Fuzzy clustering with Minkowski distance functions

Patrick J.F. Groenen[*] Uzay Kaymak[†]
Joost van Rosmalen [‡]

July 6, 2006

Econometric Institute Report EI 2006-24

Abstract

Distances in the well known fuzzy c-means algorithm of Bezdek (1973) are measured by the squared Euclidean distance. Other distances have been used as well in fuzzy clustering. For example, Jajuga (1991) proposed to use the L_1-distance and Bobrowski and Bezdek (1991) also used the L_∞-distance. For the more general case of Minkowski distance and the case of using a root of the squared Minkowski distance, Groenen and Jajuga (2001) introduced a majorization algorithm to minimize the error. One of the advantages of iterative majorization is that it is a guaranteed descent algorithm, so that every iteration reduces the error until convergence is reached. However, their algorithm was limited to the case of Minkowski parameter between 1 and 2, that is, between the L_1-distance and the Euclidean distance. Here, we extend their majorization algorithm to any Minkowski distance with Minkowski parameter greater than (or equal to) 1. This extension also includes the case of the L_∞-distance. We also investigate how well this algorithm performs and present an empirical application.

[*] Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, groenen@few.eur.nl).

[†] kaymak@few.eur.nl

[‡] vanrosmalen@few.eur.nl

1 Introduction

Since Ruspini (1969) first proposed the idea of fuzzy partitions, fuzzy clustering has grown to be an important tool for data analysis and modelling. Especially after the introduction of the fuzzy c-means algorithm (Dunn, 1973; Bezdek, 1973), objective function-based fuzzy clustering has received much attention from the scientific community as well as the practitioners of fuzzy set theory (Bezdek & Pal, 1992; Yang, 1993; Baraldi & Blonda, 1999a, 1999b; Höppner, Klawonn, Kruse, & Runkler, 1999). Consequently, fuzzy clustering has been applied extensively for diverse tasks such as pattern recognition, data analysis, data mining, image processing, and engineering systems design. Objective function-based fuzzy clustering has also become one of the key techniques in fuzzy modelling, where they are used for partitioning the feature space from which the rules of a fuzzy system can be derived (Babuška, 1998).

In general, objective function-based fuzzy clustering algorithms partition a data set into overlapping groups by minimizing an objective function derived from the distance between the cluster prototypes and the data points (or objects). The clustering results are largely influenced by how this distance is computed, since it determines the shape of the clusters. The success of fuzzy clustering in various applications may depend very much on the shape of the clusters. As a result, there is a significant amount of literature on fuzzy clustering, which is aimed at investigating the use of different distance functions in fuzzy clustering, leading to different cluster shapes.

One way of influencing the shape of the clusters is to consider prototypes with a geometric structure. The fuzzy c-varieties (FCV) algorithm uses linear subspaces of the clustering space as prototypes (Bezdek, Coray, Gunderson, & Watson, 1981b), which is useful for detecting lines and other linear structures in the data. The fuzzy c-elliptotypes (FCE) algorithm takes convex combinations of fuzzy c-varieties prototypes with fuzzy c-means prototypes to obtain localized clusters (Bezdek, Coray, Gunderson, & Watson, 1981a). Kaymak and Setnes (2002) proposed using volumes in the clustering space as the cluster prototypes.

Another way for influencing the shape of the clusters is modifying the distance measure that is used in the objective function. Distances in the well known fuzzy c-means algorithm of Bezdek (1973) are measured by the squared Euclidean distance. Gustafson and Kessel (1979) use the quadratic Mahanalobis norm to measure the distance. Jajuga (1991) proposed to use the L_1-distance and Bobrowski and Bezdek (1991) also used the L_∞-distance. Further, Hathaway, Bezdek, and Hu (2000) studied the Minkowski semi-norm as the dissimilarity function.

In this paper, we consider fuzzy clustering with the more general case of the Minkowski distance and the case of using a root of the squared Minkowski distance. The Minkowski norm provides a concise, parametric distance function that generalizes many of the distance functions used in the literature. The advantage is that mathematical results can be shown for a whole class of distance functions, and the user can adapt the distance function to suit the needs of the application by modifying the Minkowski parameter. By considering the additional case of the roots of the squared Minkowski distance, we introduce an extra parameter that can be used to control the behaviour of the clustering algorithm with respect to outliers. This provides an additional way of dealing with outliers, which is different than the “noise cluster” approach proposed in Dave (1991).

Our analysis follows the approach that Groenen and Jajuga (2001) introduced previously. Minimization of the objective function is partly done by iterative majorization. One of the advantages of iterative majorization is that it is a guaranteed descent algorithm, so that every iteration reduces the objective function until convergence is reached. The algorithm in Groenen and Jajuga (2001) was limited to the case of a Minkowski parameter between 1 and 2, that is, between the L_1-distance and the Euclidean distance. Here, we extend their majorization algorithm to any Minkowski distance with Minkowski parameter greater than (or equal to) 1. This extension also includes the case of the L_∞-distance. We also explore the behaviour of our algorithm with an illustrative example using real-world data.

The outline of the paper is as follows. We expose the formalization of the clustering problem in Section 2. The majorizing algorithm for fuzzy c-means with Minkowski distances is given in Section 3. We discuss in Section 4 the behaviour of our algorithm by using an illustrative example based on empirical data concerning attitudes about the Internet. Finally, conclusions are given in Section 5.

2 Formalization

In this paper, we focus on the fuzzy clustering problem that uses a root of the squared Minkowski distance. This problem can be formalized by minimizing the objective (or loss) function

$$L(\mathbf{F}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{k=1}^{K} f_{ik}^{s} d_{ik}^{2\lambda}(\mathbf{V}) \tag{1}$$

Table 1: Special distances obtained by specific choice of λ and p and some of their properties.

p	λ	Distance	Assumed cluster shape	Robust
1	1.0	Squared L_1	box/diamond	no
1	0.5	L_1	box/diamond	yes
2	1.0	Squared Euclidean	circular	no
2	0.5	Unsquared Euclidean	circular	yes
∞	1.0	Squared Dominance	box	no
∞	0.5	Dominance	box	yes

under the constraints

$$\begin{array}{ll} 0 \leq f_{ik} \leq 1, & i = 1, ..., n \quad k = 1, ..., K \\ \sum_{k=1}^{K} f_{ik} = 1, & i = 1, ..., n \end{array} \tag{2}$$

where n is the number of objects, K is the number of fuzzy clusters, f_{ik} is the membership grade of object i in fuzzy cluster k, s is the weighting exponent larger than 1. The distance between object i given by the i-th row of the $n \times m$ data matrix $\mathbf{X}$ and fuzzy cluster k of the $K \times m$ cluster coordinate matrix $\mathbf{V}$ is given by

$$d_{ik}^{2\lambda}(\mathbf{V}) = \left(\sum_{j=1}^{m} |x_{ij} - v_{kj}|^p \right)^{2\lambda/p}, \quad 1 \leq p \leq \infty, \quad 0 \leq \lambda \leq 1 \tag{3}$$

where λ is the power of the squared Minkowski distance $d_{ik}^{2\lambda}(\mathbf{V})$ with $1 \leq p \leq \infty$.

