Rock Mech. Rock Engng. (1999) 32 (1), 1–27

Rock Mechanics and Rock Engineering

On Distance Measures for the Fuzzy K-means Algorithm for Joint Data

By

R. E. Hammah and **J. H. Curran**

Rock Engineering Group, Department of Civil Engineering, University of Toronto, Canada

Summary

The analysis of data collected on rock discontinuities often requires that the data be separated into joint sets or groups. A statistical tool that facilitates the automatic identification of groups of clusters of observations in a data set is cluster analysis. The fuzzy *K*-means cluster technique has been successfully applied to the analysis of joint survey data. As is the case with all clustering algorithms, the results of an analysis performed with the fuzzy *K*-means algorithm for discontinuity data are highly dependent on the distance metric employed in the analysis. This paper explores the significant issues surrounding the choice and use of various distance measures for clustering joint survey data. It also proposes an analogue of the Mahalanobis distance norm (used for data in Euclidean space) for clustering spherical data. Sample applications showing the greater flexibility and power of the new distance measure over the originally proposed distance metric for spherical data are given in the paper.

1. Introduction

In the analysis of discontinuity data collected from joint surveys the rock mechanics expert seeks to classify the data according to similarities observed in the recorded attributes or characteristics. Often there is little existing *a priori* information as to the number of fracture sets present in the data set, and also as to how the discontinuities are distributed among the joint sets. It is therefore up to the data analyst to discover the structure of the data (the interrelationships between the different objects or patterns in the data set) using exploratory tools of data analysis. One such tool, which is capable of grouping data into classes without a priori information on the classes, is cluster analysis.

Clustering, in the language of pattern recognition, is an unsupervised learning procedure for classifying objects or patterns. Algorithms for performing cluster analysis can be grouped into two broad categories – hierarchical clustering methods, and partitional clustering techniques. Hierarchical methods seek to construct a hierarchy of the relationships between the pattern vectors in a data set.

The end result is a hierarchy of nested partitions, which can be pictorially viewed as a tree. Partitional clustering methods on the other hand classify the pattern vectors in a data set into a number of mutually exclusive categories.

Each of the two main categories of clustering techniques can be further broken down into several sub-groups, but this paper shall focus on a particular class of partitional clustering algorithms – the class of fuzzy *K*-means algorithms. Fuzzy *K*-means algorithms have been successfully implemented for solving problems in several research fields including medical imaging, computer vision and market segmentation. The authors of this paper proposed a variation of it for tackling the exploratory analysis of joint survey data in (Hammah and Curran, 1998a).

Pattern vectors (observations or objects) that belong to the same cluster possess similar attributes and therefore occupy a particular region of a pattern space. Because different clusters of vectors lie in different regions of the pattern space, the distance between pattern vectors serves as a measure of association between vectors. Clustering algorithms partition data into clusters based on these measures of association or similarity between vectors in a data set. As a result one of the major issues to consider in choosing of constructing a clustering algorithm is how distances are going to be measured. This step is critical to the whole clustering process because the output of a clustering algorithm will only be as meaningful as the input distances and similarities (Everitt, 1980). It is of great importance, therefore, to thoroughly explore all the issues surrounding the choice of one distance measure or the other.

The present paper shall cover the topics of pattern spaces in which the observations in a data set are believed to be embedded, namely, the definition and different types of distance measures (metrics), the correlation between variables in a data set and its effect on clustering, and the geometric significance of distance metrics. Also a new distance measure for fuzzy *K*-means clustering for spherical data shall be introduced.

2. Overview of Fuzzy *K*-means Clustering Algorithm for Discontinuity Data

The importance of cluster analysis to the delineation of discontinuity (joint or fracture) data into sets was recognized in the rock mechanics community several years ago. It was acknowledged that the use of such a statistical technique would introduce objectivity into discontinuity data analysis. One of the earliest clustering tools for separating discontinuities into sets based on orientations was supplied by Shanley and Mahtab (1976). Mahtab and Yegulalp (1982) subsequently made improvements to the algorithm. This algorithm has been widely used in rock discontinuity analysis. One of its principal weaknesses, however, is that it cannot accommodate non-orientation attributes of discontinuities in the delineation of sets.

Very recently Dershowitz et al. (1996) developed a stochastic algorithm for clustering discontinuities. This new algorithm defines clusters as statistically homogeneous groups of data, and has the capability to include non-orientation discontinuity properties in the process of identifying sets. In the algorithm, probability distributions are defined for each of the variables in the separate clusters. The

parameters of these distributions are modified iteratively until each set is statistically homogeneous. Individual observations are then assigned to sets based on their probabilities of belonging to the sets.

The stochastic algorithm of Dershowitz et al. (1996) relies on the numerical integration of all the probability distributions defined for the variables in the different sets. A more computationally attractive alternative is offered by the fuzzy *K*-means algorithm described in more detail below. All fuzzy cluster algorithms rely on elements of the fuzzy set theory first proposed by Zadeh (1965). In fuzzy set theory an object can belong to more than one set, but with varying degrees of membership to the sets. The degree of membership of an object to a set is based on the certainty of the object belonging to the set. The greater the certainty of an object belonging to a set, the closer is its membership degree to one. Fuzzy set theory makes it possible for uncertainty to be accounted for in a natural and realistic manner during data analysis (Bezdek, 1981). Using the idea of fuzzy sets, Ruspini (1969) developed an algorithm for clustering data using an objective function. Almost all subsequent fuzzy clustering methods developed can trace their roots to Ruspini's technique (Pal and Bezdek, 1992).

Fuzzy *K*-means algorithms have seen substantial growth in popularity among researchers in diverse fields, because of their versatility and ease of adaptation to the needs of different research communities. Bezdek (1981) provides an excellent treatise on the family of fuzzy *K*-means methods. Several of the classical works that have helped drive the development of fuzzy clustering methods can be found in the volume edited by Bezdek and Pal (1992).

The earliest work, known to the authors, on the application of fuzzy *K*-means clustering to the analysis of joint data was by Harrison (1992). However, the conventional fuzzy algorithm used by Harrison did not take into account the specific nature of discontinuity data. As will be discussed later in this paper, one of the fundamental problems of the conventional fuzzy *K*-means algorithm in its application to discontinuity data analysis is the difficulty it encounters in dealing with discontinuity orientations. In Hammah and Curran (1998a) modifications were made to the fuzzy *K*-means algorithm that allowed it to overcome the shortcomings. These modifications included a different distance norm for joint orientations, and a novel approach for computing the centroids (means) of clusters of orientation data.

One of the distinguishing features of discontinuity data is that part of the information is spherical data (joint orientation). The measurement of distances between points on the surface of a unit sphere cannot be treated the same way as points in Euclidean space. Another distinctive feature of joint data is the presence of variables, which are of different kinds. For example, joint spacing is very different from joint orientation in that the former is measured in Euclidean space, while the latter is spherical (or directional) data. Some of the data recorded on joints are quantitative and others qualitative, and this feature further complicates the picture. The use of an algorithm that is able to cope with the heterogeneous nature of survey data is one of the attributes that sets apart this algorithm from some of the previous tools for delineating fractures.

The fuzzy *K*-means algorithm partitions a data set of N objects or pattern

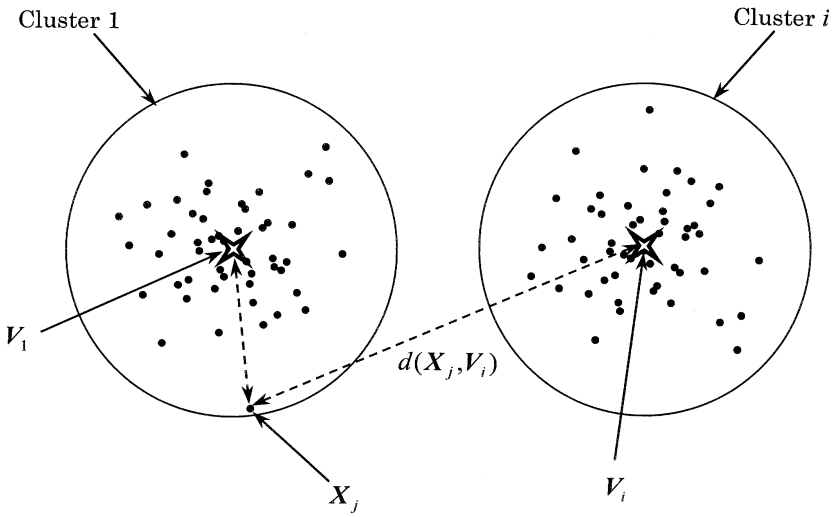

Fig. 1. Two-dimensional example illustrating the geometric meaning of cluster centroids and distances of an observation from these centroids

vectors into K clusters or groups ($K \leq N$). This partitioning is achieved by minimizing an objective function.

$$J_m(U, V) = \sum_{j=1}^{N} \sum_{i=1}^{K} (u_{ij})^m \, d^2(X_j, V_i); \quad K \leq N, \tag{1}$$

using an iterative procedure.

The quantity, $d^2(\boldsymbol{X}_j, \boldsymbol{V}_i)$, is the distance measured from a pattern vector, $\boldsymbol{X}_j$, ($j = 1 \ldots N$), to the prototype or centroid, $\boldsymbol{V}_i$, of the i-th cluster ($i = 1 \ldots K$). The prototype of a cluster is defined as the geometric mean of the pattern vectors belonging to that cluster. A geometrical representation of the cluster centroids and distances from observations to centroids is given in Fig. 1. From the objective function (Eq. 1) and the illustration in Fig. 1, it can be seen that the fuzzy K-means algorithm identifies clusters by looking for high-density regions in a pattern space. In regions of high density, the distances between the centroid of a cluster and observations belonging to the cluster are minimal.

