2}{d^2\left(\mathbf{x}_i, \mathbf{h}_g\right)}, \qquad (7)$$

where $\mathbf{U}$ and $\mathbf{H}$ are found by standard FkM and, usually, $\eta = 1$. γ_g is a cluster-specific weighting factor, which can be set as

$$\gamma_g = \gamma \sum_{i=1}^{n} u_{ig}^m, \qquad (8)$$

where, once again, the membership degrees are found by FkM and γ increases the importance of the repulsion term in the optimization problem. Finally, as clarified in [16], the parameter ξ depends on the minimal distance we are willing to accept between neighboring clusters. However, we can observe that

$$\sum_{g=1}^{k}\gamma_g \sum_{g'=1, g'\neq g}^{k} \frac{1}{\xi d^2\left(\mathbf{h}_g, \mathbf{h}_{g'}\right)} = \sum_{g=1}^{k}\frac{\gamma_g}{\xi} \sum_{g'=1, g'\neq g}^{k} \frac{1}{d^2\left(\mathbf{h}_g, \mathbf{h}_{g'}\right)} = \sum_{g=1}^{k}\gamma'_g \sum_{g'=1, g'\neq g}^{k} \frac{1}{d^2\left(\mathbf{h}_g, \mathbf{h}_{g'}\right)}, \qquad (9)$$

with $\gamma'_g = \frac{\gamma_g}{\xi}$. Hence, from a practical point of view, we can set $\xi = 1$ falling the relevance of the repulsion term on $\gamma'_g = \gamma_g$. In the following, calculi will be done with $\xi = 1$.

The main difference between PkM and PkM-R is given by the repulsion term

$\sum_{g=1}^{k} \gamma_g \sum_{g'=1, g' \neq g}^{k} \frac{1}{d^2\left(\mathbf{h}_g, \mathbf{h}_{g'}\right)}$. It acts avoiding that coincident clusters occur. In fact, if g and g' are almost coincident clusters, then $d^2\left(\mathbf{h}_g, \mathbf{h}_{g'}\right)$ is close to 0 and therefore the repulsion term tends to infinity. Obviously the repulsion term plays a relevant role only when the risk of getting coincident clusters is high. If $d^2\left(\mathbf{h}_g, \mathbf{h}_{g'}\right)$ is large $\forall g, g'$, the repulsion term is close to 0.
As for PkM, an iterative algorithm should be adopted for determining the updates of $\mathbf{H}$ and $\mathbf{U}$. Since the repulsion term can be considered as a constant when updating $\mathbf{U}$, the optimal value of u_{ig} does not vary with respect to the PkM case and, therefore, coincides with (5). Things concerning the update of $\mathbf{h}_g$ are remarkably more complex. According to Timm et al. [16], by equating the gradient $g(\mathbf{h}_g) = \frac{\partial J_{PkM-R}}{\partial \mathbf{h}_g}$ to zero the update of $\mathbf{h}_g$ is given by

$$\mathbf{h}_g = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i - \gamma_g \sum_{g'=1, g' \neq g}^{k} \frac{\mathbf{h}_{g'}}{\left[\left\|\mathbf{h}_g - \mathbf{h}_{g'}\right\|^2\right]^2}}{\sum_{i=1}^{n} u_{ig}^m - \gamma_g \sum_{g'=1, g' \neq g}^{k} \frac{1}{\left[\left\|\mathbf{h}_g - \mathbf{h}_{g'}\right\|^2\right]^2}}. \tag{10}$$

Timm et al. [16] suggest to solve (10) recursively because $\mathbf{h}_g$ appears also on the right side, i.e. $g(\mathbf{h}_g)$ is non-linear. However, in the literature, it is recognized that minimizing a function f or, equivalently, solving $g = f' = 0$ when g is non-linear should be done by means of second-order approximation methods, rather than with first-order approximations method as the gradient one (see, e.g., [3]). For this purpose, we suggest to consider the Newton method, which is briefly recalled below.

2.1 Newton method

Given $f : \mathbb{R}^p \longrightarrow \mathbb{R}$, we look for $\theta^* \in \mathbb{R}^p$ such that

$$f(\theta^*) = \min_{\theta} f(\theta). \tag{11}$$

The Newton method is an iterative procedure for determining θ^* in (11). Since the existence of θ^* is not guaranteed and it is therefore impossible to guarantee the convergence of a numerical algorithm to a global minimum, the objective reduces to find a local minimum θ^*. Generally speaking, the iterative solution of the Newton method takes the form

$$\theta^{(n+1)} = \theta^{(n)} + \varphi^{(n)} \Lambda^{(n)} \tag{12}$$

where $\theta^{(n)}$ is the value of θ at the $n-th$ iteration, $\varphi^{(n)}$ is the step length and $\Lambda^{(n)}$ is the direction of search. $\theta^{(n+1)}$ is such that $f(\theta^{(n+1)}) < f(\theta^{(n)})$. The Newton method consists in setting $\theta^{(n)} := \theta^{(n+1)}$ and repeating (12) upon convergence.
If the first and second derivatives of f exist, a second-order approximation of f can be found by Taylor expansion as