The introduction of the power λ allows to control the loss function against outliers. For large λ, e.g., $\lambda = 1$, outliers may dominate the loss function, whereas the loss function will be more robust if λ is small.

The use of Minkowski distances allows to vary the assumptions of the shape of the clusters by varying p. The most often used value is $p = 2$ that assumes a circular cluster shape. Using $p = 1$ assumes that the clusters are in the shape of a (rotated) square in two dimensions or a diamond like shape in three or more dimensions. For $p = \infty$, the clusters are assumed to be in the form of a box with sides parallel to the axes. Both $p = 1$ and $p = \infty$ can be used in cases where the data structures have "boxy" shapes, i.e., shapes with sharp "edges" (Bobrowski & Bezdek, 1991). A summary of combinations of λ and p and some properties of the distances are presented in Table 1 (taken from Groenen and Jajuga 2001).

Groenen and Jajuga (2001) note that (1) has several known fuzzy clustering models as a special case. For example, for $p = 2$ and $\lambda = 1$, the important member of fuzzy ISODATA family is obtained that corresponds to squared Euclidean distances (while assuming the identity metric). A fuzzy clustering objective function that is robust against outliers can be obtained by choosing $\lambda = 1/2$ and $p = 1$ so that the L_1-norm is used. Note that this choice implicitly assumes a "boxy" shape of the clusters. A robust version of fuzzy clustering with a circular shape can be specified by $\lambda = 1/2$ and $p = 2$ that implies the unsquared Euclidean distance. Thus, λ takes care of robustness issues and p of the shape of the clusters. Dodge and Rousson (1998) named the cluster centroids for $\lambda = 1/2$ and $p = 1$ 'L_1-medians', for $\lambda = 1$ and $p = 1$ 'L_1-means', for $\lambda = 1/2$ and $p = 2$ 'L_2-medians', and for $\lambda = 1$ and $p = 2$ the well known 'L_2-means'.

3 The Majorizing Algorithm for Fuzzy *c*-means with Minkowski Distances

Depending on the particular function, the minimization method of iterative majorization has some nice properties. The most important one is that in each iteration of the iterative majorization the loss function is decreased until this value converges. Such guaranteed descent methods are useful because no step in the wrong direction can be taken. Note that this property does not imply that a global minimum is found unless the function exhibits a special property such as convexity. Some general papers on iterative majorization are De Leeuw (1994), Heiser (1995), Lange, Hunter, and Yang (2000), Kiers (2002), and Hunter and Lange (2004); an introduction can be found in Borg and Groenen (2005).

The majorization algorithm of Groenen and Jajuga (2001) worked for all $1 \leq p \leq 2$. Below we expand their majorization algorithm to the situation of all $p > 1$. Each iteration of their algorithm consists of two steps: (1) update the cluster memberships $\mathbf{F}$ for fixed centers $\mathbf{V}$ and (2) update $\mathbf{V}$ for fixed $\mathbf{F}$. For Step (2) we use majorization. Below, we start by explaining some basic ideas of iterative majorization. Then, the update of the cluster memberships is given. This is followed by some derivations for the update of the cluster centers $\mathbf{V}$ in the case of $1 \leq p \leq 2$. Then, the update is derived for $2 < p < \infty$ and a special update for the case of $p = \infty$.

3.1 Iterative Majorization

Iterative majorization can be seen as a gradient method with a fixed step size. However, iterative majorization can also be applied to functions that are at some points nondifferentiable. Central to iterative majorization is the use of an auxiliary function similar to the first order Taylor expansion used as an auxiliary function in a gradient method and second order expansion for Newton's method. The unique feature of the auxiliary function in iterative majorization—the so-called majorizing function— is that it touches the original function or is located above it. In contrast, the auxiliary functions of the gradient method or Newton's method can be partially below and above the original function.

Let the original function be presented by $\varphi(\mathbf{X})$, the majorizing function by $\hat{\varphi}(\mathbf{X},\mathbf{Y})$, where $\mathbf{Y}$ is the current known estimate. Then, a majorizing function has to fulfil the following three requirements: (1) $\hat{\varphi}(\mathbf{X},\mathbf{Y})$ is a more simple function in $\mathbf{X}$ than $\varphi(\mathbf{X})$, (2) it touches $\varphi(\mathbf{X})$ at the known supporting point $\mathbf{Y}$ so that $\varphi(\mathbf{Y}) = \hat{\varphi}(\mathbf{Y},\mathbf{Y})$, and (3) $\hat{\varphi}(\mathbf{X},\mathbf{Y})$ is never smaller than $\varphi(\mathbf{X})$, that is, $\varphi(\mathbf{X}) \leq \hat{\varphi}(\mathbf{X},\mathbf{Y})$ for all $\mathbf{X}$. Often, the majorizing function is either linear or quadratic.

To see how a single iteration reduces $\varphi(\mathbf{X})$, consider the following. Let $\mathbf{Y}$ be some known point and let the minimum of the majorizing function $\hat{\varphi}(\mathbf{X},\mathbf{Y})$ be given by $\mathbf{X}^+$. Note that for a majorizing algorithm to be sufficiently fast, it should be easy to compute $\mathbf{X}^+$. Because the $\hat{\varphi}(\mathbf{X},\mathbf{Y})$ is always larger than or equal to the $\varphi(\mathbf{X})$, we must have $\varphi(\mathbf{X}^+) \leq \hat{\varphi}(\mathbf{X}^+,\mathbf{Y})$. This property is essential for the so-called sandwich inequality, that is, the chain

$$\varphi(\mathbf{X}^+) \leq \hat{\varphi}(\mathbf{X}^+,\mathbf{Y}) \leq \hat{\varphi}(\mathbf{Y},\mathbf{Y}) = \varphi(\mathbf{Y}) \tag{4}$$

that proves that the update $\mathbf{X}^+$ never increase the original function. For the next iteration, we simply set $\mathbf{Y}$ equal to $\mathbf{X}^+$ and compute a new majorizing function. For functions that are bounded from below or are sufficiently constrained, the majorization algorithm always gives a convergent sequence of nonincreasing function values, see, for example, Borg and Groenen (2005).

One property that we use here is that if a function consists of a sum of functions and that each of these functions can be majorized, then the sum of the majorizing functions also majorizes the original functions. For example, suppose that $\varphi(\mathbf{X}) = \sum_i \varphi_i(\mathbf{X})$ and $\varphi_i(\mathbf{X}) \leq \hat{\varphi}_i(\mathbf{X},\mathbf{Y})$ then $\varphi(\mathbf{X}) \leq \sum_i \hat{\varphi}_i(\mathbf{X},\mathbf{Y})$.