For vectors in $\boldsymbol{R}^p$ space (P-dimensional Euclidean space), new cluster centroids are computed using the formula:

$$\hat{\boldsymbol{V}}_i = \frac{\sum_{j=1}^{N} (u_{ij})^m \boldsymbol{X}_j}{\sum_{j=1}^{N} (u_{ij})^m}. \tag{2}$$

Because spherical data has its specific nature, new prototypes for it cannot be computed using formula (2) (see Hammah and Curran, 1998a). The first step in

determining new prototypes for spherical data (orientation information) involves the computation of a modified orientation matrix for a cluster i (Hammah and Curran, 1998a):

$$\boldsymbol{S}_i^* = \frac{1}{\sum_{j=1}^{N}(u_{ij})^m} \begin{bmatrix} \sum_{j=1}^{N}(u_{ij})^m x_j x_j & \sum_{j=1}^{N}(u_{ij})^m x_j y_j & \sum_{j=1}^{N}(u_{ij})^m x_j z_j \\ \sum_{j=1}^{N}(u_{ij})^m x_j y_j & \sum_{j=1}^{N}(u_{ij})^m y_j y_j & \sum_{j=1}^{N}(u_{ij})^m y_j z_j \\ \sum_{j=1}^{N}(u_{ij})^m x_j z_j & \sum_{j=1}^{N}(u_{ij})^m y_j z_j & \sum_{j=1}^{N}(u_{ij})^m z_j z_j \end{bmatrix}, \tag{3}$$

where (x_j, y_j, z_j) are the direction cosines of the vector $\boldsymbol{X}_j$.

The eigenanalysis of this matrix yields three eigenvalues τ_{i1}, τ_{i2} and τ_{i3} arranged such that $\tau_{i1} < \tau_{i2} < \tau_{i3}$, and three corresponding normalized eigenvectors $\vec{\xi}_{i1}$, $\vec{\xi}_{i2}$ and $\vec{\xi}_{i3}$. The eigenvector, $\vec{\xi}_{i3}$, corresponding to the largest eigenvalue, τ_{i3}, is the desired updated prototype, $\hat{\boldsymbol{V}}_i$, for the i-th cluster, i.e.

$$\hat{\boldsymbol{V}}_i = \vec{\xi}_{i3}. \tag{4}$$

It was proven and demonstrated in Hammah and Curran (1998a) that this approach for computing prototypes correctly deals with clusters that contain antipodal vectors (sets which wrap between upper and lower hemispheres), because it always determines cluster centroids to lie within the acute angles between vectors.

The fuzzy K-means algorithm for joint delineation starts off with K randomly generated initial prototypes (Hammah and Curran, 1998a). The generation of the K initial prototypes is such that no *a priori* information on cluster structure is needed at all. The selection of initial prototypes is then followed by the computation of distances of the pattern vectors in the data set from the initial cluster centroids.

Partitional clustering algorithms can be divided into two classes – hard partitional algorithms and soft partitional algorithms. A hard partitioning is one in which observations are classified as either belonging to or not belonging to a cluster. In this case an observation has a 1 or 0 ("yes or no") degree of membership is a cluster – it is either a member of that cluster or not. This scheme has little ambiguity associated with it when the clusters in a data set are compact and well separated. However, when the clusters overlap or are not so compact, some vectors (at least) in one cluster bear some semblance to vectors in other clusters. Such situations are quite common in data analysis, serving as an indication that degrees of membership cannot always be so definitive and must therefore not be restricted to values of either zero or one.

In soft clustering the membership degrees of pattern vectors are real number values between 0 and 1. The sum of the degrees of membership of a vector to all K clusters is always equal to 1. The family of fuzzy K-means clustering algorithms comes under this class of clustering techniques. The degree of membership, u_{ij}, present in Eqs. (1) and (3), measures the likelihood of observation $\boldsymbol{X}_j$ belonging to

cluster i. It is a function of the computed distances of vector $\boldsymbol{X}_j$ from all K cluster prototypes and is defined as:

$$u_{ij} = \frac{\left[\dfrac{1}{d^2(\boldsymbol{X}_j, \boldsymbol{V}_i)}\right]^{1/(m-1)}}{\displaystyle\sum_{k=1}^{K}\left[\frac{1}{d^2(\boldsymbol{X}_j, \boldsymbol{V}_k)}\right]^{1/(m-1)}}. \tag{5}$$

The weighting exponent m is a real number greater than 1 and controls the "fuzziness" of the cluster memberships, u_{ij}'s. The closer m is to 1 the harder are the membership values, i.e. the degrees of membership assume values close to either 0 or 1. Larger values of m lead to a smoother gradation of the degrees of membership.

After membership values have been calculated, new prototypes for each of the clusters are determined using Eqs. (2) and (4) for the analysis of heterogeneous data or Eq. (4) only for the analysis of purely spherical joint data. Distances from the pattern vectors to the updated cluster centroids are again computed and updated memberships, $\hat{u}_{ij}$, determined. If the absolute value of the smallest change between the recalculated memberships and the previous memberships is less then an established tolerance, i.e. if

$$\max_{ij}[|u_{ij} - \hat{u}_{ij}|] < \varepsilon, \tag{6}$$

then the iterations are terminated, else the procedure loops back to the step of computing updated prototypes.

Based on the final membership values vectors are assigned to respective clusters or groups. An observation, $\boldsymbol{X}_j$, is assigned to the cluster i $(i = 1 \ldots K)$, when its membership degree to that particular cluster u_{Ij}, is greater than its membership values to all other clusters. In case of ties, i.e. when the maximum membership value of an observation occurs for two or more clusters, the observation is assigned to the least-numbered cluster (Gustafson and Kessel, 1978). The sequence of steps outlined above for the fuzzy K-means algorithm is illustrated with a flow chart in Fig. 2.

Because the fuzzy K-means algorithm will usually partition a data set into K partitions, whether or not the data set actually contains K clusters, there is the need to establish a criterion or some criteria for deciding when the true structure has been recovered. (The true structure of a data set refers to its correct number of clusters and correct membership of pattern vectors to their respective clusters.) The area of cluster analysis that attempts to solve these problems is known as cluster validity and the indices for discriminating between various partitions or cluster results are known as validity or performance indices. The different cluster validity indices for the fuzzy K-means algorithm for clustering joint survey data, and their application to sample data sets, are not discussed in this paper. Detailed discussions of cluster validity indices for the fuzzy K-means clustering of discontinuity data can be found in (Hammah and Curran, 1998a; Hammah and Curran, 1998b).

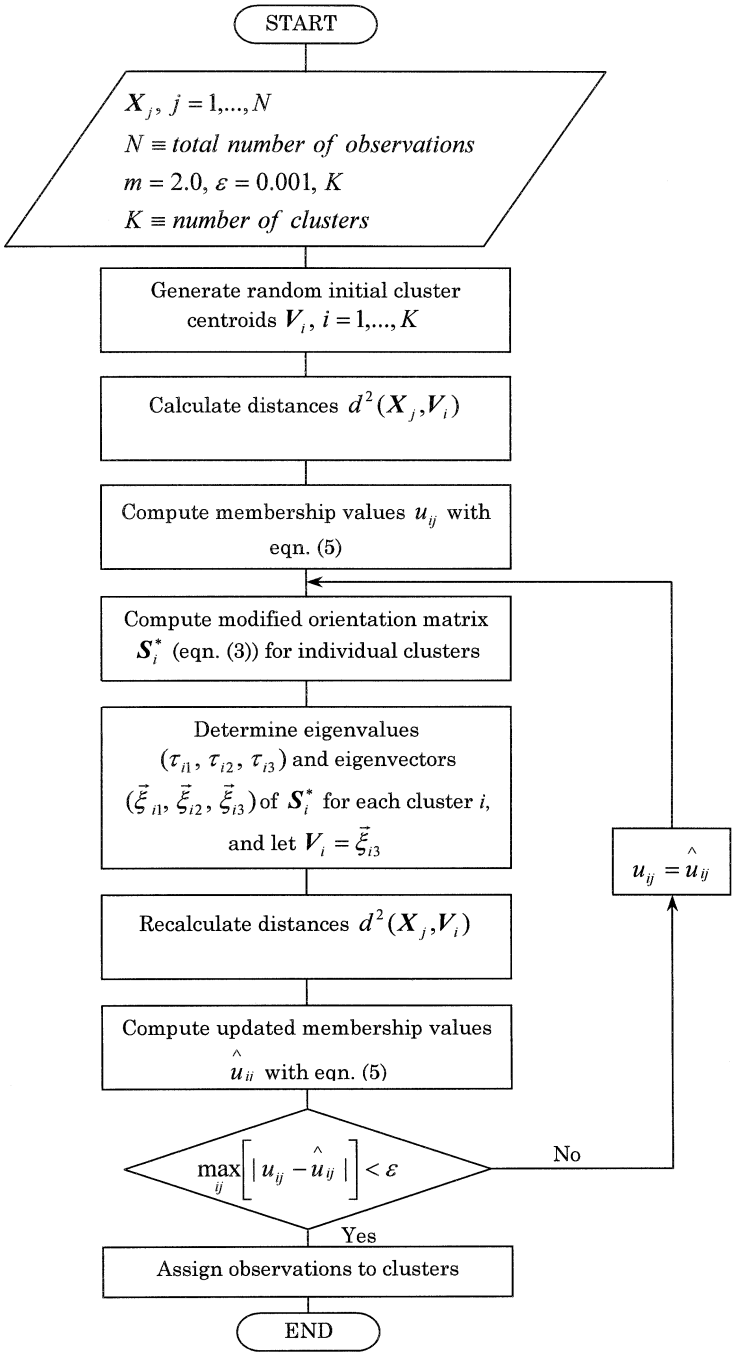

Fig. 2. Flow chart of the basic fuzzy *K*-means algorithm for clustering discontinuity orientations

3. Distance Measure

It was mentioned earlier in this paper that vectors belonging to the same cluster lie in a particular region of a pattern space. As result of this geometric property of pattern vectors, clustering algorithms must have means of establishing associations between vectors and various clusters. The association between vectors and clusters can either be in the form of similarity measures, or dissimilarity measures (distance metrics).