$$f(\theta) \approx f\left(\theta^{(n)}\right) + g\left(\theta^{(n)}\right)\left(\theta - \theta^{(n)}\right) + \frac{1}{2}\left(\theta - \theta^{(n)}\right)' H\left(\theta^{(n)}\right)\left(\theta - \theta^{(n)}\right), \quad (13)$$

where $g = f'$ is the gradient of f and $H = f''$ is the Hessian matrix of f. θ is a local optimum if $f'(\theta) = 0$ and from (13) we thus have

$$f'(\theta) = 0 + g\left(\theta^{(n)}\right)\left(\theta - \theta^{(n)}\right) + H\left(\theta^{(n)}\right)\left(\theta - \theta^{(n)}\right) = 0, \quad (14)$$

hence

$$\theta = \theta^{(n)} - \left[H\left(\theta^{(n)}\right)\right]^{-1} g\left(\theta^{(n)}\right), \quad (15)$$

where $[H\left(\theta^{(n)}\right)]^{-1}$ denotes the inverse of $H\left(\theta^{(n)}\right)$. By considering (12), we get $\varphi^{(n)} = 1$ and $\Lambda^{(n)} = \left[H\left(\theta^{(n)}\right)\right]^{-1} g\left(\theta^{(n)}\right)$.

Remark 1 (step-halving)
When the Newton algorithm runs and is close to a solution, it may occur that $f(\theta^{(n+1)}) > f(\theta^{(n)})$. This is so because the step length $\varphi^{(n)}$ is too large and, thus, the algorithm "overtakes" the minimum during the search. This drawback can be easily solved by the so-called backtracking step, which consists in reducing the step length by a fixed fraction δ each time. If $f(\theta^{(n+1)}) > f(\theta^{(n)})$ then $\theta^{(n+1)}$ should be rejected and replaced by a new solution setting $\varphi^{(n)} := \frac{\varphi^{(n)}}{\delta}$ until $f(\theta^{(n+1)}) \leq f(\theta^{(n)})$. A common choice is $\delta = \frac{1}{2}$ and is usually referred to as step-halving.

2.2 Newton method in PkM-R

In order to apply the Newton method in PkM-R for updating the prototypes, the gradient and the Hessian must be computed. It is easy to show that the gradient is

$$g(\mathbf{h}_g) = \frac{\partial J_{PkM-R}}{\partial \mathbf{h}_g} = -2\sum_{i=1}^{n} u_{ig}^{m}(\mathbf{x}_i - \mathbf{h}_g) - 2\gamma_g \sum_{g'=1, g' \neq g}^{k} \frac{(\mathbf{h}_g - \mathbf{h}_{g'})}{\left[\|\mathbf{h}_g - \mathbf{h}_{g'}\|^2\right]^2}, \quad (16)$$

and that the generic diagonal element of $\mathbf{H}_g$ is

$$\frac{\partial^2 J_{PkM-R}}{\partial^2 h_{gj}} = 2\sum_{i=1}^{n} u_{ig}^m - 2\gamma_g \sum_{g'=1,g'\neq g}^{k} \left[\frac{1}{\left[\|\mathbf{h}_g - \mathbf{h}_{g'}\|^2\right]^2} - \frac{4\left(h_{gj} - h_{g'j}\right)^2}{\left[\|\mathbf{h}_g - \mathbf{h}_{g'}\|^2\right]^3} \right], \quad (17)$$

while the generic off-diagonal one is

$$\frac{\partial^2 J_{PkM-R}}{\partial h_{gj} \partial h_{gj'}} = 8\gamma_g \sum_{g'=1,g'\neq g}^{k} \frac{\left(h_{gj} - h_{g'j}\right)\left(h_{gj'} - h_{g'j'}\right)}{\left[\|\mathbf{h}_g - \mathbf{h}_{g'}\|^2\right]^3}. \quad (18)$$

In order to update the prototypes we then use

$$\mathbf{h}_g^{(n+1)} = \mathbf{h}_g^{(n)} - [\mathbf{H}]^{-1} g\left(\mathbf{h}_g^{(n)}\right) \quad (19)$$