3.2 Updating the Cluster Membership

For fixed cluster centers $\mathbf{V}$, Groenen and Jajuga (2001) derive the update of the cluster memberships $\mathbf{F}$ as

$$f_{ik} = \frac{\left(d_{ik}^{2\lambda}(\mathbf{V})\right)^{-1/(s-1)}}{\sum_{l=1}^{K}\left(d_{il}^{2\lambda}(\mathbf{V})\right)^{-1/(s-1)}} \tag{5}$$

for fixed $\mathbf{V}$ and $s > 1$, see also Bezdek (1973). These memberships are derived by taking the Lagrangian function, setting the derivatives equal to zero, and solving the equations.

There are two special cases. The first one occurs if s is large. The larger s, the closer $-1/(s-1)$ approaches zero. As a consequence $\left(d_{ik}^{2\lambda}(\mathbf{V})\right)^{-1/(s-1)} \approx 1$ for all ik so that update (5) will yield $f_{ik} \approx 1/K$. Numerical accuracy can produce equal cluster memberships, even for not too large s, such as, $s = 10$. If this happens for all f_{ik}, then all cluster centers collapse into the same point and the algorithm gets stuck. Therefore, in practical applications s should be chosen quite small, say, $s \leq 2$. The second special case occurs if s approaches 1 from above. In that case, update (5) approaches the update for hard clustering, that is, setting

$$f_{ik} = \begin{cases} 1 & \text{if } d_{ik} = \min_l d_{il} \\ 0 & \text{if } d_{ik} \neq \min_l d_{il}, \end{cases} \tag{6}$$

where it is assumed that $\min_l d_{il}$ is unique.

3.3 Updating the Cluster Coordinates

We follow the majorization approach of Groenen and Jajuga (2001) for finding an update of the cluster coordinates $\mathbf{V}$ for fixed $\mathbf{F}$. Our loss function $L(\mathbf{F}, \mathbf{V})$ may be seen as a weighted sum of the λth root of squared Minkowski distances. Because the weights f_{ik}^s are nonnegative, it suffices for now to consider $d_{ik}^{2\lambda}(\mathbf{V})$, the root of squared Minkowski distances. Let us focus on the root for a moment. Groenen and Heiser (1996) proved that for root λ of a, with $0 \leq \lambda \leq 1$, $a \geq 0$ and $b > 0$, the following majorization inequality holds:

$$a^\lambda \leq (1-\lambda)b^\lambda + \lambda b^{\lambda-1}a, \tag{7}$$

with equality if $a = b$. Using (7), we can obtain the majorizing inequality

$$d_{ik}^{2\lambda}(\mathbf{V}) \quad \leq \quad (1-\lambda)d_{ik}^{2\lambda}(\mathbf{W}) + \lambda d_{ik}^{2(\lambda-1)}(\mathbf{W})d_{ik}^2(\mathbf{V}), \tag{8}$$

with $\mathbf{W}$ is the estimate of $\mathbf{V}$ from the previous iteration and we assume for the moment that $d_{ik}(\mathbf{W}) > 0$. Thus, the root λ of a squared Minkowski distance can be majorized by a constant plus a positive weight times the squared Minkowski distance.

The next step is to majorize the squared Minkowski distance. To do so, we distinguish three cases: (a) $1 \leq p \leq 2$, (b) $2 < p < \infty$, and (c) $p = \infty$.

For the case of $1 \leq p \leq 2$, Groenen and Jajuga (2001) use Hölder's inequality to prove that

$$\begin{aligned}
d_{ik}^2(\mathbf{V}) &\leq \frac{\sum_{j=1}^m (x_{ij} - v_{kj})^2 |x_{ij} - w_{kj}|^{p-2}}{d_{ik}^{p-2}(\mathbf{V})} \\
&= \sum_{j=1}^m a_{ijk}^{(1\leq p\leq 2)} (x_{ij} - v_{kj})^2, \\
&= \sum_{j=1}^m a_{ijk}^{(1\leq p\leq 2)} v_{kj}^2 - 2\sum_{j=1}^m b_{ijk}^{(1\leq p\leq 2)} v_{kj} + c_{ik}^{(1\leq p\leq 2)} \qquad (9)
\end{aligned}$$

where

$$\begin{aligned}
a_{ijk}^{(1\leq p\leq 2)} &= \frac{|x_{ij} - w_{kj}|^{p-2}}{d_{ik}^{p-2}(\mathbf{V})}, \\
b_{ijk}^{(1\leq p\leq 2)} &= a_{ijk}^{(1\leq p\leq 2)} x_{ij}, \\
c_{ik}^{(1\leq p\leq 2)} &= \sum_{j=1}^m a_{ijk}^{(1\leq p\leq 2)} x_{ij}^2.
\end{aligned}$$

For $p \geq 2$, (9) is reversed, so that it cannot be used for majorization. However, Groenen, Heiser, and Meulman (1999) have developed majorizing inequalities for squared Minkowski distances with $2 < p < \infty$ and $p = \infty$. We first look at $2 < p < \infty$. They proved that the Hessian of the squared Minkowski distance always has the largest eigenvalue smaller than $2(p-1)$. By numerical experimentation they even found a smaller maximum eigenvalue of $(p-1)2^{1/p}$ but they were unable to prove this. Knowing an upper bound of the largest eigenvalue of the Hessian is enough to derive a majorizing inequality if it is combined with the requirement of touching at the supporting point (that is, at this point the gradients of the squared Minkowski distance and the majorizing function must be equal and the same must hold for their function values).

This majorizing inequality can be derived as follows. For notational simplicity, we express the squared Minkowski distance as $d^2(\mathbf{t}) = (\sum_j |t_j|^p)^{2/p}$. The first derivative $\partial d^2(\mathbf{t})/\partial t_j$ can be expressed as $2t_j|t_j|^{p-2}/d^{p-2}(\mathbf{t})$. Knowing that the largest eigenvalue of the Hessian of $d^2(\mathbf{t})$ is bounded by $2(p-1)$, a quadratic majorizing can be found (Groenen et al., 1999) of the form

$$d^2(\mathbf{t}) \leq 4(p-1)\sum_{j=1}^m t_j^2 - 2\sum_{j=1}^m t_j b_j + c$$

with

$$b_j = 4(p-1)u_j - \frac{1}{2}\frac{\partial d^2(\mathbf{u})}{\partial u_j} = u_j \left[4(p-1) - \frac{|u_j|^{p-2}}{d^{p-2}(\mathbf{u})}\right],$$

$$c \;=\; d^2(\mathbf{u}) + 4(p-1)\sum_{j=1}^{m} u_j^2 - \sum_{j=1}^{m} u_j \frac{\partial d^2(\mathbf{u})}{\partial u_j} = 4(p-1)\sum_{j=1}^{m} u_j^2 - d^2(\mathbf{u}),$$

and $\mathbf{u}$ the known current estimate of $\mathbf{t}$. Substituting $t_j = x_{ij} - v_{kj}$ and $u_j = x_{ij} - w_{kj}$ gives the majorizing inequality

$$\begin{aligned} d_{ik}^2(\mathbf{V}) \;\leq\;& 4(p-1)\sum_{j=1}^{m}(x_{ij}-v_{kj})^2 \\ &-2\sum_{j=1}^{m}(x_{ij}-v_{kj})(x_{ij}-w_{kj})[4(p-1) - |x_{ij}-w_{kj}|^{p-2}/d_{ik}^{p-2}(\mathbf{W})] \\ &+4(p-1)\sum_{j=1}^{m}(x_{ij}-w_{kj})^2 - d_{ik}(\mathbf{W}). \end{aligned}$$