Similarity measures measure the relationship between two vectors or observations (both vectors are of P dimensions), based on the values of the P variables common to both objects (Everitt, 1980). Similarity measures are large when the two objects being compared share great commonality and small when they differ significantly from each other. Generally, they are real numbers between 0 (when the objects being compared have no similarities at all) and 1 (when the objects being compared share absolutely the same characteristics).

Distance measures vary inversely to similarity measures in magnitude. When two objects share the same values of the P variables common to both, the distance between them is 0. The more dissimilar the two observations are, the greater the distance is between them. Distance metrics can be greater than 1 and must satisfy some rigorous conditions (which are outlined in the next section). Because fuzzy K-means clustering routines use distance measures, the rest of this paper shall focus solely on them.

The distinctive features of joint survey data hint at the type of pattern space the analyst of joint data is confronted with. Pattern spaces can be either homogeneous or heterogeneous (Nadler and Smith, 1993). In homogeneous pattern spaces all the P variates (or dimensions) of a vector are of the same nature. For example, in a P-dimensional Euclidean space, all the axes have the same nature and therefore the space is homogeneous. Another example of a homogeneous space is the surface of a unit sphere (the space for spherical data). Spherical data or orientation information is represented by the surface of a unit sphere in $\boldsymbol{R}^3$ space. Mathematically, this space can be represented as:

$$\Omega_3 = \{\boldsymbol{X} \in \boldsymbol{R}^3 : X_1^2 + X_2^2 + X_3^2 = 1\}. \tag{7}$$

The variables X_1, X_2, X_3 are the direction cosines of vector $\boldsymbol{X}$.

Heterogeneous pattern spaces arise when a vector possesses variables that lie in two or more different homogeneous spaces. The space within which the analyst of joint orientation data often has to work is heterogeneous. Part of the space (the portion commonly analyzed on stereographic plots) is spherical, while the other part is mostly Euclidean. Whenever qualitative data that is multivalued is encountered in a data set, it can be treated as Euclidean. Those that are binary in nature, "yes or no", "present or absent", may, however, be also present in joint survey data and require different treatment. The treatment of such variables shall not be considered in this paper. The heterogeneity of discontinuity data is one of the primary reasons why few algorithms exist that are capable of incorporating all recorded features of joints for their delineation into joint sets or clusters. Heterogeneous pattern spaces pose problems in cluster analysis mainly because of the

difficulty of establishing an appropriate distance measure for measuring degrees of association between pattern vectors lying in such spaces.

For both homogeneous and heterogeneous pattern spaces, the weighting of variables is a key issue. Variable weighting refers to the means of increasing or reducing the impact of a variable of attribute in cluster analysis by multiplying its contribution to a distance metric with a weight. Weighting is an important element of cluster analysis because of the different measurement scales of variables, and the different degrees of information on cluster structure that variables, may provide. When a variable, for example, does not possess any information on the cluster structure of a data set, it would be ideal to assign it a weight of zero. However, since the purpose of this paper is to examine distance metrics and their general effects on cluster results, and to propose an elliptical distance norm for spherical data, it shall not provide in-depth coverage on the topic of weighting. A more comprehensive coverage of this topic as it relates to the cluster analysis of discontinuities is provided in (Hammah and Curran, 1997).

The choice of distance measures must be commensurate with the pattern space within which the clusters are believed to be embedded. Establishing which distance measure to use is one of the most fundamental steps in designing or choosing a clustering routine.

3.1 Properties of Distance Measures

For a distance function to be a valid measure or norm of distance between two vectors in a pattern space, it must satisfy the following four axioms (Nadler and Smith, 1993):

i) $d(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ (distances must be non-negative)
ii) $d(\boldsymbol{X}, \boldsymbol{Y}) = 0$ iff[1] $\boldsymbol{Y} = \boldsymbol{X}$ (reflexivity)
iii) $d(\boldsymbol{X}, \boldsymbol{Y}) = d(\boldsymbol{Y}, \boldsymbol{X})$ (symmetry)
iv) $d(\boldsymbol{X}, \boldsymbol{Y}) \leq d(\boldsymbol{X}, \boldsymbol{Z}) + d(\boldsymbol{Y}, \boldsymbol{Z})$ (metric inequality or triangle inequality),

where $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are pattern vectors and $d(\ ,\ )$ denotes a distance function.

3.2 Distance Metrics in Euclidean Space (R^3)

For pattern vectors in $\boldsymbol{R}^3$ space the most commonly used distance measure is the Euclidean norm:

$$d^2(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X} - \boldsymbol{Y})^T(\boldsymbol{X} - \boldsymbol{Y}) = \|\boldsymbol{X} - \boldsymbol{Y}\|^2 = \sum_{p=1}^{P}(X_p - Y_p)^2, \tag{8}$$

where X_p and Y_p are the values of the p-th variable for the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

[1] iff is read as "if and only if".

The Euclidean distance measure is easy to use and implement, if the measurement scale for the P variables involved is not of significance (Everitt, 1980). The Euclidean metric, however, can give very different results when the scale of a variable is changed and must therefore be used with caution. A number of researchers advocate the normalization of all variables in Euclidean space before using them in a cluster analysis. However, careful consideration should be given to the type of the normalization chosen. Some normalization schemes can worsen cluster results by weakening differences between clusters (Everitt, 1980).

To counter the problem of scaling when using the Euclidean distance measure, a weighted Euclidean norm:

$$d^2(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{p=1}^{P} w_p (X_p - Y_p)^2 \tag{9}$$

can be used instead. In clustering algorithms, there are various ways of determining the weighting coefficients, w_p, such as the extremal weighting method (Lumelsky, 1982), applied to the delineation of discontinuity data sets by Hammah and Curran (1997).

Although the Euclidean distance metric enjoys the most popularity in cluster analysis of data in $\boldsymbol{R}^P$ space, there exist several other distance measures for this pattern space. The usage of these metrics is dependent on the problem being solved. One such distance measure is the city-block norm:

$$d(\boldsymbol{X}, \boldsymbol{Y}) \sum_{p=1}^{P} |X_p - Y_p|. \tag{10}$$

The city-block distance metric can, for example, be used for applications involving the determination of distances traveled by vehicles within a city. Between any two points in a city, the shortest distance a vehicle can travel is not the Euclidean distance between the two points, but the distance traveled along city streets.

Both the Euclidean metric and the city-block distance measure are specific cases of a class of distance measures known as Minkowski metrics. They are defined by the general formula:

$$d^r(\boldsymbol{X}, \boldsymbol{Y}) \sum_{p=1}^{P} |X_p - Y_p|^r, \quad r = 1, 2, \ldots. \tag{11}$$

In the context of cluster analysis, all the above discussed distance measures assume that within clusters there is no correlation between the variables of pattern vectors. When this assumption is violated clustering algorithms using these measures can arrive at erroneous solutions. One of the key means of avoiding the problems posed by correlated variables in cluster analysis is the use of distance metrics that measure statistical relations. The Mahalanobis distance metric (Mahalanobis, 1936) falls into this category. It accounts for correlation between variables by including the covariance matrix of a group of pattern vectors into the distance metric, and is not affected by scale changes. The Mahalanobis distance measure

between two points in a data set is computed using the formula:

$$d^2(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X} - \boldsymbol{Y})^T \boldsymbol{C}^{-1} (\boldsymbol{X} - \boldsymbol{Y}), \tag{12}$$

where $\boldsymbol{C}$ is the covariance matrix of all the observations in the data set.

Its definition originates from the multivariate normal probability distribution function:

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{P/2} |\boldsymbol{C}|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{X} - \bar{\boldsymbol{X}})^T \boldsymbol{C}^{-1} (\boldsymbol{X} - \bar{\boldsymbol{X}}) \right], \tag{13}$$

Where $\bar{\boldsymbol{X}}$ is the mean of the distribution.

The covariance matrix of a group of N vectors is computed using the formula:

$$\boldsymbol{C} = \sum_{j=1}^{N} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T. \tag{14}$$

When the covariance matrix is equivalent to the identity matrix $\boldsymbol{I}$, i.e. when

$$\boldsymbol{C} = \boldsymbol{I} \tag{15}$$

the Mahalanobis distance metric becomes equivalent to the Euclidean distance norm. The weighted Euclidean norm is also a particular case of the Mahalanobis distance, when the covariance matrix is a diagonal matrix. The covariance matrix is diagonal when variables are not correlated. In that case the weights of the weighted Euclidean distance are just the inverses of the variances of the variables of the P-dimensional pattern vectors.

