doi:10.1007/s11589-012-0845-3

Application of fuzzy clustering method to determining sub-fault planes of earthquake from aftershocks sequence*

Fuchang Wang[1,♦] Yongge Wan[1] Huirong Cao[2]
Zhitong Jin[1] and Qingqing Ren[1]

[1] *The Institute of Disater Prevention of China Earthquake Administration, Sanhe* 065201, *China*
[2] *College of Mathematics and Information Science, Langfang Teachers' College, Langfang* 065000, *China*

Abstract Earthquake rupture process generally involves several faults activities instead of a single fault. A new method using both fuzzy clustering and principal component analysis makes it possible to reconstruct three dimensional structure of involved faults in earthquake if the aftershocks around the active fault planes distribute uniformly. When seismic events are given, the optimal faults structures can be determined by our new method. Each of sub-fault planes is fully characterized by its central location, length, width, strike and dip. The resolution determines the number of fault segments needed to describe the earthquake catalog. The higher the resolution, the finer the structure of the reconstructed fault segments. The new method successfully reconstructs the fault segments using synthetic earthquake catalogs. By taking the 28 June 1992 Landers earthquake occured in southern California as an example, the reconstructed fault segments are consistent with the faults already known on geological maps or blind faults that appeared quite frequently in longer-term catalogs.

Key words: **fault plane solution; small earthquake clustering; fuzzy clustering; principal component analysis; Landers earthquakes**

CLC number: P315.2 **Document code**: A

1 Introduction

In recent years, a large number of studies have devoted to physical mechanisms between the faulting and seismicity. However, the relations between them are still not very clear. Such study results have been proved important and effective to improve our understanding of seismic risk, especially that of potential destructive seismic events in the future. Seismic data include much underground information and study on the seismic data is one of the most important ways to invert for seismotectonic map.

Most of information available on the fault distribution comes from surface mapping with different scales. However, the information only from the surface rupture is often incomplete because earthquake catalogs generally include large quantities of events which seem associated with blind faults instead of those we can see on Earth surface. For example, the community fault model (CFM) of the Southern California Earthquake Center (SCEC), obtained a three-dimensional structure with strike-slip, blind-thrust, and oblique-reverse faults of southern California by integratively using surface traces, seismicity, seismic reflection profiles, borehole data, and other subsurface imaging techniques (Plesch and Shaw, 2002). In general, a fault only represented by a simple surface cannot meet the needs of the fine structure in fault zones and in drilling experiments of active faults (Scholz, 2002; Faulkner et al., 2003). These results suggest that fault zone should be actually consisted of narrow earthquake-generating cores, possibly accompanied by small subfaults. Detailed structure and seismicity analyses also revealed that many events still

* **Received** 22 November 2011; **accepted in revised form** 6 March 2012; **published** 10 April 2012.
♦ **Corresponding author. e-mail**: 13832667401@126.com

cannot be attributed to any known brittle structure (Guzofski et al., 2007). Using distribution of the small earthquakes, Wan et al. (2008) proposed an algorithm to determine fault parameters and regional stress field and applied it to Tangshan earthquake sequence. However, their aftershocks clusters partition method is subjective for only using naked eyes. Wang et al. (2008, 2010a) also proposed some methods for fault plane parameters estimation based on aftershocks, but these methods deal with multiple fault planes simultaneously. Ouillon and Sornette (2011) applied Gaussian Mixture Modelling(GMM) and expectation maximization(EM) algorithm to divide the Mount Lewis event (1986, M_{L}=5.7) in California, however, the procedure is quite complex.

It is necessary to use more effective and objective data mining tools for reliable seismological interpretations. The clustering analysis is an efficient tool to decrease the dimensionality of information and to extract hidden information and pattern included in huge amount of observations (Berkhin, 2002), and to reduce the uncertainty of the focal mechanism solutions in many other applications (Ekström and England, 1989, Hardebeck and Shearer, 2002; Wang et al., 2009).

Hard clustering and fuzzy clustering are both very important data mining methods. The main difference between them is as follows. In the hard clustering, each datum can belongs to one and only one cluster, while in fuzzy clustering, each datum belongs to different clusters with different membership degrees. In this paper, we intend to examine the efficiency of Gustafson and Kessel (GK) fuzzy clustering algorithm by clustering the partial data of Landers aftershock sequence.

We propose a general method to identify and locate active faults in seismically active region by using a rigorous approach based on mathematics and seismicity, aiming to gain a better understanding of the relationship between fault structures and earthquakes.

In section 2, the principles of fuzzy clustering and mathematical frame of unsupervised clustering algorithm proposed by Gustafson and Kessel (1979) are reviewed. In section 3, we test the effectivity of this new method through a synthetic data set. In section 4, we apply GK fuzzy clustering and principal component analysis method to partial Landers aftershock sequence and present the number of clusters and parameters of sub-fault planes corresponding to the clusters. In section 5, the discussion and conclusion are given, and the potential use is also discussed.