Some rewriting yields

$$d_{ik}^2(\mathbf{V}) \leq a^{(2<p<\infty)}\sum_{j=1}^{m} v_{kj}^2 - 2\sum_{j=1}^{m} b_{ijk}^{(2<p<\infty)} v_{kj} + \sum_{j=1}^{m} c_{ijk}^{(2<p<\infty)}, \qquad (10)$$

where

$$\begin{aligned} a^{(2<p<\infty)} &= 4(p-1) \\ b_{ijk}^{(2<p<\infty)} &= a^{(2<p<\infty)} w_{kj} - (x_{ij}-w_{kj})|x_{ij}-w_{kj}|^{p-2}/d_{ik}^{p-2}(\mathbf{W}) \\ c_{ik}^{(2<p<\infty)} &= a^{(2<p<\infty)}\sum_{j=1}^{m} w_{kj}^2 - d_{ik}^2(\mathbf{W}) + 2\sum_{j=1}^{m} x_{ij}(x_{ij}-w_{kj})|x_{ij}-w_{kj}|^{p-2}/d_{ik}^{p-2}(\mathbf{W}). \end{aligned}$$

If p gets larger, $a^{(2<p<\infty)}$ also becomes larger, thereby making the majorizing function steeper. As a result, the steps taken per iteration will be smaller. For the special case of $p = \infty$, Groenen et al. (1999) also provided a majorizing inequality. This one can be (much) faster than using (10) with a large p. It depends on the difference between the two largest values of $|x_{ij} - w_{kj}|$ over the different j.

Let us for the moment focus on $d^2(\mathbf{t})$ again. And let φ_j be an index that orders the values $|t_j|$ decreasingly, so that $|t_{\varphi_1}| \leq |t_{\varphi_2}| \leq \ldots \leq |t_{\varphi_m}|$. The majorizing function for $p = \infty$ becomes

$$d^2(\mathbf{t}) \;\leq\; a\sum_{j=1}^{m} t_j^2 - 2\sum_{j=1}^{m} t_j b_j + c$$

with

$$a \;=\; \begin{cases} \dfrac{|u_{\varphi_1}|}{|u_{\varphi_1}| - |u_{\varphi_2}|} & \text{if } |u_{\varphi_1}| - |u_{\varphi_2}| > \varepsilon \\ \dfrac{\epsilon + |u_{\varphi_1}|}{\varepsilon} & \text{if } |u_{\varphi_1}| - |u_{\varphi_2}| \leq \varepsilon \end{cases}$$

$$
\begin{aligned}
b_j &= \begin{cases} a\dfrac{|u_{\varphi_2}|u_j}{|u_j|} & \text{if } j = \varphi_1 \\ au_j & \text{if } j \neq \varphi_1 \end{cases} \\
c &= d^2(\mathbf{u}) + 2\sum_j u_j b_j - a\sum_j u_j^2.
\end{aligned}
$$

Note that the definition of a for $|u_{\varphi_1}| - |u_{\varphi_2}| \leq \varepsilon$ takes care of ill conditioning, that is, values of a getting too large. Strictly speaking, majorization is not valid anymore, but for small enough ϵ the monotone convergence is retained.

Backsubstitution of $t_j = x_{ij} - v_{kj}$ and $u_j = x_{ij} - w_{kj}$ gives the majorizing inequality

$$
\begin{aligned}
d_{ik}^2(\mathbf{V}) &\leq a_{ik}^{(p=\infty)}\sum_j (x_{ij} - v_{kj})^2 - 2\sum_j (x_{ij} - v_{kj})b_{ijk}^{(p=\infty)} + c_{ik}^{(p=\infty)} \\
&= a_{ik}^{(p=\infty)}\sum_j v_{kj}^2 - 2\sum_j v_{kj}b_{ijk}^{(p=\infty)} + c_{ik}^{(p=\infty)} \qquad (11)
\end{aligned}
$$

where

$$
\begin{aligned}
a_{ik}^{(p=\infty)} &= \begin{cases} \dfrac{|x_{i\varphi_1} - w_{k\varphi_1}|}{|x_{i\varphi_1} - w_{k\varphi_1}| - |x_{i\varphi_2} - w_{k\varphi_2}|} & \text{if } |x_{i\varphi_1} - w_{k\varphi_1}| - |x_{i\varphi_2} - w_{k\varphi_2}| > \varepsilon, \\ \dfrac{\varepsilon + |x_{i\varphi_1} - w_{k\varphi_1}|}{\varepsilon} & \text{if } |x_{i\varphi_1} - w_{k\varphi_1}| - |x_{i\varphi_2} - w_{k\varphi_2}| \leq \varepsilon, \end{cases} \\
b_{ijk}^{(p=\infty)} &= \begin{cases} a_{ik}^{(p=\infty)}\left[x_{ij} - \dfrac{|x_{i\varphi_2} - w_{k\varphi_2}|(x_{i\varphi_1} - w_{k\varphi_1})}{|x_{i\varphi_1} - w_{k\varphi_1}|}\right] & \text{if } j = \varphi_1, \\ a_{ik}^{(p=\infty)}w_{kj} & \text{if } j \neq \varphi_1, \end{cases} \\
c_{ik}^{(p=\infty)} &= d_{ik}^2(\mathbf{W}) - 2\sum_j b_{ijk}^{(p=\infty)}(x_{ij} - w_{kj}) - \sum_j a_{ik}^{(p=\infty)}w_{kj}^2 + 2\sum_j a_{ik}^{(p=\infty)}x_{ij}^2.
\end{aligned}
$$