For the fuzzy K-means algorithm a modification of the Mahalanobis norm that makes local variations of the correlation between the variables of vectors in a cluster possible is given by the formula:

$$d^2(\boldsymbol{X}_j, \boldsymbol{V}_i) = |\boldsymbol{F}_i|^{1/P} (\boldsymbol{X}_j - \boldsymbol{V}_i) \boldsymbol{F}_i^{-1} (\boldsymbol{X}_j - \boldsymbol{V}_i)^T, \tag{16}$$

where $\boldsymbol{F}_i$ is the covariance matrix of the i-th cluster. This metric is commonly known as the scaled Mahalanobis distance (Krishnapuram and Keller, 1993). The matrix $\boldsymbol{F}_i$ is called the fuzzy covariance matrix (Gustafson and Kessel, 1978) and is defined as:

$$\boldsymbol{F}_i = \frac{\sum_{j=1}^{N} u_{ij}^m (\boldsymbol{X}_j - \boldsymbol{V}_i)(\boldsymbol{X}_j - \boldsymbol{V}_i)^T}{\sum_{j=1}^{N} u_{ij}^m}. \tag{17}$$

3.3 Distances Measured Between Points on a Unit Sphere

In the space defined as the surface of a unit sphere in $\boldsymbol{R}^3$, the use of the Euclidean norm for measuring distances would not be natural. As an example let us look at the case of three discontinuities with orientations 84/000, 86/000 and 88/180 (given in the dip/dip direction convention). Expressed in direction cosines (it is

assumed that all the poles are being plotted in the upper hemisphere), the poles of the three discontinuities have the following coordinates: $(-0.99452, 0.0, 0.104528)$, $(-0.99756, 0.0, 0.069756)$, and $(0.999391, 0.0, 0.034899)$, respectively. When plotted in the *x*-*z* plane, it can be seen that the second of these three unit vectors is the bisector of the acute angle between the other two, i.e. the second orientation is the mean of the first and third orientations. As such the distances measured from the second pole to the other two should be the same. If the square of the distance of the second vector from the first vector is measured using the Euclidean norm the answer obtained is 0.0012183. The squared distance of the third vector from the second using the same norm is 3.98903! The only way of avoiding absurd results would be to reverse the sign of the third vector and then use these new coordinates to determine its squared distance from vector 2. Thus the square of the Euclidean metric can be used for such data only by choosing distances between two vectors as the minimum of these two values:

a) the square of the magnitude of the sum of the vectors, and
b) the square of the magnitude of the difference between the two vectors.

A more natural metric for the vectors should involve the angles between the vectors. An appropriate metric for Ω_3 is the square of the sine of the angle between two vectors (the cosine is a similarity measure (Anderberg, 1973)) and is written as:

$$d^2(\boldsymbol{X}, \boldsymbol{Y}) = 1 - (\boldsymbol{X} \cdot \boldsymbol{Y})^2, \tag{18}$$

where $\boldsymbol{X} \cdot \boldsymbol{Y}$ is the dot product of the two vectors (the cosine of the angle between the vectors). This distance measure (which shall be called the sine-squared measure from this point forth) satisfies all four conditions demanded of a distance metric. When it is used in the example given above the answer obtained for both distances is 0.001218. This does not require the reversal of signs in order to arrive at the right answers. The sine-squared metric is ideal for clustering orientations since the angle between two orientations never exceeds 90°. This is because orientations are actually axial data or undirected lines in space (Fisher, Lewis and Embleton, 1987), and so their measure of closeness is the acute angle between them.

The metric supplied by Eq. (18) makes it possible for the fuzzy *K*-means algorithm for discontinuity data to correctly delineate joint sets that wrap between hemispheres. The example of the three orientations examined above is a simple instance of this ability of the algorithm. More detailed cases on the correct handling of wrapped discontinuity sets be the algorithm can be found in (Hammah and Curran, 1998a).

The sine-squared measure enjoys additional advantages over the Euclidean metric in the clustering of orientations. It forms the fundamental reason why the method of computing the centroids of orientation clusters through the eigenanalysis of the orientation matrix, unambiguously determines means to lie within the acute-angled cone defined by the vectors of a cluster (Hammah and Curran, 1997a). In addition, this metric makes it possible to define validity or performance measures appropriate for determining the optimality (correctness) of cluster partitions (Hammah and Curran, 1998a).

One major difference between the sine-squared measure of distance in Ω_3 and the metrics used in Euclidean space is that the sine-squared measure has an upper bound of 1. The constraint in the magnitude of this measure can result in significant differences in scale, and in some cases the distances measured on a unit sphere can be overshadowed in a cluster analysis of discontinuities by the Euclidean distances. For example, in the cluster analysis of a discontinuity data set involving orientations and spacing of the order of tens of metres, the contribution of orientation to overall distances computed between observations will be very small compared to that of spacing. Subsequently, the impact of the orientations of discontinuities on cluster results would be much minimized. This problem would be worsened if the spacing were to be measured in centimetres instead of metres. It is for these reasons that it sometimes becomes wiser to standardize (normalize) Euclidean variables before entering them into a cluster analysis. One other way of reducing the impact of this variable scale or range would be to apply different weights to the different variables in an analysis. The role of variable weighting in the clustering of discontinuity data is covered in (Hammah and Curran, 1997).

4. Geometric Significance of Distance Measures

Every distance measure induces a topology on the pattern space in which it is employed (Bezdek, 1981). A distance measure defines a "unit ball" of a prescribed geometry or shape in the space in which it is defined. This phenomenon significantly contributes to the observation made by Everitt (1980) that the output of a clustering will only be as meaningful as the input distance metric.

The Euclidean metric describes hyperspheres in P-dimensional pattern space. In Fig. 3a the contours of equal distance (or constant probability) from a cluster centroid described by the Euclidean metric in two-dimensional space are shown (Bow, 1992). These contours take the shape of concentric circles. For three-dimensional Euclidean data, contour profiles of distance are spherical surfaces. This indicates that distance metrics are associated with geometric shapes in space, and therefore control the shapes of clusters that an algorithm can identify.

The Euclidean distance measure, therefore, should be used in cases when the clusters embedded in the pattern space are expected or suspected to be hyperspheres. Statistically, this would be equivalent to searching for clusters of data in pattern spaces in which the variables are uncorrelated and have equal variances.

As noted earlier on, the weighted Euclidean distance norm is equivalent to the Mahalanobis distance with a diagonal covariance matrix. This measure describes hyper-ellipsoids in P-dimensional space. The principal axes of these hyper-ellipsoids are parallel to the axes of the space in which variables are measured. Fig. 3b is an example of two-dimensional ellipsoids (ellipses) defined by this distance measure in two-dimensional space (Bow, 1992). The major and minor principal axes of the ellipses shown on the figure are parallel to the variable axes x_1 and x_2. The weighted Euclidean metric should be used for clustering data for which the variables involved are uncorrelated, but possess different variances.

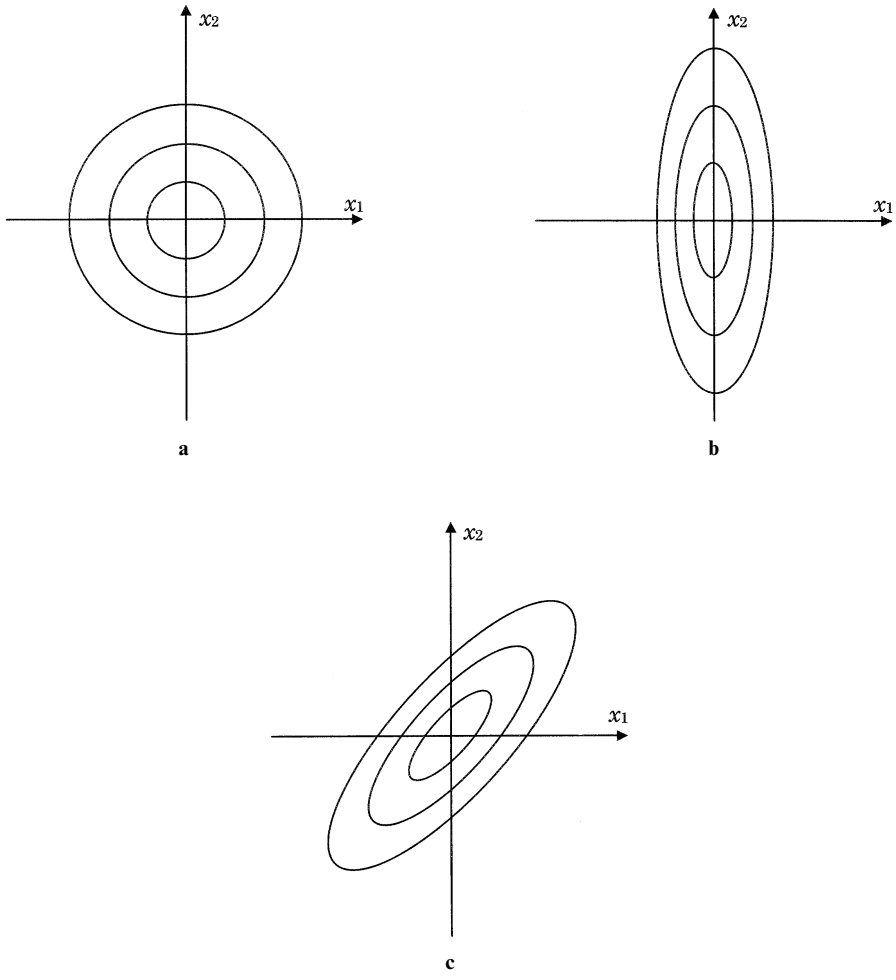

Fig. 3a–c. Contours of constant probability for three different cluster shapes in $\boldsymbol{R}^2$ pattern space. **a** Cluster shape when covariance matrix, $\boldsymbol{C}$, is equal to the identity matrix, $\boldsymbol{I}$. **b** Cluster shape when covariance matrix, $\boldsymbol{C}$, is diagonal, and $\text{var}(x_2) > \text{var}(x_1)$. **c** Cluster shape when covariance matrix, $\boldsymbol{C}$, is fully populated

The unit geometric shapes defined by the Mahalanobis distance with a fully populated covariance matrix are also hyperellipsoids, but the principal axes of these are not aligned with the axes of the space. It is a more general distance metric for Euclidean space. Shown is Fig. 3c are the contours described by the Mahalanobis distance with a fully populated covariance matrix for two-dimensional Euclidean space (Bow, 1992).