3 Fuzzy data and metrics for fuzzy data

Suppose now that the observed data are affected by imprecision and cannot be expressed in terms of single values. A useful approach consists in managing them in terms of fuzzy sets [20]. In this work, we limit our attention to the class of LR symmetric fuzzy numbers. In this case, the generic fuzzy datum $\widetilde{X}$ can be defined in terms of a pair of parameters, namely the center c and the spread $s(>0)$, and the so-called membership function giving the degree of membership of x in $\widetilde{X}$:

$$\mu_{\widetilde{X}}(x) = L\left(\frac{c-x}{s}\right), \qquad x \leq c \quad (s>0), \quad (20)$$

where the function $L : \mathbb{R} \to [0,1]$ is a convex upper semi-continuous function so that $L(0) = 1$ and $L(z) = 0$, for all $z \in \mathbb{R} \setminus [0,1]$. If $L(z) = 1 - z$ for $0 \leq z \leq 1$, then $\widetilde{X}$ is a symmetric triangular fuzzy number. A fuzzy datum can be seen as an interval of values between $c - s$ and $c + s$ with weights given by the membership function.
When the available data refer to a set of n objects on which p LR symmetric fuzzy variables are collected, we have a fuzzy data matrix

$$\widetilde{\mathbf{X}} = \left\{\widetilde{x}_{ij} \equiv (c_{ij}, s_{ij})_L, i = 1, ..., n, j = 1, ..., p\right\}, \quad (21)$$

where $\widetilde{x}_{ij} \equiv (c_{ij}, s_{ij})_L$ represents the LR symmetric fuzzy variable j observed on the i-th object with center c_{ij} and spread s_{ij}. Also we can define the matrices of the centers ($\mathbf{C}$) and of the spreads ($\mathbf{S}$) of order $(n \times p)$ with generic elements c_{ij} and s_{ij}, respectively.

3.1 Metrics for fuzzy data

A crucial point when dealing with fuzzy data concerns how to compute a suitable dissimilarity measure. The Euclidean distance could be applied to the centers of the fuzzy data but the conclusions are in general misleading since the spread information is ignored. In order to compare two objects i and i' characterized by one LR symmetric fuzzy variable ($p = 1$), Yang and Ko [18] suggest the following (squared) distance:

$$\begin{aligned} d_F^2(\tilde{x}_i, \tilde{x}_{i'}) &= (c_i - c_{i'})^2 + [(c_i - \lambda s_i) - (c_{i'} - \lambda s_{i'})]^2 + [(c_i + \lambda s_i) - (c_{i'} + \lambda s_{i'})]^2 \\ &= 3(c_i - c_{i'})^2 + 2(\lambda s_i - \lambda s_{i'})^2, \end{aligned} \tag{22}$$

with $\tilde{x}_i \equiv (c_i, s_i)_L$, $\tilde{x}_{i'} \equiv (c_{i'}, s_{i'})_L$ and $\lambda = \int_{\mathbb{R}} L(z)dz$ is a parameter taking values in [0,1] based on the shape of the membership function (in the triangular case $\lambda = \frac{1}{2}$). The intuition behind the Yang and Ko [18] distance is to compare two fuzzy numbers by the sum of the squared Euclidean distances between their centers and their "trimmed" lower bounds and upper bounds (bearing in mind that the spreads are scaled by $\lambda < 1$).
In (21) objects i and i' are characterized by $p > 1$ LR symmetric fuzzy variables. In particular, we denote the fuzzy vectors of length p for objects i and i' as $\tilde{\mathbf{x}}_i \equiv (\mathbf{c}_i, \mathbf{s}_i)_L$ and $\tilde{\mathbf{x}}_{i'} \equiv (\mathbf{c}_{i'}, \mathbf{s}_{i'})_L$, respectively. Note that $\tilde{\mathbf{x}}_i$, $\mathbf{c}_i$ and $\mathbf{s}_i$ are the i-th rows of $\widetilde{\mathbf{X}}$, $\mathbf{C}$ and $\mathbf{S}$, respectively. In this case, the objects are no longer represented by intervals (with weights). If $p = 2$ every object can be seen as a rectangle in $\mathbb{R}^2$ with four vertices and, if $p > 2$, as a hyperrectagle in R^p with 2^p vertices. In these cases, the (squared) distance between the objects i and i' can be computed as the sum of the squared Euclidean distances between the centers and all the 2^p vertices [8]. It can be shown that, after a little algebra, such a distance can be rewritten as

$$d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'}) = (2^p + 1)\|\mathbf{c}_i - \mathbf{c}_{i'}\|^2 + 2^p \|\boldsymbol{\lambda} * \mathbf{s}_i - \boldsymbol{\lambda} * \mathbf{s}_{i'}\|^2, \tag{23}$$

where $\boldsymbol{\lambda}$ is the p-vector with generic element $\lambda_j = \int_{\mathbb{R}} L_j(z)dz$, where $L_j(z)$ is the membership function of the LR symmetric fuzzy variable j and $*$ denotes the Hadamard product (elementwise product between vectors of the same length). In the following we shall consider the case in which all the variables have the same membership function. Thus, $\lambda_j = \lambda$, $j = 1, \ldots, p$, and (23) can be simplified as

$$d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'}) = (2^p + 1)\|\mathbf{c}_i - \mathbf{c}_{i'}\|^2 + 2^p \lambda^2 \|\mathbf{s}_i - \mathbf{s}_{i'}\|^2. \tag{24}$$

It is worth mentioning that, if $p = 1$, (23) and (24) reduce to (22). See, for more details, [8]. As it will be clarified in the next section, the metric in (24) shall be considered for extending PkM-R to fuzzy data.