Recapitulating, the loss function is a weighted sum of the root of the squared Minkowski distance. The root can be majorized by (8) that yields a function of squared Minkowski distances. For the case $1 \leq p \leq 2$, (9) shows how the squared Minkowski distance can be majorized by a quadratic function in $\mathbf{V}$ (see, Figure 1), (10) how this can be done for $2 < p < \infty$, and (11) for $p = \infty$. These results can be combined to obtain the following majorizing function for $L(\mathbf{F}, \mathbf{V})$, that is,

$$
L(\mathbf{F}, \mathbf{V}) \leq \lambda\sum_{j=1}^m\sum_{k=1}^K v_{kj}^2\sum_{i=1}^n a_{ijk} - 2\lambda\sum_{j=1}^m\sum_{k=1}^K v_{kj}\sum_{i=1}^n b_{ijk} + c + \sum_{i=1}^n\sum_{k=1}^K c_{ik}, \qquad (12)
$$

where

$$
a_{ijk} = \begin{cases} f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W})a_{ijk}^{(1\leq p\leq 2)} & \text{if } 1 \leq p \leq 2 \\ f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W})a^{(2<p<\infty)} & \text{if } 2 < p < \infty \\ f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W})a_{ik}^{(p=\infty)} & \text{if } p = \infty \end{cases}
$$

$$
\begin{aligned}
b_{ijk} &= \begin{cases} f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W}) b_{ijk}^{(1\leq p\leq 2)} & \text{if } 1 \leq p \leq 2 \\ f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W}) b_{ijk}^{(2<p<\infty)} & \text{if } 2 < p < \infty \\ f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W}) b_{ijk}^{(p=\infty)} & \text{if } p = \infty \end{cases} \\
c_{ik} &= \begin{cases} f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W}) c_{ik}^{(1\leq p\leq 2)} & \text{if } 1 \leq p \leq 2 \\ f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W}) c_{ik}^{(2<p<\infty)} & \text{if } 2 < p < \infty \\ f_{ik}^s d_{ik}^{2(\lambda-1)}(\mathbf{W}) c_{ik}^{(p=\infty)} & \text{if } p = \infty \end{cases} \\
c &= \sum_{i=1}^{n} \sum_{k=1}^{K} f_{ik}^s (1-\lambda) d_{ik}^{2\lambda}(\mathbf{W}).
\end{aligned}
$$

It is easily recognized that (12) is a quadratic function in the cluster coordinate matrix $\mathbf{V}$ that reaches its minimum for

$$
v_{kj}^{+} = \frac{\sum_{i=1}^{n} b_{ijk}}{\sum_{i=1}^{n} a_{ijk}}. \tag{13}
$$

3.4 The Majorization Algorithm

The majorization algorithm can be summarized as follows.

1. Given a data set $\mathbf{X}$. Set $0 \leq \lambda \leq 1$, $1 \leq p \leq \infty$, and $s \geq 1$. Choose ε, a small positive constant.

2. Set the membership grades $\mathbf{F} = \mathbf{F}^0$ with $0 \leq f_{ik}^0 \leq 1$ and $\sum_{k=1}^{K} f_{ik}^0 = 1$ and the cluster coordinate matrix $\mathbf{V} = \mathbf{V}_0$. Compute $L_{\text{prev}} = L(\mathbf{F}, \mathbf{V})$.

3. Update $\mathbf{F}$ by (5) if $s > 1$ or by (6) if $s = 1$.

4. Set $\mathbf{W} = \mathbf{V}$. Update $\mathbf{V}$ by (13).

5. Stop if $(L_{\text{prev}} - L(\mathbf{F}, \mathbf{V}))/L(\mathbf{F}, \mathbf{V})) < \varepsilon$.

6. Set $L_{\text{prev}} = L(\mathbf{F}, \mathbf{V})$ and go to Step 3.

4 Internet Attitudes

To show how fuzzy clustering can be used in practice, we apply it to an empirical data set. Our data set is based on a questionnaire on attitudes towards the Internet. It consists of evaluations of 22 statements about the Internet by 194 students gathered around 2002 before the wide availability

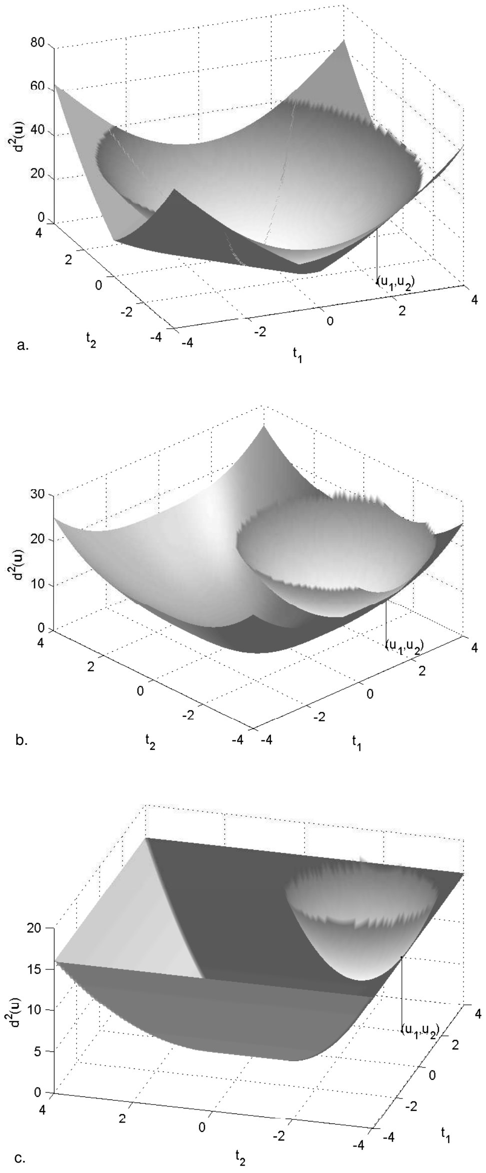


Figure 1: The original function $d^2(\mathbf{t})$ and the majorizing functions for $p = 1$, $p = 3$, and $p = \infty$ using supporting point $\mathbf{u} = [2, -3]'$.

Table 2: Results of fuzzy clustering for internet data set using $K = 3$.

				Cluster volumes		
λ	s	p	$L(\mathbf{F}, \mathbf{V})$	Cluster 1	Cluster 2	Cluster 3
.5	1.2	1	4087	1.315	1.411	1.267
.5	1.2	2	1075	1.227	1.326	1.268
.5	1.2	∞	421	1.355	1.408	1.341
.5	1.5	1	2983	0.965	0.965	0.979
.5	1.5	2	773	0.920	0.920	0.920
.5	1.5	∞	310	0.973	0.973	0.973
1	1.2	1	103115	1.281	1.363	1.236
1	1.2	2	7358	1.177	1.328	1.173
1	1.2	∞	1123	1.257	1.588	1.284
1	1.5	1	83101	0.965	0.997	0.979
1	1.5	2	5587	0.920	0.920	0.920
1	1.5	∞	878	0.977	0.977	0.977

of broadband Internet access. The statements were evaluated using a seven-point Likert scale, ranging from 1 (completely disagree) to 7 (completely agree).[1]

The respondents are clustered using the fuzzy clustering algorithm to study their attitudes towards the Internet. We use $K = 3$. The convergence criterion ε of the majorization algorithm was set to 10^{-8}. The monotone convergence of the majorization algorithm generally leads to a local minimum. However, depending on the data and the different settings of p, s, and λ several local minima may exist. Therefore, in every analysis, we applied 10 random starts and report the best one. We tried three different values of p $(1, 2, \infty)$ to examine the cluster shape, two values of s (1.2, 1.5) to study the sensitivity for the fuzziness parameter s, and two values for λ (.5, 1.0) to check the sensitivity for outliers.