From the above discussions, it can be seen that distance metrics control the shape of the clusters that can be identified by a cluster algorithm. An algorithm founded on the Euclidean distance norm, for example, could experience difficulties in identifying clusters in a data set that were elliptically shaped, with principal axes

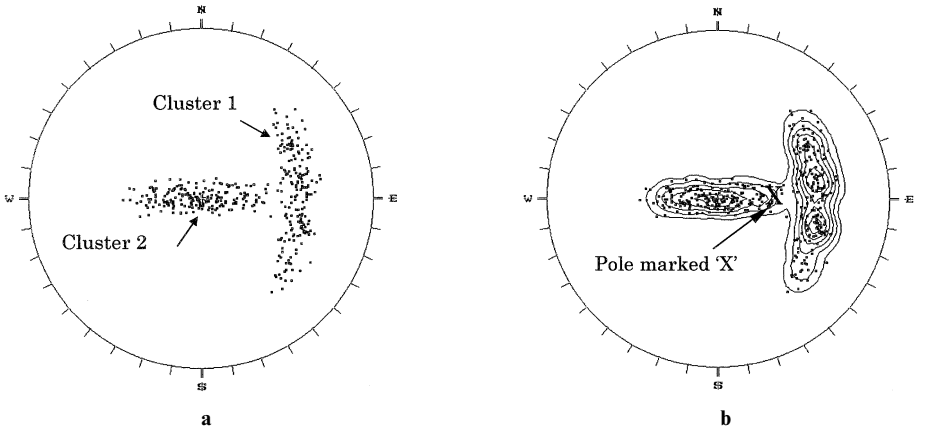

Fig. 4a,b. Stereographic plots of 400 poles in a simulated data set of two approximately elliptical clusters of joints at right angles to each other. **a** Pole plot of the joints in the data set. **b** Contour plot of the data

not parallel to the axes of the measurement space of the variables, i.e. clusters of the type shown in Fig. 3c.

The fuzzy clustering algorithm proposed by Gustafson and Kessel (1978) employs the scaled Mahalanobis distance to account for differences in the shapes of the clusters in an $\boldsymbol{R}^{\mathrm{P}}$ pattern space. The fuzzy covariance matrix allows for these local differences to be accounted for during clustering. Bezdek (1981) observed after running this algorithm on a sample data set that fuzzy clustering with this distance norm yields results consistent with theoretical predictions.

Topology of Sine-Squared Distance Measure on the Unit Sphere

The distance measure (Eq. 18) for spherical data proposed by the author (Hammah and Curran, 1998a) describes spherical contours of constant probability on a unit sphere (the next section explains why this is so). The sine-squared measure works well for circular distributions on a sphere (e.g. Fisher distributed joints). It performs well even when the clusters to be recovered have elliptical shapes, for which corresponding principal axes are approximately parallel. (In such cases, the major principal axes of the clusters are sub-parallel, i.e. are close to being parallel. This automatically implies that the minor principal axes are likewise subparallel.) However, in cases where elliptical distributions of joints are unfavorably oriented with respect to each other, it would prove to be inadequate. Fig. 4 is a stereographic plot[2] of two elliptically shaped (approximately) distributions, each consisting of 200 poles, at right angles to each other (the data are simulated). Were the

[2] All stereographic plots shown in the paper were produced using DIPS, a software package for the interactive analysis of orientation data, developed by the Rock Engineering Group of the University of Toronto (Diederichs and Hoek, 1996).

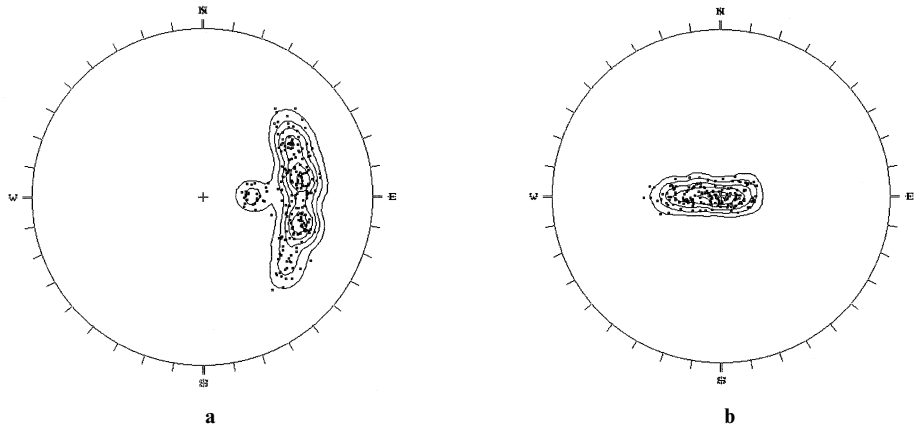

Fig. 5a,b. Results of fuzzy *K*-means clustering using the sine-squared distance metric. **a** Contour plot of poles assigned to cluster 1. **b** Contour plot of poles assigned to cluster 2

two distributions to be parallel to each other, the algorithm would place the decision boundary midway between the two clusters and thus would assign poles correctly. However, under the current configuration of the clusters, the sine squared distance measure would encounter some problems in establishing the correct decision boundary. Looking at the pole marked with an 'X' in Fig. 4(b), it can be seen on the contour profile that it has a greater likelihood of belonging to cluster 2 than to cluster 1. However, if the square of the sine of the angle between that pole and the cluster centroids is used as the distance measure, the point would be nearer to the prototype of cluster 1 than to that of cluster 2. Therefore that pole would be assigned to cluster 1 instead of the cluster 2.

The results of the fuzzy *K*-means algorithm run on the data set with the sine-squared measure confirm the theoretical predictions. The cluster partitions identified by the algorithm can be seen in Figs 5a and 5b. 29 poles, which belong to cluster 2, are incorrectly assigned to cluster 1, because those points were deemed to be closer to the centroid of cluster 1 than to that of cluster 2. This happened because the distance metric, in its 'ignorance', failed to recognize that distances were not only dependent on the angle between vectors, but also depended on the shapes of the clusters. These incorrectly assigned points show up on the contour plot of cluster 1 as the circular bulge, on the left side of the cluster (Fig. 5a).

5. Elliptical Distribution for Spherical Data

The most widely encountered statistical model for spherical data is the Fisher distribution:

$$f(\boldsymbol{X}) = c(\kappa)\exp(\boldsymbol{X}\cdot\vec{\lambda}), \tag{19}$$

where $c(\kappa)$ is the normalization constant and is equal to

$$c(\kappa) = \frac{\kappa}{2\pi(\exp(\kappa) - \exp(-\kappa))} = \frac{\kappa}{4\pi \sinh(\kappa)}. \tag{20}$$

The Fisher distribution is a two parameter model[3] (the parameters are κ and λ) and is the spherical analogue of the bivariate normal distribution with equal variances. The contours of constant probability described by the Fisher distribution are circular in shape. Note that the quantity $\boldsymbol{X} \cdot \vec{\lambda}$ is the cosine of the angle between a vector $\boldsymbol{X}$ and the mean of the distribution, $\vec{\lambda}$, and is the same as the dot product term found in the sine-squared distance measure.

The spherical analogue of the general bivariate normal distribution is the Kent distribution (Kent, 1982):

$$f(\boldsymbol{X}) = c(\kappa, \beta) \exp\{\kappa(\boldsymbol{X} \cdot \vec{\xi}_3) + \beta[(\boldsymbol{X} \cdot \vec{\xi}_2)^2 - (\boldsymbol{X} \cdot \vec{\xi}_1)^2]\}. \tag{21}$$

The contours described by the Kent distribution are near ellipses close to the mean of the distribution, but are generally oval in shape. It is a five parameter model with the parameters being two shape parameters κ and β, and the triple of vectors $\vec{\xi}_1$, $\vec{\xi}_2$, and $\vec{\xi}_3$. $\vec{\xi}_3$ is the mean of the probability distribution. In the plane perpendicular to $\vec{\xi}_3$, the contour profiles of the distribution are oval in shape. $\vec{\xi}_2$ and $\vec{\xi}_1$ are the major and minor principal axes of the distribution profile, respectively. The density of the distribution profile is highest along the major principal axis $\pm\vec{\xi}_2$ and least along the minor axis $\pm\vec{\xi}_1$ (Fisher, Lewis and Embleton, 1987). The Kent distribution has rotational asymmetry about the mean and is consequently more flexible than the Fisher distribution (Fisher, Lewis and Embleton, 1987). For a specified cluster i of elliptical shape, the geometric representation of its three eigenvectors relative to the position of the cluster on a unit sphere is shown on Fig. 6. Contours of equal probability of the Kent distribution representing the cluster, rather than the individual points in the cluster, are drawn on Fig. 6 to help arrive at a better appreciation of the shape of the cluster, and the geometric elements of the distribution.

The shape parameter κ (also known as the concentration parameter) controls the concentraion of the poles about the mean vector. The larger the value of κ the less is the scatter of the distribution towards the mean, $\vec{\xi}_3$. β is a measure of the ratio of the density along the major principal axis to the density alony the minor principal axis. The greater the value of β the greater is the departure of the distribution profile from circular symmetry. β is thus called the 'ovalness parameter'. The Fisher distribution is a limiting case of the Kent distribution when β is zero.

The Kent distribution has two main forms – a unimodal form and a bimodal form (Kent, 1982). These forms are determined by the ration of κ to β. The distribution is unimodal when $\kappa/\beta \geq 2$, and bimodal when $\kappa/\beta < 2$. It is the unimodal form, which holds great interest for us (primarily because it is for this form that the Kent distribution correctly behaves as the spherical analogue of the bivariate normal distribution (Kent, 1982)). The normalizing constant c(κ, β) is

[3] The original formulation of the Fisher distribution has three parameters κ, α and β. The colatitude α and longitude β of the mean orientation of the distribution, together, are equivalent to the vector of direction cosines, λ, in the two parameter formulation of the distribution (Fisher et al. 1987).