2 Fuzzy clustering methods

In order to divide the aftershocks sequence data into several clusters, we firstly introduce fuzzy clustering method. In this paper, we assume that the data set $\boldsymbol{X}$ consists of latitude, longitude and depth. Because clusters can be seen as subsets of the data set, clustering methods can be classified into fuzzy or hard clustering methods according to whether the subsets are fuzzy or hard. Hard clustering method is based on classical set theory, i.e., an object only belong to a single cluster. Hard clustering in a data set $\boldsymbol{X}$ means that the data set is into several mutually exclusive subsets, as we see in the traditional K-means clustering method. Fuzzy clustering method allows an object belong to several clusters simultaneously, with different membership degrees. In fact, fuzzy clustering is generally closer to practice than hard clustering, because objects on the boundaries between several clusters are not always included to one of the clusters, but usually belong to several clusters with different membership degrees between 0 and 1. Because objective function of hard clustering method isn't differentiable, it is difficult to design a fast convergence and high quality algorithm.

2.1 Fuzzy c-means (FCM) clustering algorithm

The fuzzy c-means clustering algorithm is to minimize an objective function called c-means function. It is defined as

$$\min J_m(\boldsymbol{X};\, \boldsymbol{U},\, \boldsymbol{V},\, \boldsymbol{A}) = \sum_{i=1}^{c}\sum_{k=1}^{n} (\mu_{ik})^m (D^2_{ikA}) \quad (1)$$

$$s.t.\ \mu_{ik} \in [0,1],\ \text{for}\ \forall i,\ k;\ \sum_{i=1}^{c} \mu_{ik} = 1,\ \text{for}\ \forall k;$$
$$0 < \sum_{k=1}^{n} \mu_{ik} < n,\ \text{for}\ \forall i; \quad (2)$$

where $\boldsymbol{U}=(\mu_{ik})_{c\times n}$ is partition matrix, μ_{ik} is membership function value between the k-th object and i-th fuzzy subset of $\boldsymbol{X}$, the sum of each column of $\boldsymbol{U}$ is 1, $\boldsymbol{V}=[\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \cdots,\ \boldsymbol{v}_c]$, $\boldsymbol{v}_i \in R^p$ is a vector group of cluster centers, which have to be computed, and $D^2_{ikA}=\| \boldsymbol{x}_k - \boldsymbol{v}_i \|^2_A = (\boldsymbol{x}_k - \boldsymbol{v}_i)^{\mathrm{T}} A (\boldsymbol{x}_k - \boldsymbol{v}_i)$ is a squared inner-product distance norm.

Statistically, equation (1) can be seen as a measure of the total variance from $\boldsymbol{x}_i$ to $\boldsymbol{v}_k$. The minimization of the c-means function (1) is a nonlinear optimization

problem that can be solved by various optimization methods, such as simulated annealing, genetic algorithms and particle swarm algorithm. However, the most popular method is Picard iteration. This method tries to find stationary points of objective function (1) by the first-order conditions. This method is called as the fuzzy c-means (FCM) algorithm.

The stationary points of the objective function (1) can be found by adjoining the constraint (2) to J_m by means of the Lagrange multiplier method:

$$\bar{J}(\boldsymbol{X};\boldsymbol{U},\boldsymbol{V},\boldsymbol{A},\boldsymbol{\lambda}) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m D_{ikA}^2 + \sum_{k=1}^{n}\lambda_k\Big(\sum_{i=1}^{c}\mu_{ik}-1\Big) \tag{3}$$

and by setting gradients of $\bar{J}$ with respect to $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{\lambda}$ as zero. If $D_{ikA}^2>0$, then for $\forall i$, k and $m>1$, ($\boldsymbol{U}$, $\boldsymbol{V}$) may minimize equation (1) only if

$$\mu_{ik} = \begin{cases} \dfrac{1}{\sum\limits_{j=1}^{c}\left(\dfrac{D_{ikA}}{D_{jkA}}\right)^{\frac{2}{m-1}}}, & D_{jkA}\neq 0,\ i\neq j \\ 1, & D_{jkA}=0,\ i=j. \end{cases} \tag{4}$$

and

$$\boldsymbol{v}_i = \frac{\sum\limits_{k=1}^{n}\mu_{ik}^m\boldsymbol{x}_k}{\sum\limits_{k=1}^{n}\mu_{ik}^m},\ i=1,\,2,\,\cdots,\,c. \tag{5}$$

This solution also satisfies the remaining constraints of equation (2). The $\boldsymbol{v}_i$ in equation (5) is weighted mean of the i-th cluster. The weights are the membership degrees. That is why the algorithm is called "c-means". In fact, the FCM algorithm is a simple iteration method through formulas (4) and (5).

The FCM algorithm computes with the standard Euclidean distance norm, which induces hyper-spherical clusters. Hence it can only detect clusters with the same shape and orientation. In this case, the common choice of norm inducing matrix is $\boldsymbol{A}=\boldsymbol{I}$. More generally, the norm inducing matrix $\boldsymbol{A}$ also can be chosen as a diagonal matrix that accounts for different variances in the directions of the coordinate axes of $\boldsymbol{X}$:

$$\boldsymbol{A}_D = \begin{bmatrix} (1/\sigma_1)^2 & 0 & \cdots & 0 \\ 0 & (1/\sigma_2)^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & (1/\sigma_p)^2 \end{bmatrix}. \tag{6}$$

In order to detect hyper-planar cluster, the norm inducing matrix $\boldsymbol{A}$ can be defined as the inverse of the $n\times n$ covariance matrix $\boldsymbol{A}=\boldsymbol{F}^{-1}$, with

$$\boldsymbol{F} = \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{x}_k-\bar{\boldsymbol{x}})(\boldsymbol{x}_k-\bar{\boldsymbol{x}})^{\mathrm{T}}, \tag{7}$$

where $\bar{\boldsymbol{x}}$ is defined as the sample mean of the data $\boldsymbol{X}$. In this case, $\boldsymbol{A}$ represents the Mahalanobis norm on $\boldsymbol{R}^p$. The algorithm is described in detail as follows.