Table 2 shows some results for this data set using different values for λ, s, and p. The final value of the loss function and the volumes of the three clusters are calculated in every instance. As there is no natural standardization for $L(\mathbf{F}, \mathbf{V})$, the values can only be used to check for local minima within a particular choice of λ, s, and p.

The labelling problem of clusters refers to possible permutations of the clusters among different runs. To avoid this problem, we took the $\mathbf{V}$ obtained

[1] We would like to thank Peter Verhoef for making these data available.

by $\lambda = 1, p = 1$, and $s = 1.2$ as a target solution $\mathbf{V}^*$ and tried all permutation matrices $\mathbf{P}$ of the rows of $\mathbf{V}$ (with $\mathbf{V}^{(\text{Perm})} = \mathbf{PV}$) for other combinations of λ, p, and s and choose the one that minimizes the sum of the squared residuals

$$\sum_k \sum_j (v^*_{kj} - v^{(\text{Perm})}_{kj})^2 = \|\mathbf{V}^* - \mathbf{PV}\|^2. \tag{14}$$

The permutation $\mathbf{P}$ that minimizes (14) is also applied to the cluster memberships, so that $\mathbf{F}^{(\text{Perm})} = \mathbf{FP}'$. By using this strategy, we assume that the clusters are the same among the different analyses.

To measure the size of a cluster, we consider its volume by computing the cluster covariance matrix with elements

$$\mathbf{G}_k = \frac{\sum_{i=1}^n f^s_{ik}(\mathbf{x}_i - \mathbf{v}_k)'(\mathbf{x}_i - \mathbf{v}_k)}{\sum_{i=1}^n f^s_{ik}},$$

where $\mathbf{x}_i$ is the $1 \times j$ row vector of row i of $\mathbf{X}$ and $\mathbf{v}_k$ row i of $\mathbf{V}$. Then, as a measure of the volume of cluster k one can use $\det(\mathbf{G}_k)$. However, we take $\det(\mathbf{G}_k)^{1/m}$, which can be interpreted as the geometric mean of the eigenvalues of $\mathbf{G}_k$ and has the advantage that it is not sensitive to m. Table 2 shows that for $s = 1.5$ the cluster volumes are all the same with a slight difference among the clusters of $p = 1$. For $s = 1.5$, Cluster 2 is generally the largest and the other two have about the same size. The more robust setting of $\lambda = .5$, generally shows slightly larger clusters, but the effect does not seem large. Therefore, outliers do not seem to be a problem of this data set.

To interpret the clusters, we have to look at $\mathbf{V}$. As it is impossible to show the clusters in a 22-dimensional space, they are represented by parallel coordinates (Inselberg, 1981, 1997). Every cluster k defines a line through the cluster centers v_{kj}, see Figure 2 for $s = 1.2$ and $\lambda = 1$. Note that the order of the variables is unimportant. This figure can be interpreted by considering the variables that have different scores for the clusters. The patterns for $p = 1, 2$, and ∞ are similar and $p = 1$ shows them the clearest.

For $p = 1$ and $\lambda = 1$, each cluster center is a (weighted) median of a cluster. Because all elements of the internet data set are integers, the cluster centers necessarily have integer values. The left most panel in Figure 2 shows the parallel coordinates for $p = 1$. The solid line represents Cluster 1 and is characterized by respondents saying that the Internet is easy, safe, addictive, and who seem to form an active user community (positive answers to variables 16 to 22). However, the strongest difference of Cluster 1 to the others is given by their total rejection of regulation of content on the Internet. We call this

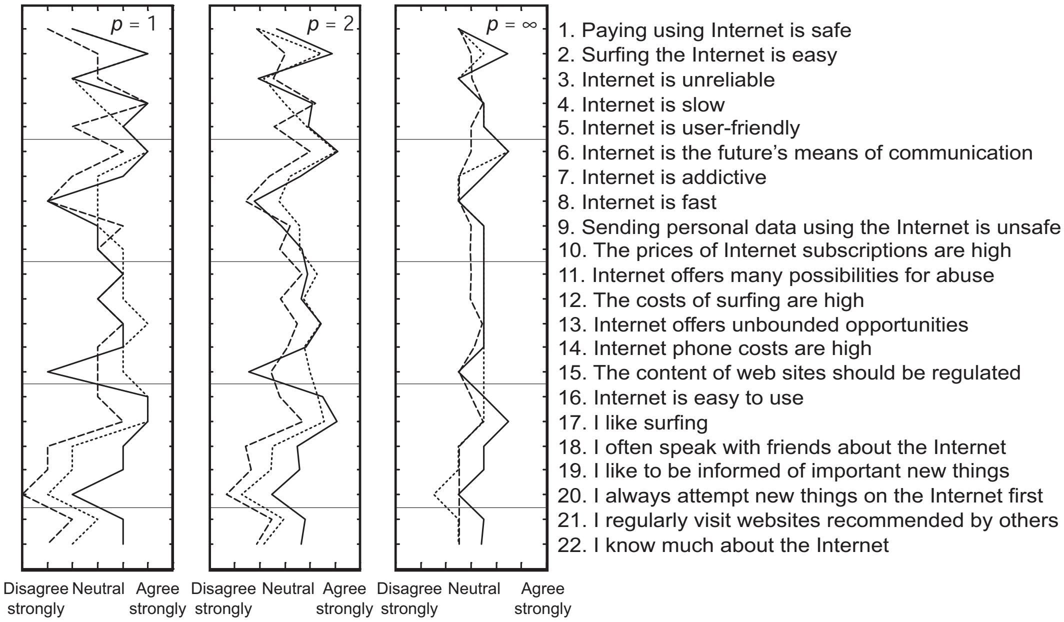
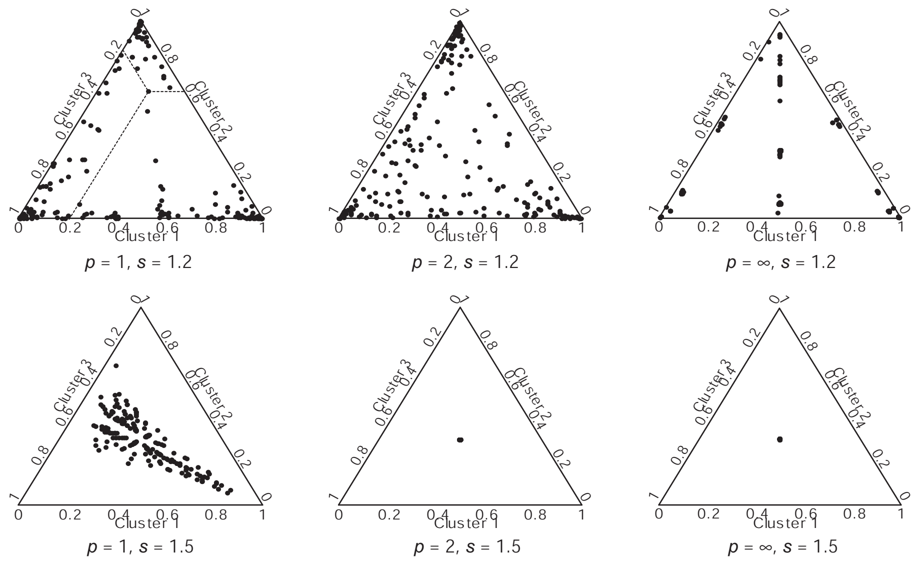