4 Fuzzy clustering with repulsion constraints for fuzzy data

In this section we propose a generalization of PkM-R to fuzzy data. In the last decade, a few attempts to generalize fuzzy clustering for fuzzy data have been suggested in the literature. For an overview see [7].
In the case of fuzzy data, it must be underlined that the concept of homogeneous clusters is related to the positions (i.e. the centers) and/or the sizes (i.e. the spreads) of the hyperrectangles. It follows that outliers can be determined not only with respect to the location of the objects, as in the non-fuzzy data case, but also with respect to the associated imprecision. Although this is a relevant problem in the domain of cluster analysis for fuzzy data, a limited number of robust variants of FkM for such a kind of data exists [4, 10]. In [10], a robust version of FkM for fuzzy data is introduced such that, following the metric approach, an exponential-based distance for fuzzy data is considered in the loss function. Note that this proposal works only in the univariate case ($p = 1$). A possibilistic clustering model for fuzzy data (hereinafter PkM-F) has been suggested in [4]. Similarly to standard PkM, also in [4] the use of the FkM for fuzzy data as starting point is recommended to solve empirically the coincident cluster problem. Here, we are going to introduce a new possibilistic clustering model for fuzzy data involving repulsion constraints between pairs of prototypes. This represents a more elegant and fruitful way to handle the risk of obtaining coincident clusters. To do it, we assume that the prototypes inherit the imprecision of the observed data and hence are LR symmetric fuzzy numbers. Thus, the prototype matrix is $\widetilde{\mathbf{H}} \equiv \left(\mathbf{H}^C, \mathbf{H}^S\right)_L$ with generic element $\tilde{h}_{gj} \equiv \left(h_{gj}^C, h_{gj}^S\right)_L$, where $\mathbf{H}^C$ and $\mathbf{H}^S$ are the matrices of the centers and of the spreads with generic elements h_{gj}^C and h_{gj}^S, respectively. Furthermore, in order to take into account properly the features of the observed data and of the prototypes, the distance in (24) is adopted.
The Possibilistic k-Means with Repulsion constraints for Fuzzy data (PkM-

RF) can be formalized as

$$\begin{aligned}\min_{\mathbf{U},\widetilde{\mathbf{H}}} J_{PkM-RF} &= \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d_F^2\left(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{h}}_g\right) + \sum_{g=1}^{k}\eta_g \sum_{i=1}^{n}\left(1-u_{ig}\right)^m \\ &\quad + \sum_{g=1}^{k}\gamma_g \sum_{g'=1, g'\neq g}^{k} \frac{1}{d_F^2\left(\widetilde{\mathbf{h}}_g, \widetilde{\mathbf{h}}_{g'}\right)} \qquad (25a) \\ \text{s.t.} \quad & u_{ig} \in [0,1], i = 1,\ldots,n, \forall g = 1,\ldots,k. \qquad (25b)\end{aligned}$$

Note that in (25a) all the ingredients are based on d_F^2 in (24) rather than on the standard d^2. The parameters η_g and γ_g play the same role as for PkM-R and can be defined as reported in (7) and (8), respectively, provided that the prototypes and the membership degrees are obtained considering a suitable FkM for fuzzy data. In our analyses, we used the FkM for Fuzzy data (FkM-F) proposed in [6].
The loss in (25a) must be minimized with respect to $\mathbf{U}$, $\mathbf{C}$ and $\mathbf{S}$. The update of the elements of $\mathbf{U}$, keeping fixed $\mathbf{C}$ and $\mathbf{S}$ is given by (5), provided that d^2 is replaced by d_F^2. The updates of the centers and spreads of the prototypes is computed by Newton algorithm since the gradients $g(\mathbf{h}_g^C)$ and $g(\mathbf{h}_g^S)$ are non-linear.

Update of the centers of the prototypes:
The update of the prototype centers $\mathbf{h}_g^C$ can be done by

$$\mathbf{h}_g^{C(n+1)} = \mathbf{h}_g^{C(n)} - \left[\mathbf{H}_c\right]^{-1} g\left(\mathbf{h}_g^{C(n)}\right), \qquad (26)$$

where

$$\begin{aligned} g\left(\mathbf{h}_g^C\right) &= \frac{\partial J_{PkM-RF}}{\partial \mathbf{h}_g^C} = -2\left(2^p+1\right)\sum_{i=1}^{n} u_{ig}^m\left(\mathbf{c}_i - \mathbf{h}_g^C\right) \\ &\quad -2\left(2^p+1\right)\gamma_g \sum_{g'=1,g'\neq g}^{k} \frac{\left(\mathbf{h}_g^C - \mathbf{h}_{g'}^C\right)}{\left[\left(2^p+1\right)\left\|\mathbf{h}_g^C - \mathbf{h}_{g'}^C\right\|^2 + 2^p\lambda^2\left\|\mathbf{h}_g^S - \mathbf{h}_{g'}^S\right\|^2\right]^2}. \qquad (27)\end{aligned}$$