For the data set $\boldsymbol{X}$, we choose the number of clusters $1<c<n$, the weighting exponent $m>1$, the termination tolerance $\varepsilon>0$ and the norm-inducing matrix $\boldsymbol{A}$, and then initialize partition matrix randomly to meet equation (2), finally repeat the below steps for l=1, 2, ...

Step 1 Compute the cluster means:

$$\boldsymbol{v}_i^{(l)} = \frac{\sum\limits_{i=1}^{n}\left(\mu_{ik}^{(l-1)}\right)^m\boldsymbol{x}_k}{\sum\limits_{i=1}^{n}\left(\mu_{ik}^{(l-1)}\right)^m},\ 1\le i\le c.$$

Step 2 Compute the distances:

$$D_{ikA}^2 = (\boldsymbol{x}_k-\boldsymbol{v}_i)^{\mathrm{T}}A(\boldsymbol{x}_k-\boldsymbol{v}_i),\quad 1\le i\le c,\ 1\le k\le n$$

Step 3 Update the partition matrix:

$$\mu_{ik}^{(l)} = \frac{1}{\sum\limits_{j=1}^{c}\left(\dfrac{D_{ikA}}{D_{jkA}}\right)^{\frac{2}{m-1}}}$$

Until

$$\left\|\boldsymbol{U}^{(l)}-\boldsymbol{U}^{(l-1)}\right\|<\varepsilon$$

When the algorithm terminates, the center of each cluster and partition matrix $\boldsymbol{U}$ are attained. If the number of the objects in the data set X is n, the clusters number is chosen as c, the iterations number is q, and partitioning the data set X with the standard FCM algorithm, then the time complexity is $O(qcn^2)$. Moreover, although FCM algorithm is an unsupervised clustering algorithm, it requires a priori knowledge about the number of clusters. Otherwise FCM algorithm will produce a misleading result. So in this case, the algorithm's unsupervised learning and adaptability is damaged.

2.2 Principles of Gustafson and Kessel (GK) fuzzy clustering algorithm

In order to overcome the drawbacks of the standard FCM algorithm, Gustafson and Kessel (1979) revised the FCM algorithm by employing an adaptive distance norm, which can detects clusters of different geometrical shapes in one data set. Each of the clusters has its

own norm-inducing matrix $\boldsymbol{A}_i$, which yields the following inner-product norm:

$$D^2_{ikA_i} = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T \boldsymbol{A}_i (\boldsymbol{x}_k - \boldsymbol{v}_i), \quad 1 \le i \le c,\ 1 \le k \le n \tag{8}$$

The matrices $\boldsymbol{A}_i$ are used as optimization variables in the c-means function, allowing each cluster to adapt the distance norm to the local topological structure of the data. Let $\boldsymbol{A}$ denote a c-tuple of the norm-inducing matrices: $\boldsymbol{A}=(\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_c)$. The objective function of the GK algorithm is defined as

$$\min J_{GKm}(\boldsymbol{X}; \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}) = \sum_{i=1}^{c}\sum_{j=1}^{n} (\mu_{ij})^m D^2_{ik\boldsymbol{A}_i} \tag{9}$$

For a fixed $\boldsymbol{A}$, condition (2) can be directly applied. However, the objective function (9) cannot be directly minimized with respect to $\boldsymbol{A}_i$, since it is linear in $\boldsymbol{A}_i$. This means that J can be as small as desired by simply making $\boldsymbol{A}_i$ less positively definite. To obtain a feasible solution, $\boldsymbol{A}_i$ must be constrained in some way. The usual way is to constrain the determinant of $\boldsymbol{A}_i$. Allowing the matrix $\boldsymbol{A}_i$ to vary with its determinant fixed corresponds to optimizing the cluster's shape while its volume remains constant:

$$\|\boldsymbol{A}_i\| = \rho_i,\ \rho_i > 0 \tag{10}$$

where ρ_i is fixed for each cluster. Using the Lagrange multiplier method, the following expression for $\boldsymbol{A}_i$ is obtained:

$$\boldsymbol{A}_i = [\rho_i \det(\boldsymbol{F}_i)]^{\frac{1}{n}} F_i^{-1} \tag{11}$$

where F_i is the fuzzy covariance matrix of the i-th cluster defined as

$$\boldsymbol{F}_i = \frac{\sum_{i=1}^{n} (\mu_{ik})^m (\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^T}{\sum_{i=1}^{N} (\mu_{ik})^m}, \tag{12}$$

Substituting equations (11) and (12) into equation (8) yields a generalized squared Mahalanobis distance norm between $\boldsymbol{x}_k$ and the cluster mean $\boldsymbol{v}_i$, in which the covariance is weighted by the membership degrees in $\boldsymbol{U}$. For the purpose of avoiding that the cluster covariance matrices are singular, Babuška et al. (2002) proposed a modified Gustafson-Kessel algorithm. The algorithm is described as follows.