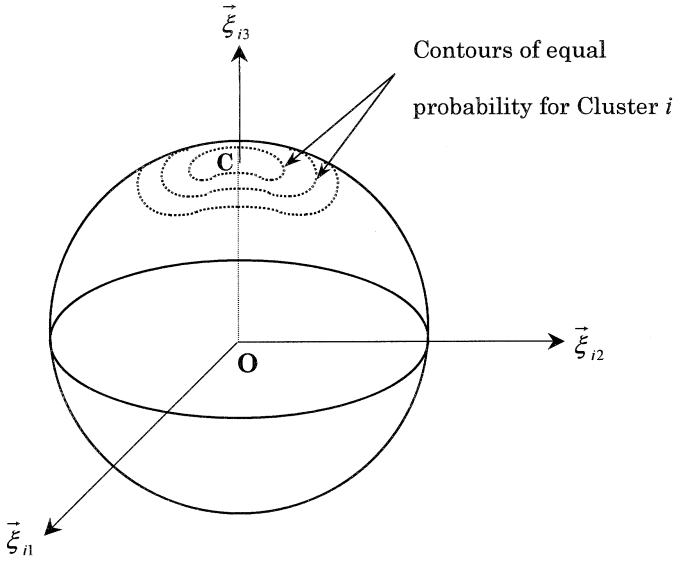

Fig. 6. Position of the eigenvectors of a cluster orientation matrix, relative to the distribution of the points (discontinuity poles) in the cluster

calculated for the unimodal form of the Kent distribution from the formula:

$$c(\kappa,\beta) = 1 \Big/ (2\pi)^{3/2}\kappa^{-1/2}\sum_{r=0}^{\infty}\frac{(2r)!}{r!r!}\left(\frac{\beta}{\kappa}\right)^{2r} I_{2r+1/2}(\kappa), \tag{22}$$

where $I_{2r+1/2}(\kappa)$ are modified Bessel functions.

The maximum value of the Kent distribution occurs at the mean direction of the distribution. A vector located at the mean would be orthogonal to the principal axes $\vec{\xi}_2$ and $\vec{\xi}_1$ thus eliminating the term $\beta[(\boldsymbol{X}\cdot\vec{\xi}_2)^2 - (\boldsymbol{X}\cdot\vec{\xi}_1)^2]$ at the mean. A method for estimating the parameters of the unimodal Kent didstribution, outlined in (Kent, 1982), is provided in the appendix of this paper.

6. Spherical Analogue of Mahalanobis Distance

Just as the exponential power term of the multivariate normal distribution forms the basis of the Mahalanobis distance measure, the corresponding term of the Kent distribution

$$\kappa(\boldsymbol{X}\cdot\vec{\xi}_3) + \beta[\boldsymbol{X}\cdot\vec{\xi}_2)^2 - (\boldsymbol{X}\cdot\vec{\xi}_1)^2]$$

can be used to define a new distance metric for spherical data. The proposed distance measure for the fuzzy *K*-means clustering of spherical data has the definition:

$$d^2(\boldsymbol{X}_j,\vec{\xi}_{i3}) = 1 - \frac{\{\kappa_i(\boldsymbol{X}\cdot\vec{\xi}_{i3}) + \beta_i[(\boldsymbol{X}\cdot\vec{\xi}_{i2})^2 - (\boldsymbol{X}\cdot\vec{\xi}_{i1})^2]\}^2}{\kappa_i^2}. \tag{23}$$

The reason for normalizing the second term in Eq. (23) by dividing it with κ_i^2 is to ensure that distances are always positive (from Eq. (21) it can be noted that the maximum value of the exponential term in the unimodal Kent distribution is κ). When $\beta_i = 0$ the new distance metric reduces to the sine-squared distance measure.

This new metric has approximate ellipses as its basic shape and is thus more general than the sine-squared distance norm. It uses estimated values of the concentration parameter κ_i and 'ovalness' parameter β_i of the local Kent distributions, describing the clusters present in a spherical (directional) data set, to calculate distances.

The procedure for estimating the local parameters of the clusters follows the conceptual model originally proposed by Kent (see Appendix). To enable the incorporation of the elliptical distance measure into fuzzy *K*-means algorighm, the model parameters are estimated from the eigenanalysis of the modified orientation matrices for the clusters in a data set. It was mentioned earlier that the eigenanalysis of a cluster orientation matrix $\boldsymbol{S}_i^*$ produces three orthogonal vectors. The position of the eigenvectors relative to the cluster was illustrated in Fig. 6.

Watson (1966) furnished the geometric interpretation of the eigenvectors of the orientation matirx of a unitary cluster (Eq. (A1) in the Appendix). For a data set consisting of a singled cluster, all membership values, u_{ij}, in the fuzzy clustering framework are equal to 1.) The eigenvector $\vec{\xi}_3$, corresponding to the largest eigenvalue of the orientation matrix τ_3 of such a cluster of points, is the mean of the cluster. The other two eigenvectors, $\vec{\xi}_2$ and $\vec{\xi}_1$, indicate, respectively, the major and minor principal axes of the elliptical contours of the cluster. Thus the three vectors obtained from the eigenanalysis of an orientation matrix are equivalent to the triple of vectors present in the Kent distribution. Following this result, the eigenvectors of the modified orientation matrix, $\boldsymbol{S}_i^*$, of a cluster i, can be interpreted as the approximations of the axes of the Kent distribution describing the cluster (Hammah and Curran, 1998a).

Besides developing the Kent distribution, Kent (1982) also provided a method for estimating the shape parameters of the distribution from a data sample, drawn from the distribution. This estimation procedure is provided in the Appendix. In this process, an intermediate value, Q (Eq. (A11) of the Appendix), is calculated and used in evaluating the parameters κ and β of the Kent distribution. This value Q is very closely approximated by the quantity, Q', computed from the formula

$$Q' = \tau_2 - \tau_1. \tag{24}$$

Therefore, in estimating the local parameters κ_i and β_i of Kent distributions for the clusters in a data set, the Q_i's computed from the above formula using local eigenvalues τ_{i1} and τ_{i2}, can be used in place of the values of Q. The steps needed to calculate the parameters of local Kent distributions in the cluster analysis of discontinuity orientations are outlined next.

For a cluster i of orientations, estimation of shape parameters is performed as follows:

i) Find the mean resultant length, R_i, of the cluster using the formula

$$\bar{R}_i = \left(\sqrt{R_{ix}^2 + R_{iy}^2 + R_{iz}^2}\right) \Big/ \sum_{j=1}^{N} (u_{ij})^m, \tag{25}$$

where

$$R_{ix} = \sum_{j=1}^{N} (u_{ij})^m x_j, \quad R_{iy} = \sum_{j=1}^{N} (u_{ij})^m y_j, \quad \text{and} \quad R_{iz} = \sum_{j=1}^{N} (u_{ij})^m z_j. \tag{26}$$

ii) Determine the value Q_i' for the cluster from the Eq. (24).

iii) The shape parameters κ_i and β_i of the cluster are subsequently determined from the formulae:

$$\hat{\kappa}_i = (2 - 2\bar{R}_i - Q_i')^{-1} + (2 - 2\bar{R}_i + Q_i')^{-1}, \quad \text{and} \tag{27}$$

$$\hat{\beta}_i = \tfrac{1}{2}\{(2 - 2\bar{R}_i - Q_i')^{-1} - (2 - 2\bar{R}_i + Q_i')^{-1}\}. \tag{28}$$

The elliptical distance metric (Eq. (23)), founded on the Kent distribution, can be incorporated into the fuzzy *K*-means algorithm of Hammah and Curran (1998a) in a straightforward manner. From the flow chart of the algorithm in Fig. 2, it can be seen that the only time distances are computed before the calculation of cluster centroids is at the beginning of the algorithm. Thereafter, they are always calculated after new centroids for clusters have been computed. Because the eigenanalysis used in determining new cluster prototypes yields the principal directions of the local Kent distributions for clusters, it is not necessary to repeat it during the subsequent computation of distances.

After starting the fuzzy *K*-means algorithm with the generation of random cluster centroids, distances are first calculated using the sine-squared measure, since at this stage no eigenvalues and eigenvectors exist for the computation of parameters for the elliptical metric (Fig. 2). All other calculations for distance use the results of the eigenanalysis of the orientation matrix, and the above-outlined steps for estimating the parameters of the Kent distribution of a cluster in the elliptical metric.

It is important to note at this point that the algorithm does not require any prior information on the parameters of the Kent distributions for the clusters in a data set. As can be deduced from the discussions above on the inclusion of the elliptical metric into the fuzzy *K*-means algorithm, this new variation of the method estimates all the necessary distance parameters entirely by itself in the course of iterations.