Figure 2: Parallel coordinates representation of clusters with $\lambda = 1$, $p = 1$, and $s = 1.2$. The lines correspond to clusters 1 (solid line), 2 (dashed line), and 3 (dotted line).

cluster the experts. Cluster 2 (the dashed line) refers to respondents that are not active users (negative answers to variables 18 to 22), find the Internet not user friendly, unsafe to pay, not addictive, and they are neutral on the issue of regulation of the content of websites. This cluster is called the novices. Cluster 3 looks in some respects like Cluster 1 (surfing is easy, paying is not so safe) but those respondents do not find the Internet addictive, are neutral on the issue of the speed of the Internet connection, and seem to be not such active users as those of Cluster 1. They are mostly characterized by finding the costs of Internet high and allowing for some content regulation. This cluster represents the cost-aware Internet user.

As we are dealing with three clusters and the cluster memberships sum to one, they can be plotted in a triangular 2D scatterplot—called a triplot—as in Figure 3. To reconstruct the fuzzy memberships from this plot, the following should be done. For Cluster 1, one has to project a point along a line parallel to the axis of Cluster 3 onto the axis of Cluster 1. We have done this with dotted lines for respondent 112 for the case of $p = 1$, $s = 1.2$, and $\lambda = 1$. We can read from the plot that this respondent has fuzzy memberships f_{i1} of about .20. Similarly, for Cluster 2, we have to draw a line horizontally (parallel to the axis of Cluster 1) and project it onto the axis of Cluster 2



p = 1, s = 1.2

p = 2, s = 1.2

p = ∞, s = 1.2

p = 1, s = 1.5

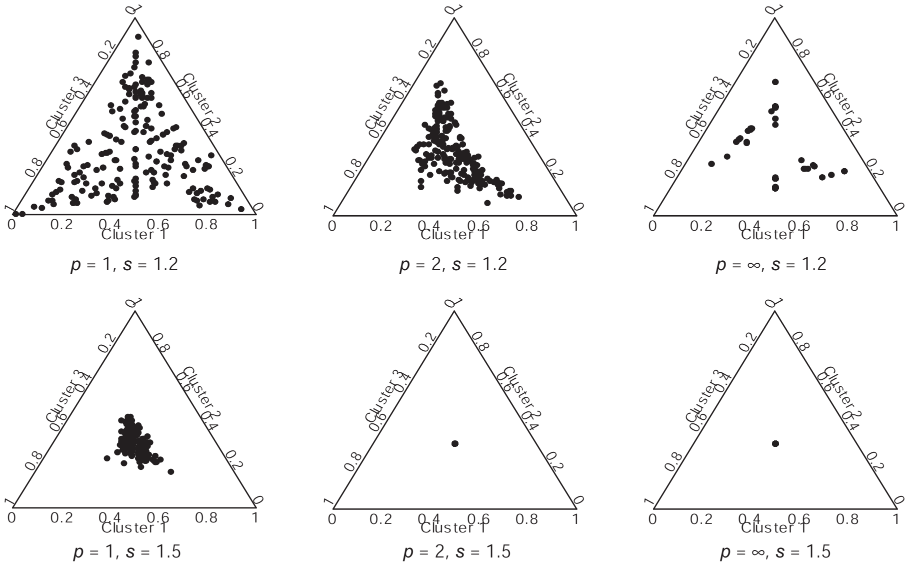
p = 2, s = 1.5

p = ∞, s = 1.5

Figure 3: Triplot showing the cluster membership in $\mathbf{F}$ for each respondent for $s = 1.2, 1.5$, $\lambda = 1$, and $p = 1, 2, \infty$.

showing f_{i2} of about .65. Finally, f_{i3} is obtained by projecting onto the axis of Cluster 3 along a line parallel to Cluster 2, yielding f_{i3} of about .15. In four decimals, these values are .2079, .6457, and .1464. Thus, a point located close to a corner implies that this respondent has almost solely assigned to this cluster. Also, a point exactly in the middle of the triangle implies an equal memberships of 1/3 to all three clusters. Finally, points that are on a straight line from a corner orthogonal to a cluster axis have equal cluster memberships of two clusters. For the case of $p = \infty$, Figure 3 shows a vertical line (starting in Cluster 2 and orthogonal to the Cluster 1 axis) implying that the memberships for Clusters 1 and 3 are the same for those respondents.

For the choice $p = 2$ and $s = 1.5$ and $p = 2$ or ∞, all clusters centers are in close proximity to each other in the center. In other words, all fuzzy memberships are about 1/3 and consequently the three cluster centers are the same. Therefore, $s = 1.5$ is too large for $p = 2$ or ∞. This finding is an indication of overlapping clusters. For a value of $s = 1.2$, the triplot for $p = 1$ shows more pronounced clusters because most of the respondents are in the corners. For $p = 2$ and $s = 1.2$, the memberships are more evenly distributed over the triangle although many respondents are still located in the corners. For $p = \infty$ and $s = 1.2$, some respondents are on the vertical line (combining equal memberships to Clusters 1 and 3 for varying membership of Cluster 2). The points that are located close to the Cluster 1 axis at .5 have a membership of .5 for Clusters 1 and 3, those close to .5 at the Cluster



p = 1, s = 1.2 p = 2, s = 1.2 p = ∞, s = 1.2

p = 1, s = 1.5 p = 2, s = 1.5 p = ∞, s = 1.5

Figure 4: Triplot showing the cluster membership in $\mathbf{F}$ for each respondent for $s = 1.2, 1.5$, $\lambda = .5$, and $p = 1, 2, \infty$..