When the data set $\boldsymbol{X}$ is given, we choose the number of clusters $1<c<n$, the weighted exponent $m\in[1.5, 2.5]$, the termination tolerance $\varepsilon>0$ and the norm-inducing matrix $\boldsymbol{A}$, and then initialize the partition matrix randomly. Subsequently, we repeat the following steps for l=1, 2, $\cdots$

Step 1 Calculate the cluster centers.

$$\boldsymbol{v}_i^{(l)} = \frac{\sum_{i=1}^{n} (\mu_{ik}^{(l-1)})^m \boldsymbol{x}_k}{\sum_{i=1}^{n} (\mu_{ik}^{(l-1)})^m},\ 1 \le i \le c;$$

Step 2 Compute the cluster covariance matrices.

$$\boldsymbol{F}_i^{(l)} = \frac{\sum_{i=1}^{n} (\mu_{ik}^{(l-1)})^m (\boldsymbol{x}_k - \boldsymbol{v}_i^{(l)})(\boldsymbol{x}_k - \boldsymbol{v}_i^{(l)})^T}{\sum_{i=1}^{c} (\mu_{ik}^{(l-1)})^m}, \quad 1 \le i \le c$$

Add a scaled identity matrix:

$$\boldsymbol{F}_i := (1-\gamma)F_i + \gamma \det(F_0)^{1/n} I.$$

Extract $\boldsymbol{F}_i$ from eigenvalues λ_{ij} and eigenvectors ϕ_{ij}, find $\lambda_{i\max}=\max_{1\le j\le n} \lambda_{ij}$ and set $\lambda_{i\max}=\lambda_{ij}/\beta$, $\lambda_{i\max}/\lambda_{ij}>\beta$ for $\forall j$. In this paper, we set $\beta=10^{15}$.

Reconstruct $\boldsymbol{F}_i$ by

$$\boldsymbol{F}_i = [\varphi_{i1}, \varphi_{i2}, \cdots, \varphi_{in}]\mathrm{diag}(\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{in}) \cdot [\varphi_{i1}, \varphi_{i2}, \cdots, \varphi_{in}]^{-1},$$

Step 3 Compute the distances.

$$D^2_{ikA}(\boldsymbol{x}_k, \boldsymbol{v}_i) = (\boldsymbol{x}_k - \boldsymbol{v}_i^{(l)})^T [\rho_i \det(F_i)^{1/n} F_i^{-1}] \cdot (\boldsymbol{x}_k - \boldsymbol{v}_i^{(l)}),\ 1 \le i \le c,\ 1 \le k \le n.$$

Step 4 Update the partition matrix

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} \left[\frac{D_{ikA_i}(x_k,\ v_i)}{D_{jkA_i}(x_k,\ v_j)}\right]^{2/(m-1)}}, \quad 1 \le i \le c,\ 1 \le k \le n.$$

until

$$\left\|\boldsymbol{U}^{(l)} - \boldsymbol{U}^{(l-1)}\right\| < \varepsilon$$

2.3 Cluster validity

It is very important to determine the cluster numbers (Pal and Bezdek, 1995; Bensaid et al., 1996; Xie and Beni, 1991) in advance. When the number of clusters is fixed, Bensaid et al. (1996), Xie and Beni (1991) proposed cluster validity indexes for finding the best partition of the data set $\boldsymbol{X}$ by the GKFCM algorithm.

Although there are several different validity indexes proposed in the literatures, none of them can copy with all the situations. Therefore, we use three indexes

in this paper, in order to have mutual confirmation each other. These indexes are described as follows:

Partition index (S_{C}): Partition index (S_{C}) is the ratio of the sum of compactness to separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster (Bensaid et al., 1996).

$$S_{\mathrm{C}}(c)=\sum_{i=1}^{c}\frac{\sum_{j=1}^{n}(\mu_{ij})^{m}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2}{n_i\sum_{k=1}^{c}\|\boldsymbol{v}_k-\boldsymbol{v}_i\|^2}. \tag{13}$$

S_{C} is useful when different partitions have equal number of clusters. A lower value of S_{C} indicates a better partition.

Separation index (S): On the contrary of partition index (S_{C}), the separation index S uses a minimum-distance separation for partition validity (Bensaid et al., 1996).

$$S(c)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^{m}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2}{n\min\limits_{i,k}\|\boldsymbol{v}_k-\boldsymbol{v}_i\|^2}. \tag{14}$$

Xie and Beni's index (XB): This index aims to quantify the ratio of the total variation within clusters to the separation of clusters (Xie and Beni, 1991).

$$S_{\mathrm{XB}}(c)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^{m}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2}{n\min\limits_{i,j}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2}. \tag{15}$$

The optimal number of clusters should minimize the values of these validity indexes.