The generic diagonal element of the Hessian matrix can be computed as

$$\begin{aligned}\frac{\partial^2 J_{PkM-RF}}{\partial^2 h_{gj}^C} &= 2\left(2^p+1\right)\sum_{i=1}^n u_{ig}^m \\ &\quad -2\left(2^p+1\right)\gamma_g \sum_{g'=1,g'\neq g}^k \left[\frac{1}{\left[(2^p+1)\left\|\mathbf{h}_g^C-\mathbf{h}_{g'}^C\right\|^2+2^p\lambda^2\left\|\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right\|^2\right]^2}\right. \\ &\quad \left.-\frac{4(2^p+1)\left(h_{gj}^C-h_{g'j}^C\right)^2}{\left[(2^p+1)\left\|\mathbf{h}_g^C-\mathbf{h}_{g'}^C\right\|^2+2^p\lambda^2\left\|\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right\|^2\right]^3}\right]\end{aligned} \tag{28}$$

and the off-diagonal one as

$$\frac{\partial^2 J_{PkM-RF}}{\partial h_{gj}^C \partial h_{gj'}^C} = 8\left(2^p+1\right)^2\gamma_g \sum_{g'=1,g'\neq g}^k \frac{\left(h_{gj}^C-h_{g'j}^C\right)\left(h_{gj'}^C-h_{g'j'}^C\right)}{\left[\left(2^p+1\right)\left\|\mathbf{h}_g^C-\mathbf{h}_{g'}^C\right\|^2+2^p\lambda^2\left\|\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right\|^2\right]^3}. \tag{29}$$

Update of the spreads of the prototypes:
Similarly, we can update the prototype spreads $\mathbf{h}_g^S$ as

$$\mathbf{h}_g^{S(n+1)} = \mathbf{h}_g^{S(n)} - \left[\mathbf{H}_s\right]^{-1} g\left(\mathbf{h}_g^{S(n)}\right) \tag{30}$$

with

$$\begin{aligned} g\left(\mathbf{h}_g^S\right) &= \frac{\partial J_{PkM-RF}}{\partial \mathbf{h}_g^S} = -2\left(2^p\lambda^2\right)\sum_{i=1}^n u_{ig}^m\left(\mathbf{s}_i-\mathbf{h}_g^S\right) \\ &\quad -2\left(2^p\lambda^2\right)\gamma_g\sum_{g'=1,g'\neq g}^k \frac{\left(\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right)}{\left[(2^p+1)\left\|\mathbf{h}_g^C-\mathbf{h}_{g'}^C\right\|^2+2^p\lambda^2\left\|\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right\|^2\right]^2}, \end{aligned} \tag{31}$$

$$\begin{aligned}\frac{\partial^2 J_{PkM-RF}}{\partial^2 h_{gj}^S} &= 2\left(2^p\lambda^2\right)\sum_{i=1}^n u_{ig}^m \\ &\quad -2\left(2^p\lambda^2\right)\gamma_g \sum_{g'=1,g'\neq g}^k \left[\frac{1}{\left[(2^p+1)\left\|\mathbf{h}_g^C-\mathbf{h}_{g'}^C\right\|^2+2^p\lambda^2\left\|\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right\|^2\right]^2}\right. \\ &\quad \left.-\frac{4\left(2^p\lambda^2\right)\left(h_{gj}^S-h_{g'j}^S\right)^2}{\left[(2^p+1)\left\|\mathbf{h}_g^C-\mathbf{h}_{g'}^C\right\|^2+2^p\lambda^2\left\|\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right\|^2\right]^3}\right]\end{aligned} \tag{32}$$

and

$$\frac{\partial^2 J_{PkM-RF}}{\partial h_{gj}^S \partial h_{gj'}^S} = 8\left(2^p\lambda^2\right)^2\gamma_g \sum_{g'=1,g'\neq g}^k \frac{\left(h_{gj}^S-h_{g'j}^S\right)\left(h_{gj'}^S-h_{g'j'}^S\right)}{\left[\left(2^p+1\right)\left\|\mathbf{h}_g^C-\mathbf{h}_{g'}^C\right\|^2+2^p\lambda^2\left\|\mathbf{h}_g^S-\mathbf{h}_{g'}^S\right\|^2\right]^3}. \tag{33}$$

Remark 2 (starting point of the PkM-RF algorithm)
The starting point of the PkM-RF algorithm can be chosen either randomly (i.e. randomly generating the membership degree matrix $\mathbf{U}^{(0)}$ fulfilling (25b)) or rationally (i.e. using $\mathbf{U}^{(0)}$ obtained from FkM-F). We saw that the former choice should be preferred (note that the solution of FkM-F must always be computed since it is required for defining η_g and γ_g). In fact, a severe local optimum problem was observed starting in a random manner. It is worth mentioning that if random starting points are considered, then the algorithm usually converges to a local optimum and the corresponding solution is not characterized by coincident clusters. Therefore, our suggestion to use a rational starting point is not connected with the attempt to solve empirically such a problem. Rather, it is related to the more general problem of hitting the global optimum whenever iterative algorithms are implemented.