The fuzzy *K*-means cluster algorithm for discontinuity data, in addition to the orientations of discontinuities, can include discontinuity properties such as spacing in the delineation of joint sets. This is accomplished by computing distances between observations and cluster prototypes as sums of two components – a spherical component (contribution of orientation to distance) calculated from either Eq. (18) or (23), and a Euclidean component (contribution of additional variables) determined from equations such as (8). An example of the analysis of a data set

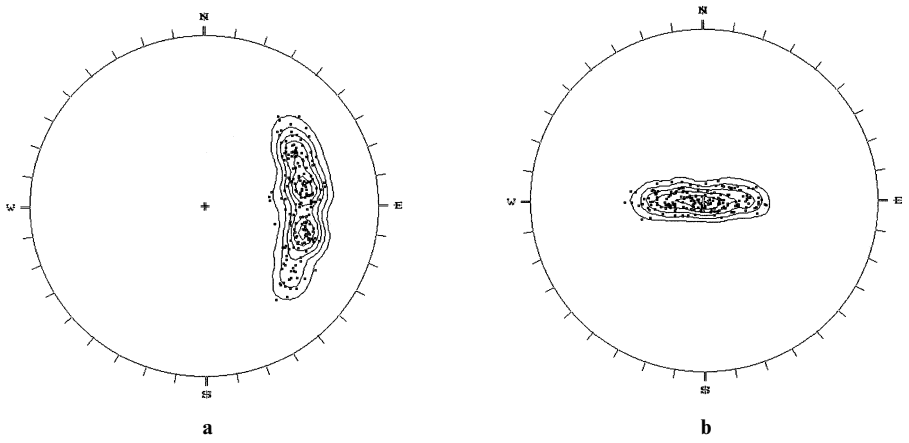

Fig. 7a,b. Results of fuzzy K-means clustering with the elliptical distance measure for spherical data. **a** Contour plot of poles assigned to cluster 1. **b** Contour plot of poles assigned to cluster 2

involving orientations and an additional discontinuity characteristic can be found in (Hammah and Curran, 1998a). In the current paper, however, because the primary focus is on gaining an understanding of the general importance of distance metrics to cluster results, the specifics as related to spherical data, and the derivation of an elliptical metric for spherical data, such examples will not be discussed.

7. Examples

Example 1: The sample data set analyzed previously with the sine-squared distance measure was reanalyzed in this example, but this time using the elliptical distance metric (Eq. (36)) instead. The algorithm was initialized with two unit vectors, chosen randomly from space of a unit sphere. No information, whatsoever, on the principal axes of the ellipses of the clusters was provided to the algorithm.

With the new distance metric, only six points belonging to cluster 2 were misclassified by the fuzzy *K*-means algorithm as coming from cluster 1. Figure 7 shows stereographic plots of these results. The circular bulge to the left of the resulting cluster 1 that was visible (see Fig. 5a), when the sine-squared distance measure was used, no longer exists on the plot of the new cluster 1 (Fig. 7a). For this particular example, the six errors in classification are not principally the shortcoming of the algorithm, but can be attributed to the fact that in the generation of the two clusters, a slight overlap of the clusters resulted. The wrong assignment of the points in the overlap region would have been avoided only if additional information had been supplied as to how the clusters differed in that region.

The values of the local parameters determined at the end of the run corresponded to the expected answers. The principal directions for both clusters were correctly computed.

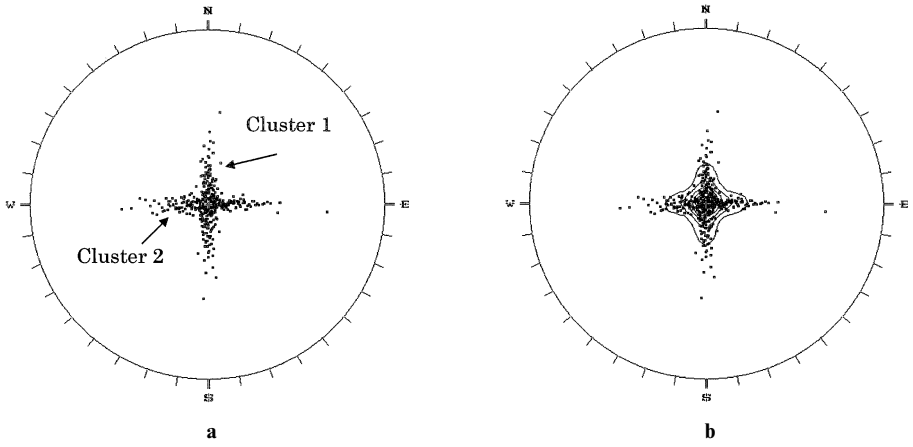

Fig. 8a,b. Plot on stereogram of 400 joints in a simulated data set (data set 2) of two approximately elliptical clusters with same centroid, but perpendicular principal axes. **a** Pole plot of the data. **b** Contour plot of the data

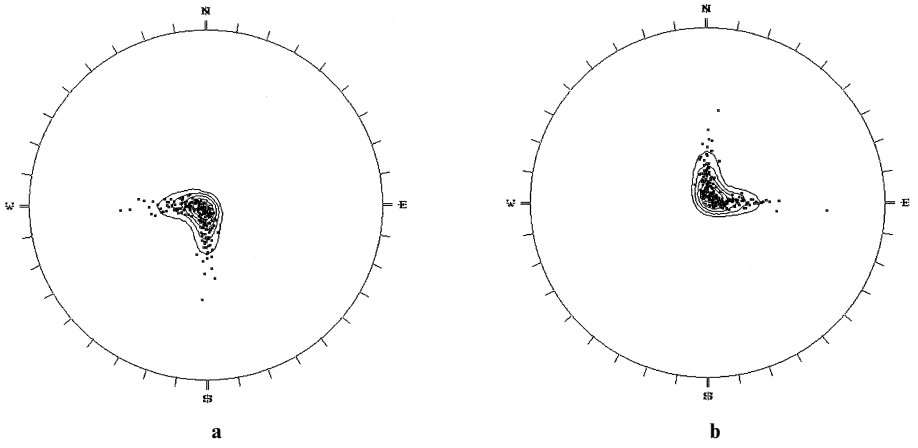

Fig. 9a,b. Structure of data set 2 recovered by fuzzy *K*-means algorithm with sine square metric as distance norm. **a** Contour plot of poles in cluster 1. **b** Contour plot of poles in cluster 2

Example 2: The effect of the incorporating the elliptical distance measure (36) into the fuzzy *K*-means algorithm was studied on a second simulated data set of 400 poles the stereographic plot of which is shown in Fig. 8. The centers of the two elliptical clusters, at right angles to each other, are exactly the same. This configuration of the two clusters implies that for the algorithm to correctly identify the clusters it must use shape information to discriminate between the two.

First the algorithm was run on the data set using the sine-squared distance measure. Predictably, it was unable to correctly recover the structure of the data set converging instead on the solution shown in Fig. 9 (note that the algorithm,

based on the initial seed points, could have arrived at a partition symmetric to the one shown in the figure). The resulting clustering is incorrect, because the distance metric was fundamentally wrong due to its inability to account for cluster shape, or estimate distances in a probabilistic sense. (It must be noted from the discussions on the Mahalanobis distance metric and the elliptical metric for spherical data that the shape parameters in distance metrics allow them to compute distances as probabilistic measures.)

The second run of the algorithm had the elliptical distance in lieu of the sine-squared distance measure. This time the basic form of the resulting assignments of the poles to the clusters is correct (Fig. 10). Note that in the region where the two clusters overlap, the assignment of the poles does not conform to the original clusters. This occurred only because a simple rule, that a point belonged to a cluster if its membership value exceeded 50%, was used in assigning the observations to the clusters. With this simple rule, unambiguous assignment of the points in the overlap region would have been possible only if additional information (an additional variable) had been provided that helped distinguish the points in that region. It is possible to devise more sophisticated assignment rules that would allow points in a region of overlap to belong to more than one cluster.

All this, however, does not detract from the fact that shape information has enabled the algorithm to perform very well. The inclusion of the elliptical distance measure enabled the algorithm to use shapes to distinguish between the two clusters. The direction cosines for the major and minor principal directions for the clusters at the end of the run were:

$$\vec{\xi}_{21} = (-0.0123782, -0.999918, 0.0006185)$$

$$\vec{\xi}_{11} = (-0.999918, -0.0123782, 0.0008104) \quad \text{with } \kappa = 72.48 \quad \text{and}$$

$$\beta = 30.85 \quad \text{for cluster 1,}$$

and

$$\vec{\xi}_{22} = (-0.999792, -0.0202372, 0.0008707)$$

$$\vec{\xi}_{12} = (-0.0203754, -0.999792, 0.0010402) \quad \text{with } \kappa = 74.44 \quad \text{and}$$

$$\beta = 31.75 \quad \text{for cluster 2.}$$

The corresponding principal directions are perpendicular to each other, as should be the cases.

The recovered structure of the data set is shown on the stereographic plots is Fig. 10. These results indicate that the new elliptical distance metric for spherical data being proposed can enhance the clustering of data of this kind.

8. Conclusion

The fuzzy *K*-means algorithm adapted to be able to incorporate both the spherical data component and Euclidean component of joint information offers a considerable advantage for the automatic identification of joint sets in survey data. How-

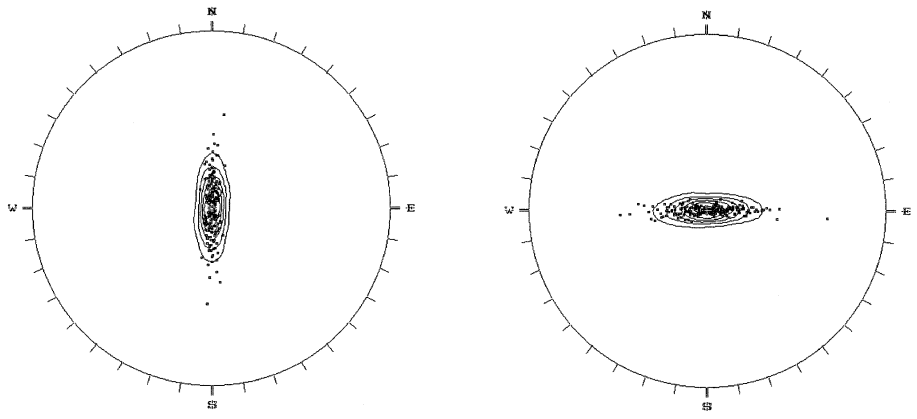

Fig. 10a,b. Structure of data set 2 recovered by fuzzy K-means algorithm using elliptical distance measure **a** Contour plot of poles assigned to cluster 1. **b** Contour plot of poles assigned to cluster 2

ever, the ability of the algorithm to recover the true structure of data sets is highly dependent on the distance metric at the basis of the algorithm. Distance measures are needed by clustering algorighms to establish the extent of the associations between observations in a data set and the centroids of the clusters believed to exist in the data. The fuzzy *K*-means algorithm computes membership degrees of pattern vectors to clusters based on the distances of the pattern vectors from the cluster prototypes.