In order to assess the goodness of the obtained partitions, we compute the above three validity indexes for different numbers of clusters c. After an upper bound $c_{\max}$ of c is estimated, the GKFCM algorithm and the above three validity indexes have to run with each c=2, 3, $\cdots$, $c_{\max}$. We compare directly with the validity indexes and select the most suitable numbers of the clusters. The criterion is that a lower index value indicates a better partition of the data set X.

3 Simulations

To test whether the algorithm is capable to discover the characteristics of faults related to earthquake sequence, we simulate an earthquake sequence using synthetic catalogs. In 3D space, two hundreds earthquakes in synthetic catalogs distribute randomly uniformly on three vertical planes, two of which are parallel to each other and the both are perpendicular to the third plane, each plane is 20 km long and 10 km wide (Figure 1). We assume each earthquake have distance error to this plane no more than 100 m. We then compute cluster validity index as shown in Figure 2. It can be seen the appropriate cluster number is three or four.

Figures 3 and 4 show simulating results for cluster number of 3 and 4 respectively. We can find that the fault planes by simulation are consistent with the three planes. In Figure 4, two of the assumed planes are found to be consistent with simulation fault, while the third one has simply been split into two sub-planes, but the positions and orientations of all planes still show a very nice inversion of the data, which would not modify a tectonic interpretation.

We find that the strike, dip, length, width and center of the reconstructed fault structure have high accuracy for each case. Our synthetic test has been performed under the condition of seismicity uniformly distributed on the fault planes.

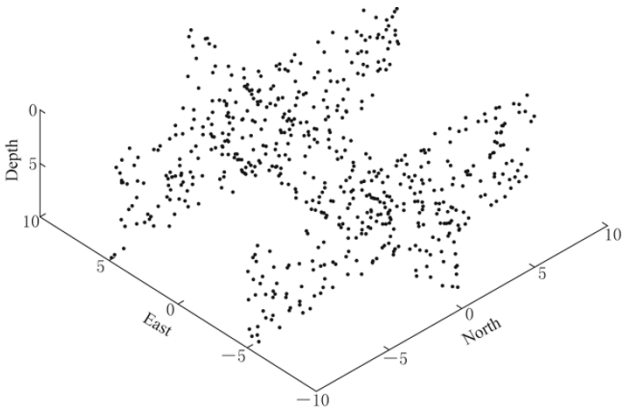

Figure 1 The simulated 3-D data set.

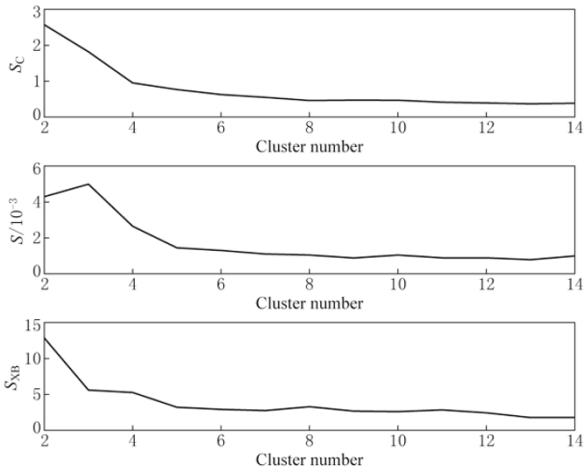

Figure 2 Values of partition index and separation index, Xie and Beni's index.

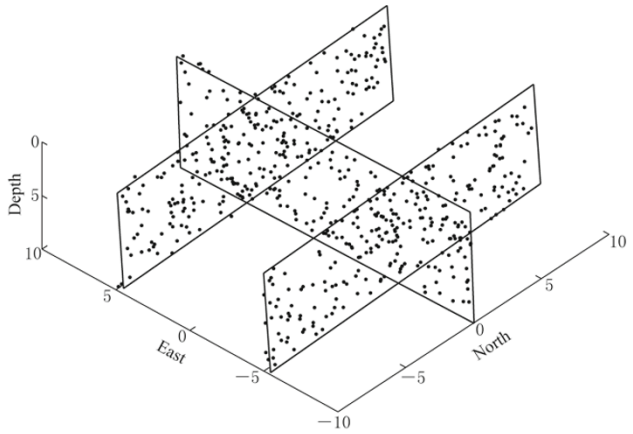

Figure 3 Simulation result with three planes.

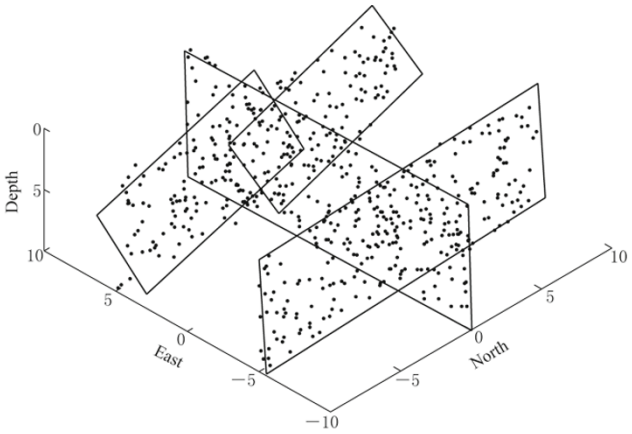

Figure 4 Simulation result with four planes.