2 axis have .5 for Clusters 1 and 2, those close to the Cluster 3 axis at .5 have .5 for Clusters 2 and 3.

For the robust case of $\lambda = 1/2$, the triplots of the fuzzy memberships are given in Figure 4. One of the effects of setting $\lambda = 1/2$ seems to be that the f_{ik} are more attracted to the center and, hence, respondents are less attracted to a single cluster than in the case of $\lambda = 1$. Again, for $s = 1.5$ and $p = 2$ and ∞, all clusters merge into one cluster and the parallel coordinates plots of the clusters would show a single line. For $s = 1.2$, the parallel coordinates plots of the clusters resemble Figure 2 reasonably well. For $s = 1.2$ and $p = 2$, the lines in the parallel coordinates plot are closer together than for $\lambda = 1$.

For this data set, the clusters cannot be well separated because for a relatively small s of 1.5, the clusters coincide (except for $p = 1$). The cluster centers seem to be better separated when p is small, especially for $p = 1$. The fuzziness parameter s needs to be chosen small in this data set to avoid clusters collapsing into a single cluster. The effect of varying λ seems to be that the cluster memberships are less extreme for $\lambda = 1/2$ than for $\lambda = 1$.

5 Conclusions

We have considered objective function based fuzzy clustering algorithms using a generalized distance function. In particular, we have studied the extension of the fuzzy c-means algorithm to the case of the parametric Minkowski distance function and to the case of the root of the squared Minkowski distance function. We have derived the optimality conditions for the membership values from the Lagrangian function. For cluster centers, however, we have used iterative majorization to derive the optimality conditions. One of the advantages of iterative majorization is that it is a guaranteed descent algorithm, so that every iteration reduces the objective function until convergence is reached. We have derived suitable majorization functions for the distance function that we study. Extending results from Groenen and Jajuga (2001), we have given a majorization algorithm for *any* Minkowski distance with Minkowski parameter greater than (or equal to) 1. This extension also included the case of the L_∞-distance and the roots of the squared Minkowski distance.

By adapting the Minkowski parameter p, the user influences the distance function to take specific cluster shapes into account. We have also introduced an additional parameter λ for computing the roots of the squared Minkowski distance. This parameter can be used to protect the clustering algorithm against outliers. Hence, more robust clustering results can be obtained.

We have illustrated some key aspects of the behaviour of our algorithm using empirical data regarding attitudes about the Internet. With this particular data set, we have observed extremely overlapping clusters, already with a fuzziness parameter of $s = 1.5$. This finding deviates from the general practice in fuzzy clustering, where this parameter is often selected equal to 2. Apparently, the choice of s and p has to be done with some care for a given data set.

References

Babuška, R. (1998). *Fuzzy modeling for control.* Boston, MA: Kluwer Academic Publishers.

Baraldi, A., & Blonda, P. (1999a, December). A survey of fuzzy clustering algorithms for pattern recognition — part I. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, *29*(6), 778–785.

Baraldi, A., & Blonda, P. (1999b, December). A survey of fuzzy clustering algorithms for pattern recognition — part I. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, *29*(6), 786–801.

Bezdek, J. C. (1973). *Fuzzy mathematics in pattern classification.* Unpublished doctoral dissertation, Cornell University, Ithaca.

Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981a). Detection and characterization of cluster substructure, II. fuzzy c-varieties and convex combinations thereof. *SIAM Journal of Applied Mathematics*, *40*(2), 358–372.

Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981b). Detection and characterization of cluster substructure, I. linear structure: fuzzy c-lines. *SIAM Journal of Applied Mathematics*, *40*(2), 339–357.

Bezdek, J. C., & Pal, S. K. (1992). *Fuzzy models for pattern recognition.* New York: IEEE Press.

Bobrowski, L., & Bezdek, J. C. (1991). c-Means clustering with the l_1 and l_∞ norms. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*, 545–554.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York: Springer.

Dave, R. N. (1991, November). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, *12*(11), 657–664.

De Leeuw, J. (1994). Block relaxation algorithms in statistics. In H.-H. Bock, W. Lenski, & M. M. Richter (Eds.), *Information systems and data analysis* (pp. 308–324). Berlin: Springer.

Dodge, Y., & Rousson, V. (1998). Multivariate L_1 mean. In A. Rizzi, M. Vichi, & H. Bock (Eds.), *Advances in data science and classification* (pp. 539–546). Berlin: Springer.

Dunn, J. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57.

Groenen, P. J. F., & Heiser, W. J. (1996). The tunneling method for global optimization in multidimensional scaling. *Psychometrika*, *61*, 529–550.

Groenen, P. J. F., Heiser, W. J., & Meulman, J. J. (1999). Global optimization in least-squares multidimensional scaling by distance smoothing. *Journal of Classification*, *16*, 225–254.

Groenen, P. J. F., & Jajuga, K. (2001). Fuzzy clustering with squared Minkowski distances. *Fuzzy Sets and Systems*, *120*, 227–237.

Gustafson, D. E., & Kessel, W. C. (1979). Fuzzy clustering with a fuzzy covariance matrix. In *Proc. ieee cdc* (pp. 761–766). San Diego, USA.

Hathaway, R. J., Bezdek, J. C., & Hu, Y. (2000, October). Generalized fuzzy c–means clustering strategies using l_p norm distances. *IEEE Transactions on Fuzzy Systems*, *8*(5), 576–582.

Heiser, W. J. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In W. J.

Krzanowski (Ed.), *Recent advances in descriptive multivariate analysis* (pp. 157–189). Oxford: Oxford University Press.

Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). *Fuzzy cluster analysis: methods for classification, data analysis and image recognition.* New York: Wiley.

Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, *39*, 30–37.

Inselberg, A. (1981). *N-dimensional graphics, part I: Lines and hyperplanes* (Tech. Rep. No. G320-2711). Los Angeles (CA): IBM Los Angeles Scientific Center.

Inselberg, A. (1997). Multidimensional detective. In *Proc. ieee symp. information visualization* (p. 100-107).

Jajuga, K. (1991). L_1-norm based fuzzy clustering. *Fuzzy Sets and Systems*, *39*, 43–50.

Kaymak, U., & Setnes, M. (2002, December). Fuzzy clustering with volume prototypes and adaptive cluster merging. *IEEE Transactions on Fuzzy Systems*, *10*(6), 705–712.

Kiers, H. A. L. (2002). Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. *Computational Statistics and Data Analysis*, *41*, 157–170.

Lange, K., Hunter, D. R., & Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, *9*, 1–20.

Ruspini, E. (1969). A new approach to clustering. *Information and Control*, *15*, 22–32.

Yang, M.-S. (1993). A survey of fuzzy clustering. *Mathematical and Computer Modelling*, *18*(11), 1–16.