Algorithm PkM-RF

Step 0: Define η_g and γ_g according to (7) and (8), respectively, and $\mathbf{U}^{(0)}$ using the FkM-F solution. Set $t := 1$.

Step 1: Update the prototypes $\mathbf{H}^{(t)}$ according to (26) and (30) keeping fixed $\mathbf{U}^{(t-1)}$.

Step 2: Update the membership degree matrix $\mathbf{U}^{(t)}$ according to (5) provided that d^2 is replaced by d_F^2 keeping fixed $\mathbf{H}^{(t)}$.

Step 3: If $\left\|\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)}\right\| < \varepsilon$, where ε is a small number fixed in advance (we use $\varepsilon = 10^{-4}$), the algorithm has converged, otherwise set t:=t+1 and go to Step 1.

5 Applications

This section is devoted to two applications of PkM-RF to synthetic and real data. A comparison with the results obtained using PkM-F [4] is also discussed.

5.1 Synthetic data

We generated $n = 22$ objects with respect to $p = 2$ LR symmetric fuzzy variables such that $k = 2$ clusters with 10 objects each and two outliers can be found. The centers and the spreads of the objects (n.1–n.10) belonging to Cluster 1 were generated randomly from U[0,1]. Also the spreads of the objects (n.11–n.20) belonging to Cluster 2 came from U[0,1], whereas the centers from U[0,1]+w. The two outliers were such that the first one (n.21) had spreads from U[0,1] and centers from U[0,1]+$\frac{1}{2}w$ and the latter one (n.22) spreads from U[0,1]+$\frac{1}{2}w$ and centers from U[0,1]. The data with $w = 0.5, 1.0, 2.0$ (partially overlapped, separated and well separated clusters, respectively) were analyzed by PkM-RF and PkM-F setting $m = 2$ and using the solution from FkM-F as starting point. Obviously, the models were expected to work better as w increased. The results are summarized in Table 1. The first two rows of Table 1 give the number of times (out of ten)

Table 1: Solutions obtained using PkM-RF and PkM-F

	$w = 0.5$		$w = 1.0$		$w = 2.0$	
	PkM-RF ($\gamma = 3$)	PkM-F	PkM-RF ($\gamma = 2$)	PkM-F	PkM-RF ($\gamma = 4$)	PkM-F
$\sum_{i=1}^{10} I_{\{u_{i1} \geq 0.5\}}$	5	2	8	3	9	9
$\sum_{i=1}^{10} u_{i1} \Big/ 10$	0.52	0.20	0.58	0.34	0.62	0.68
$\sum_{i=11}^{20} I_{\{u_{i2} \geq 0.5\}}$	3	3	6	4	8	8
$\sum_{i=11}^{20} u_{i2} \Big/ 10$	0.52	0.30	0.59	0.43	0.64	0.73
$\sum_{i=21}^{22} \sum_{g=1}^{2} I_{\{u_{ig} \geq 0.5\}}$	0	0	0	0	0	0
$\sum_{i=21}^{22} \sum_{g=1}^{2} u_{ig} \Big/ 4$	0.34	0.00	0.27	0.02	0.14	0.04
$\sum_{g=1}^{2} d^2\left(\hat{\mathbf{h}}_g^C, \bar{\mathbf{h}}_g^C\right)$	0.10	0.23	0.01	0.13	0.01	0.00
$\sum_{g=1}^{2} d^2\left(\hat{\mathbf{h}}_g^S, \bar{\mathbf{h}}_g^S\right)$	0.02	0.09	0.00	0.10	0.00	0.01