4 Applications to the Landers aftershock sequence

On the basis of the above-mentioned high accuracy simulation, we apply this method to 1992 Landers earthquake occurred in southern California. We use the locations provided by Lin et al. (2009) and consider only 5000 events occurred in the region: latitude from 34.3° to 34.7° and longitude from −116.7° to −116.3° for computation time saving.

4.1 Aftershocks data processing

For each event, we define $(\theta_i,\ \varphi_i,\ h_i)$ to represent its epicentral location, where θ_i is latitude in degree, φ_i is longitude in degree, and h_i is focal depth in kilometer. Before clustering analysis, the units of latitude and longitude must be translated into kilometer. The approximate conversion formula is as follows:

The length of each degree in latitude is approximated as c_l=111.199 km, the means of latitude, longitude and depth are respectively expressed as $\theta_0=\frac{1}{n}\sum_{i=1}^{n}\theta_i$, $\varphi_0=\frac{1}{n}\sum_{i=1}^{n}\varphi_i$ and $h_0=\frac{1}{n}\sum_{i=1}^{n}h_i$, then $\theta_r=(\theta_i+\theta_0)\times\pi/360$, $x_i=c_l(\varphi_i-\varphi_0)\cos(\theta_r)$, $y_i=c_l(\theta_i-\theta_0)$, $z_i=h_i-h_0$, $i=1,\ 2,\ \cdots,\ n$.

4.2 Method for parameters determination of fault plane

We employ GK-FCM cluster analyses on data set $D=\{(x_i,\ y_i,\ z_i),\ i=1,\ 2,\ \cdots,\ n\}$ and obtain a series of sub-fault planes. The parameters of fault planes can be determined as follows. For each sub-data set $S_i=\{\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_n\}$, in which $\boldsymbol{x}_j=\{(x_i,\ y_i,\ z_i),\ j=1,\ 2,\ \cdots,\ n_i\}$, we run principal compnents analysis (PCA), compute $\hat{\boldsymbol{\mu}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j$, $\hat{\Sigma}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}(\boldsymbol{x}_j-\hat{\boldsymbol{\mu}}_j)^{\mathrm{T}}(\boldsymbol{x}_j-\hat{\boldsymbol{\mu}}_j)$. Diagonalizing the covariance matrix $\hat{\boldsymbol{\Sigma}}_i$ can get three eigenvalues $\lambda_1\geq\lambda_2\geq\lambda_3\geq0$ and their associated eigenvectors $\boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \boldsymbol{u}_3$. The largest eigenvalue λ_1 gives information on the largest dimension of the cluster (thereafter considered as its length of fault), and its associated eigenvector $\boldsymbol{u}_1$ provides the direction of vector. The second largest eigenvalue λ_2 (and its eigenvector $\boldsymbol{u}_2$) gives the width of the cluster (and direction), while the smallest eigenvalue λ_3 (and its eigenvector $\boldsymbol{u}_3$) gives information on the thickness of the cluster (and direction).

The relationship between an eigenvalue and the dimension of the fault along the associated direction is related to the specific distribution of data points on the plane. In the following, for simplicity, we assume that data points are distributed uniformly over a plane. It is easily shown that if random variable ξ is a 1-D random variable distributed uniformly over the interval [0, L], then the variance of random variable ξ is $\xi=L^2/12$,, and L can be estimated as $L=\sqrt{12\xi}$. We extend the above derivation to a 2-D space, and obtain L is equal to $\sqrt{12\lambda_1}$ and W is equal to $\sqrt{12\lambda_2}$ if events are uniformly distributed over a fault of length L and width W.

The square root of the third eigenvalue is the standard deviation of the event locations in the direction perpendicular to the fault plane. Therefore, for the simulated fault plane can agree to real fault plane as better as possible, the sum of all λ_3 values for each clustering should be minimum. This procedure defines a set of fault-like objects in the most natural way possible. For each cluster, the third eigenvector $\boldsymbol{u}_3$ is sufficient to determine the strike and dip of the fault.

If fault plane's strike is $\phi\,(0\leq\phi\leq2\pi)$ and dip is $\delta\,(0\leq\delta\leq\pi/2)$, its unit normal vector is $\boldsymbol{n}^0=(\sin\varphi\sin\delta,\ -\cos\varphi\sin\delta,\ \cos\delta)$ (Wan et al., 2000). Ac-

cording to principal components analysis (PCA), $\boldsymbol{u}_3=(u_{31}, u_{32}, u_{33})$ is another unit normal vector of fault plane and $\boldsymbol{u}_3$ is parallel to $\boldsymbol{n}^0$. So, when $u_{33}>0$, let $\boldsymbol{u}_3=\boldsymbol{u}^0$ and $u_{33}<0$, let $-\boldsymbol{u}_3=\boldsymbol{u}^0$. Suppose $u_{33}>0$ and then $\boldsymbol{u}_3=\boldsymbol{u}^0$, therefore, strike ϕ and dip δ can be computed by the following formula:

$$\delta = \arccos(u_{33}),$$

$$\varphi = \begin{cases} \pi - \arctan(u_{31}/u_{32}), & u_{32} > 0, \\ 2\pi - \arctan(u_{31}/u_{32}), & u_{31} < 0, u_{32} < 0, \\ -\arctan(u_{31}/u_{32}), & u_{31} > 0, u_{32} < 0, \\ \pi/2, & u_{31} > 0, u_{32} = 0, \\ 3\pi/2, & u_{31} < 0, u_{32} = 0. \end{cases} \quad (16)$$

The mean of $\hat{\boldsymbol{\mu}}_i$, λ_1, $\boldsymbol{u}_1$, λ_2 and $\boldsymbol{u}_2$ can be used to estimate the center, length, width and edges and four vertexes of fault plane, respectively.