Distance norms are to be chosen for an analysis first and foremost as a result of careful consideration of the space in which the sought for clusters are believed to be embedded. Homogeneous spaces are pattern spaces in which all variables have the same nature. In heterogeneous pattern spaces the variables describing observations belong to two or more homogeneous sub-spaces. Distances in pattern spaces (sub-spaces) fundamentally depend on the characteristics of the variables of the space. Norms, which may be useful in one space (sub-spaces), may be completely inappropriate in another.

Distance metrics play a key role in defining a clustering technique's performance. They impose subtle geometric constraints on the structure recovered by the algorithm. Each distance measure defines a "unit ball" (Bezdek and Pal, 1992) of a particular shape in the pattern space and this controls the shape of clusters an algorithm can extract from a data set. The shapes defined by distance norms depend on whether or not they take into consideration correlation among variables. They also depend on the variances of the different variables.

The sine-squared distance measure originally proposed for the fuzzy *K*-means algorithm for joint data measures only isotropic distances on the unit sphere (circular probability contours). When the coordinates or the points belonging to a cluster are correlated so that the contours of constant probability of the cluster are non-circular, it experiences problems delineating clusters correctly. As a result the authors have proposed a new distance measure based on the Kent probability

distribution. This new measure is able to account for differences in the local shapes (directions of principal axes and shape parameters) of clusters. Effectively, it accounts for cluster shape differences by adjusting the distance norm for individual clusters so that measured distances from cluster centroids reflect the local shapes of the clusters in a data set. The analysis of a sample data set, in which two elliptical clusters differed only in the directions of their principal axes, confirmed the enhancements the new measure brings to the automatic identification of joint sets using the fuzzy *K*-means algorithm.

Appendix

Parameter Estimation for the Unimodal Kent Distribution

Kent (1982) outlines a moment estimation method for estimating the parameters of the Kent distribution when the ratio $\kappa/\beta > 2$. The algorithm for estimating these parameters given the direction cosines $(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)$ of N vectors from a Kent distribution is as follows:

i) compute the orientation matrix $\boldsymbol{S}$:

$$\boldsymbol{S} = \frac{1}{N}\begin{bmatrix} \sum_{j=1}^{N} x_j x_j & \sum_{j=1}^{N} x_j y_j & \sum_{j=1}^{N} x_j z_j \\ \sum_{j=1}^{N} x_j y_j & \sum_{j=1}^{N} y_j y_j & \sum_{j=1}^{N} y_j z_j \\ \sum_{j=1}^{N} x_j z_j & \sum_{j=1}^{N} y_j z_j & \sum_{j=1}^{N} z_j z_j \end{bmatrix} \tag{A1}$$

and the mean of the data $(\hat{x}, \hat{y}, \hat{z})$ from the formula:

$$(\hat{x}, \hat{y}, \hat{z}) = (R_x/R, R_y/R, R_z/R), \tag{A2}$$

where

$$R_x = \sum_{i=1}^{N} x_i, \quad R_y = \sum_{i=1}^{N} y_i, \quad R_z = \sum_{i=1}^{N} z_i. \tag{A3}$$

Also calculate the mean resultant length of the N vectors using the formula:

$$\bar{R} = (\sqrt{R_x^2 + R_y^2 + R_z^2})/N \tag{A4}$$

ii) compute the orthogonal matrix $\boldsymbol{H}$

$$\boldsymbol{H} = \begin{bmatrix} \sin\hat{\theta}_0 \cos\hat{\phi}_0 & -\sin\hat{\phi}_0 & \cos\hat{\theta}_0 \cos\hat{\phi}_0 \\ \sin\hat{\theta}_0 \sin\hat{\phi}_0 & \cos\hat{\phi}_0 & \cos\hat{\theta}_0 \sin\hat{\phi}_0 \\ -\cos\hat{\theta}_0 & 0 & \sin\hat{\theta}_0 \end{bmatrix}, \tag{A5}$$

where $\hat{\theta}_0$ and $\hat{\phi}_0$ are the plunge and trend of the mean $(\hat{x}, \hat{y}, \hat{z})$.

Calculate a new matrix $\boldsymbol{B}$, which is defined as the matrix product

$$\boldsymbol{B} = \boldsymbol{H}^T \boldsymbol{S} \boldsymbol{H} \tag{A6}$$

with elements

$$\begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

and define an angle $\hat{\psi}$ such that

$$\hat{\psi} = \frac{1}{2} arc \tan \left[\frac{2b_{12}}{b_{11} - b_{22}} \right]. \tag{A7}$$

iii) Compute the rotation matrix

$$\boldsymbol{K} = \begin{bmatrix} \cos\hat{\psi} & -\sin\hat{\psi} & 0 \\ \sin\hat{\psi} & \cos\hat{\psi} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{A8}$$

and then the matrix

$$\hat{\Gamma} = \boldsymbol{H}\boldsymbol{K} = (\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3). \tag{A9}$$

Finally, compute the matrix

$$\boldsymbol{V} = \hat{\Gamma}^T \boldsymbol{S} \hat{\Gamma}, \tag{A10}$$

with elements

$$\begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix}$$

and compute the value Q defined as

$$Q = |v_{11} - v_{22}|. \tag{A11}$$

iv) For large values of κ, the parameters κ and β can be determined approximately as

$$\hat{\kappa} = (2 - 2\overline{R} - Q)^{-1} + (2 - 2\overline{R} + Q)^{-1} \tag{A12}$$

$$\hat{\beta} = \tfrac{1}{2}\{(2 - 2\overline{R} - Q)^{-1} - (2 - 2\overline{R} + Q)^{-1}\}. \tag{A13}$$

References

Anderberg, M. R. (1973): Cluster analysis for applications, Academic Press, New York.

Bow, S.-T. (1992): Pattern recognition and image processing, Marcel Dekker, New York.

Bezdek, J. C. (1981): Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York.

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

Dershowitz, W., Busse, R., Geier, J., Uchida, M. (1996): A stochastic approach for fracture set definition. In: Aubertin, M., Hassani, F., Mitri, H. (eds.), Proc., 2nd NARMS, Rock Mechanics Tools and Techniques, Montreal, Balkema, Rotterdam, 1809–1813.

Diederichs, M., Hoek, E. (1997): Dips User's Guide (Version 4.0). Department of Civil Engineering, University of Toronto.

Everitt, B. (1980): Cluster analysis, Halstad Press, New York.

Fisher, N. I., Lewis, T., Embleton, B. J. (1987): Statistical analysis of spherical data, Cambridge University Press, New York.

Gath, I., Geva, A. B. (1989): Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Machine Intell. 14 (7), 773–781.

Gustafson, D. E., Kessel, W. C. (1978): Fuzzy clustering with a fuzzy covariance matrix. In: Proc., IEEE Conference on Decision and Control, San Diego, 761–766.

Hammah, R. E., Curran, J. H. (1997): Standardization and weighting of variables for the fuzzy K-means clustering of discontinuity data. Submitted for publication.

Hammah, R. E., Curran, J. H. (1998a): Fuzzy cluster algorithm for the automatic delineation of joint sets. Int. J. Rock Mech. Min. Sci. 35 (7), 889–905.

Hammah, R. E., Curran, J. H. (1998b): Validity measures for the fuzzy cluster analysis of orientations. Submitted for publication.

Harrison, J. P. (1992): Fuzzy objective functions applied to the analysis of discontinuity orientation data. In: Hudson, J. A. (ed.), ISRM Symp. Eurock '92, Rock Characterization, British Geotechnical Society, London, 25–30.

Jain, A. K. (1986): Cluster analysis. In: Handbook of pattern recognition and image processing, Academic Press, Orlando, 33–57.

Kent, J. T. (1982): The Fisher-Bingham distribution on the sphere. J. R. Statist. Soc. B, 44 (1), 71–80.

Krishnapuram, R., Keller, J. M. (1993): A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. 1 (2), 98–110.

Lumelsky, V. (1982): A combined algorithm for weighting the variables in the clustering problem. Pattern Recogn. 15 (6), 53–60.

Mahalanobis, P. C. (1936): On the generalized distance in statistics. Proc., Natn. Inst. Sci. Calcutta, 12, 49–55.

Mahtab, M. A., Yegulalp, T. M. (1982) A rejection criterion for definition of clusters in orientation data. In: Goodman, R.E., Heuze, F. E. (eds.), Proc., 22nd Symp. Rock Mech., Issues in Rock Mechanics, Berkeley, American Institute of Mining Metallurgy and Petroleum Engineers, 116–123.

Nadler, M., Smith, E. P. (1993): Pattern recognition engineering, John Wiley, Toronto.

Ruspini, E. H. (1969): A new approach to clustering. Inform. Control 15, 22–32.

Shanley, R. J., Mahtab, M. A. (1976): Delineation and analysis of clusters in orientation data. J. Math. Geol. 8 (3), 9–23.

Watson, G. S. (1966): Statistics of orientation data. J. Geol. 74 (5), 786–797.

Zadeh, L. A. (1965): Fuzzy sets. Inform. Control 8, 338–353.

Authors' address: Prof. John Curran, Rock Engineering Group, Department of Civil Engineering, University of Toronto, Toronto, ON M5S 1A4, Canada.