in which objects n.1–n.10 were well assigned to Cluster 1 in the hard clustering sense (I denotes the indicator function) and the average membership degree of objects n.1–n.10 in Cluster 1. The same information for objects n.11–n.20 and for Cluster 2 can be found in the third and fourth rows of the table. The next two rows contain the number of times in which an outlier was assigned to a cluster in the hard clustering sense and the average membership degree of the outliers in all the clusters. Finally, in the last two rows we report the squared Euclidean distances for, respectively, the centers (superscript C) and the spreads (superscript S) between the known-in-advance prototypes ($\bar{\mathbf{h}}_g^C = \left\{ \bar{h}_{gj}^C = \sum_{i=1+10(g-1)}^{1+10(g-1)} c_{ij} \Big/ 10, j = 1,2 \right\}, g = 1,2$ and $\bar{\mathbf{h}}_g^S = \left\{ \bar{h}_{gj}^S = \sum_{i=1+10(g-1)}^{1+10(g-1)} s_{ij} \Big/ 10, j = 1,2 \right\}, g = 1,2$) and the estimated ones ($\hat{\mathbf{h}}_g^C = \left\{ \hat{h}_{gj}^C, j = 1,2 \right\}, g = 1,2$ and $\hat{\mathbf{h}}_g^S = \left\{ \hat{h}_{gj}^S, j = 1,2 \right\}, g = 1,2$).
By inspecting Table 1, we can see that PkM-RF worked better than PkM-F when $w = 0.5$ and $w = 1.0$. When $w = 2.0$, the two models well assigned the same numbers of objects to the known-in-advance clusters, but the membership degrees are higher for PkM-F. This can be explained noting that a peculiarity of PkM-F is its tendency to produce membership degrees close to zero unless there exists a strong evidence of assigning an object to a given cluster. On the contrary, PkM-RF tends to produce a fuzzier membership degree matrix. In other words, if an object does not belong to a cluster, then the corresponding membership degree from PkM-F is very close to 0, whereas that from PkM-RF is low, but, in general, sensibly higher than 0. This implies that the solutions of PkM-F were such that some non-outlier objects were correctly assigned to a cluster with a high membership degree and some others wrongly had low membership degrees to both the clusters, whereas the outlier objects had membership degrees close to zero in both the clusters. This comment does not hold for the PkM-RF solutions in which the non-outlier objects were often assigned correctly to one of the two clusters, sometimes with a not remarkably high membership degree, and the outliers had low (within 0 and 0.50) membership degrees in all the clusters. The last two rows of Table 1 show that PkM-RF worked better than PkM-F in recovering the prototypes. Finally note that coincident clusters were never found.
In the previous analyses, γ was found by running PkM-RF considering dif-

ferent values of γ and choosing the best solution from a subjective point of view. However, we now give an objective hint that can help to choose γ. Consider the data with $w = 0.5$. As we can see from the last two rows of Table 1, using $\gamma = 2$ the prototypes obtained from PkM-RF are very accurate estimates of the known-in-advance ones. When we decreased γ setting $\gamma = 0.1$ or $\gamma = 0.01$ we found that the two clusters (the two prototypes) were almost coincident. If γ decreases, the role of the repulsion term in the loss function is negligible (if $\gamma = 0$, the repulsion term vanishes) and the algorithm produces coincident clusters unless well-separated clusters exist. When we increased γ, for instance choosing $\gamma = 20$, an abnormal role of the repulsion term can be observed. In this case, the estimated prototypes tend to be located outside from the two clouds of objects belonging to Clusters 1 and 2, respectively. It occurs because the repulsion term plays a very relevant role in the loss and the algorithm attains the minimum trying to minimize mainly the repulsion term by finding prototypes very far from each other regardless the observed data. This usually implies that the objects are far from all the prototypes and, therefore, the membership degrees in all the clusters go to zero. Summing up, the choice of γ should be such that the prototypes are representative of the objects assigned to the corresponding clusters and such that neither coincident clusters occur nor the membership degrees of the objects in all the clusters tend to zero.

5.2 Students data

We consider now the Students data in [4]. The data are the answers of 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 a questionnaire about the the global economic and financial crisis. Every student gave his/her opinion and perception about six specific questions concerning the influence of financial speculation on the crisis (Q1), the opinion about the utility of a new regolarization of financial markets (Q2), the feeling about the need for a drastic change of the economic system (Q3), the opinion on the adequacy of the EU economic measures to handle the crisis (Q4), the opinion about the Italian economic measures (Q5) and, finally, the perception about the trend of the Italian economy during the following three years (Q6). The responses are fuzzified following the scheme explained in [4].
In order to compare the results with those obtained by means of PkM-F

we have set $m = 1.5$ and $k = 3$. Setting $\gamma = 2$ the prototypes and the membership degrees obtained from PkM-RF are reported, respectively, in Table 2 and Table 3.

Table 2: Prototypes from PkM-RF (data take the form (center, spread)).

Cluster	Q1	Q2	Q3	Q4	Q5	Q6
Cluster 1	(0.89, 0.06)	(0.73, 0.12)	(0.69, 0.12)	(0.58, 0.12)	(0.29, 0.11)	(0.40, 0.12)
Cluster 2	(0.95, 0.04)	(0.06, 0.04)	(0.98, 0.02)	(0.09, 0.06)	(0.05, 0.04)	(0.03, 0.03)
Cluster 3	(0.61, 0.12)	(0.51, 0.12)	(0.51, 0.12)	(0.45, 0.12)	(0.54, 0.12)	(0.51, 0.12)

The prototypes reported in Table 2 are more or less the same obtained from PkM-F. The students belonging to Cluster 1 are anxious for the crisis and have a quite good opinion on the adequacy of the EU economic measures to face up it, better than the Italian measures. Cluster 2 is characterized by students with a radical position, that is, they are worried about the crisis and think the the EU and Italian economic measures are not appropriate, requiring a drastic change of the economic systems. Students belonging to Cluster 3 are not particularly anxious for the crisis and are satisfied with the EU and Italian economic measure to handle it. By inspecting the membership degrees, we can conclude that PkM-RF and PkM-F detect the same outliers. Moreover, differently from PkM-F, the proposed method assigns a larger number of non-outlier students to the clusters in the hard clustering sense. Also we note that the average membership degrees of the non-outlier students in the clusters resulting from PkM-RF are higher than those from PkM-F.