After the above parameters are solved, we convert the parameters data (x, y, z) to origin format (θ, φ, h) by the following formulas $\theta=\theta_0+y/c_l$, $\varphi=\varphi_0+x/(c_l\cos(\theta_r))$, $h=z+h_0$.

4.3 Implementation

4.3.1 Determining the number of clustering

Before partition, we should firstly determine clustering number. Using Gustafson-Kessel algorithm in section 2.3, we set $m=2$, $\varepsilon=10^{-8}$, $\rho=1$ for each cluster, c is from 2 to 14, and obtain indexes values of S_{C}, S and S_{XB} as shown in Figure 5.

We have to mention again the clustering number using only one validation index is not reliable, thus it is necessary to obtain the optimum clustering number through comparison among the three results. We consider that partition with fewer clustering number is better, when the differences between the values of validation indexes are minor.

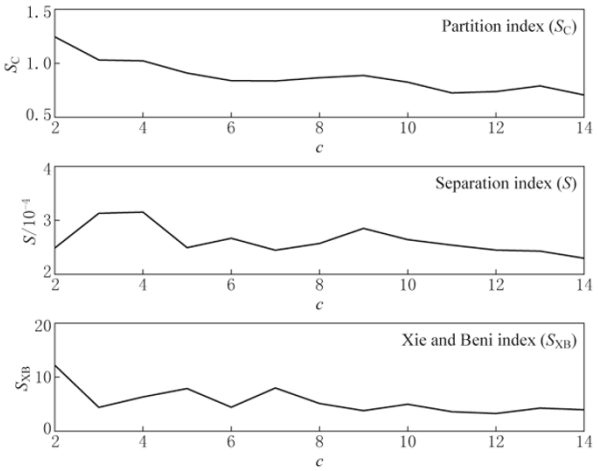

Figure 5 Values of partition index, separation index and Xie and Beni's index.

Figure 5 shows that S_{C} and S hardly decrease at the c=12. The S_{XB} index reaches this local minimum at c=11. Considering that S_{C} and S are more useful and comparing different clustering methods with the same c, we choose the optimal number of clusters to 12.

4.3.2 The sub-fault planes and parameters of partial Landers earthquake

When GK-FCM procedure is completed, we can get the membership matrix $\boldsymbol{U}$ of earthquake data set and divide the data set into c clusters based on $\boldsymbol{U}$. In some cases, one event belongs to one certain cluster. In practice, however we find this partition method is sensitive to outliers, that is to say, some events seem not to belong to any fault plane. To eliminate the outliers' effects, we give a threshold T=0.3 which is related to the cluster number. When the event's membership value to v_i is bigger than threshold value T, it should be assigned to the i-th cluster. In Figure 6b, the aftershock sequence and fault planes' projection on the ground distribution are given. The clusters aftershocks are marked with 'circles' and the outliers aftershocks are marked with 'cross'. The fault planes scope and aftershocks agree well.

Figure 6a shows the optimal 12 faults structure by fitting the data set of aftershock epicenters. The northern part now appears to be fitted very nicely, while the southern part looks quite complex. It is important to realize that some fault planes are nearly vertical, which make them barely visible in the projection shown in Figure 6. In order to describe the fault structure, it is convenient to label the 12 faults from A to L, which allows us to discuss this pattern fault by fault. The parameters of the 12 fault planes (position, size and orientation) are given in Table 1.

These fault planes will now be classified into three different categories, namely, (1) misleading planes (which have no apparent significance), (2) previously known planes (that correspond to mapped faults), and (3) unknown planes (that may correspond to blind faults or structures unmapped for whatever reason).

Table 1 reveals that most planes dip close to vertical. Two planes, E and H, have rather abnormal dips, which leads us to suspect that they are misleading. Indeed, E and H are nearly perpendicular to plane A, G and J, which are located in a zone with rather weak seismicity in the direction perpendicular to those planes. It is likely that those planes have no tectonic significance, and are found due to the algorithm to satisfy the