6 Concluding remarks

A Possiblistic k-Means clustering model with Repulsion constraints for Fuzzy data (PkM-RF) has been proposed. Its peculiarities involve the use of a suitable distance measure for fuzzy data allowing us to detect clusters of homogeneous objects in terms of the location and the size of the hyperrectangles associated with the objects. To solve the well-known coincident cluster problem, a repulsion term among prototypes has been considered and extended to the fuzzy data case. We provided the iterative solution of the algorithm by means of the Newton algorithm since a non-linear problem for updating the prototypes of the clusters must be solved. The applicability of the proposed

Table 3: Membership degrees from PkM-RF.

Student (n.)	Cluster 1	Cluster 2	Cluster 3
1	0.03	0.15	0.01
2	0.02	0.98	0.01
3	0.16	0.00	0.85
4	0.82	0.00	0.36
5	0.25	0.00	0.05
6	0.03	0.71	0.01
7	0.02	0.89	0.01
8	0.92	0.01	0.13
9	0.04	0.00	0.21
10	0.01	0.94	0.01
11	0.22	0.00	0.98
12	0.33	0.00	0.86
13	0.45	0.01	0.05
14	0.75	0.01	0.05
15	0.88	0.01	0.24
16	0.63	0.01	0.10
17	0.60	0.00	0.07
18	0.94	0.01	0.15
19	0.22	0.01	0.91
20	0.09	0.00	0.70
21	0.06	0.00	0.59
22	0.05	0.02	0.12
23	0.02	0.00	0.02
24	0.03	0.49	0.02
25	0.03	0.61	0.01
26	0.10	0.00	0.09
27	0.02	0.75	0.02

technique has been analyzed by means of two examples, which showed the usefulness of PkM-RF.

References

[1] Barni, M., Cappellini, V., Mecocci, A., 1996. Comments on 'a possibilistic approach to clustering'. IEEE Transactions on Fuzzy Systems 4, 393–396.

[2] Bezdek, J.C. (1981). Pattern recognition with fuzzy objective function algorithm, Plenum Press.

[3] Bonnans, J.F., Gilbert, J.C., Lemarèchal, C., Sagastizábal, C.A. (2006) Numerical Optimization, Theoretical and Practical Aspects. Springer-Verlag Berlin Heidelberg.

[4] Coppi, R., D'Urso, P., Giordani, P., (in press). Fuzzy and possibilistic clustering for fuzzy data. Computational Statistics and Data Analysis, doi: 10.1016/j.csda.2010.09.013.

[5] Davè, R., (1991) Characterization and detection of noise in clustering, Pattern Recognition Letters, 12, 657–664.

[6] D'Urso, P., Giordani, P., (2006). A weighted fuzzy c-means clustering model for fuzzy data. Computational Statistics and Data Analysis 50, 1496–1523.

[7] 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.

[8] Giordani, P., Kiers, H., (2004). Principal Component Analysis of symmetric fuzzy data, Computational Statistics and Data Analysis 45, 519–548.

[9] 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.

[10] 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.

[11] Krishnapuram, G.L., Keller, J., (1993). A possibilistic approach to clustering, IEEE Transactions on Fuzzy Systems 1, 98–110.

[12] Krishnapuram, G.L., Keller, J., (1996). The possibilistic c-means algorithm: insights and recommendations, IEEE Transactions on Fuzzy Systems 4, 385–393.

[13] Masson, M.-H., Denœux, T., 2008. ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recognition 41, 1384–1397.

[14] Ménard, M., Courboulay, V., Dardignac, P.A., (2003). Possibilistic and probabilistic fuzzy clustering: unification within the framework of the non-extensive thermostatistics. Pattern Recognition 36 1325–1342.

[15] 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, 332–336.

[16] Timm, H., Borgelt, C., Döring, C., Kruse, R., (2004). An extension to possibilistic fuzzy cluster analysis, Fuzzy Sets and Systems 147, 3–16.

[17] Winkler, R., Rehm, F., Kruse, R., (2010). Clustering with repulsive prototypes in: Fink, A., Lausen, B., Seidel, W., Ultsch, A. (Eds.): Data Analysis, Data Handling and Business Intelligence, 207–215. Springer-Verlag Berlin Heidelberg.

[18] 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.

[19] Yang, M.S., Wu, K.L., 2006. Unsupervised possibilistic clustering. Pattern Recognition 39, 5–21.

[20] Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.