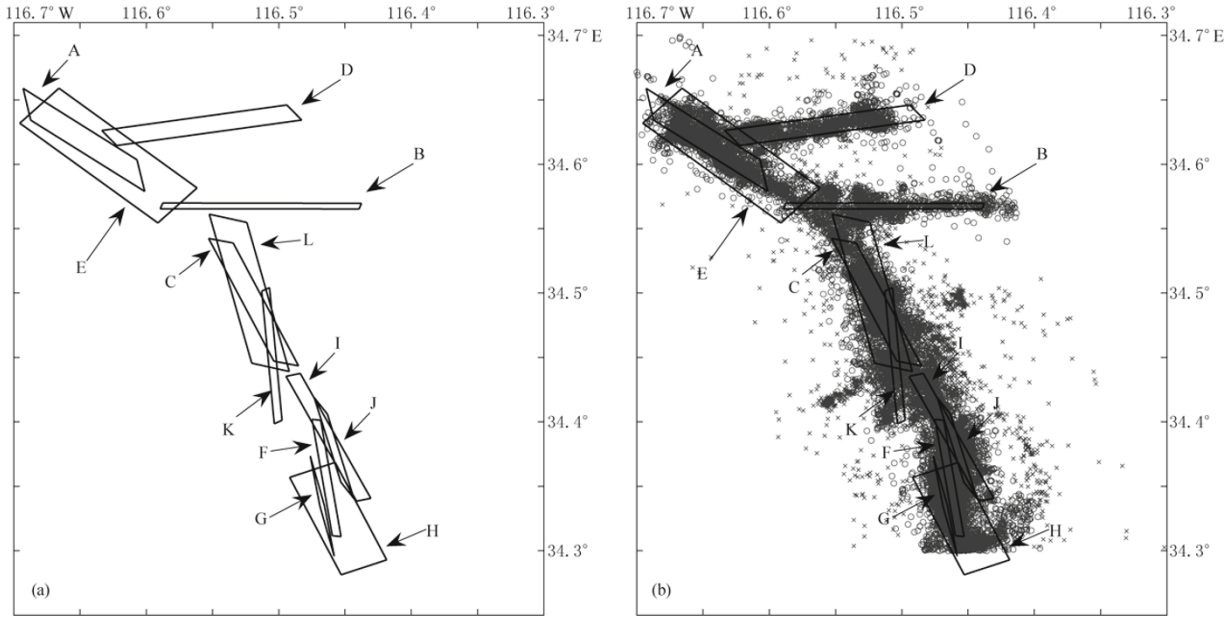

Figure 6 Simulation with 12 fault planes of the partial 1992 Landers aftershock data set. (a) Simulation with twelve fault planes. (b) Epicentral distribution of the full catalog available for the Landers area, in which the clusters aftershock are marked with circles, and those located in boundary are maked with cross.

Table 1 Comparison between fitting planes of the Landers earthquake sequence obtained with the GKFCM method and known faults in the same area

Plane label	Fault name	Long. /°W	Lat. /°N	Depth /km	Strike /°	Dip /°	Length /km	Width /km	$\sqrt{\lambda_3}$/km	Event number
A	Emerson	116.65	34.62	6.66	313.4	73.5	9.18	4.53	0.91	1 773
B	blind	116.51	34.57	6.40	0.0	87.1	11.63	6.76	0.76	1 255
C	Emerson	116.52	34.49	6.99	119.7	67.7	9.07	2.98	0.57	1 313
D	blind	116.56	34.63	6.47	188.4	80.3	11.02	7.27	0.58	1 971
E	misleading fault	116.63	34.61	3.89	318.0	38.4	10.38	3.70	0.95	1 913
F	Homestead valley	116.46	34.36	3.32	277.8	82.3	8.25	4.15	0.60	3 023
G	Homestead valley	116.47	34.33	8.28	99.5	87.4	6.35	4.14	0.40	3 439
H	misleading fault	116.45	34.32	9.43	298.0	19.0	7.56	2.69	0.68	1 790
I	Homestead Fault	116.46	34.39	8.90	113.0	76.0	8.61	2.98	0.47	2 522
J	Homestead valley	116.46	34.38	5.17	285.0	84.7	5.75	3.75	0.52	2 725
K	Emerson	116.51	34.45	4.95	273.1	80.1	9.46	4.74	0.90	1 233
L	Emerson	116.52	34.51	9.50	69.1	69.1	8.57	3.86	0.79	1 186

Note: Each fitting plane is named with the label given in Figure 6. The name of the corresponding natural fault (if it is known) is given. Misleading indicates that the plane has no tectonic significance, while blind means that the natural fault does not intersect the free surface and was not known before this study. Each plane is characterized by the latitude, longitude, and depth of its epicenter, as well as by its strike, dip, and dimensions in kilometer. The values are given of λ_3 of the corresponding cluster and the total number of events in that cluster.

criterion. All other 10 fault planes out of the 12 have dips larger than 50°, so we have no reason to doubt their existence. Figure 6 shows the 2-D epicentral distribution of the two misleading faults and the other 10 planes.

5 Discussion and conclusions

We introduce a new powerful method based on GK-FCM and PCA. The method allows us to divide earthquake epicenters data into distinct clusters which

can be interpreted as fault planes. The method is first applied to synthetic data set, and confirmed be able to recover correctly the existing faults with high accuracy. Subsequently, the method is applied to the partial aftershocks sequence of the 1992 Landers earthquake in California. It is possible that by using this method we can find not only the fault which is in accordance with that shown by surface mapping, but also some blind faults. It should be emphasized that, although the clustering analyses are unsupervised methods and powerful data mining tools which can be used for better and more reliable interpretations about observations, it is not enough only using the method for a reliable and correctly geological and seismological interpretation, field investigation should also be considered. These methods only provide mathematical, objective tools for better geological and seismological interpretations.

Because the focal depth of aftershocks have significant impact on the dips, and the depths of aftershocks are not very precise in general, thus the dips have lower accuracy than strikes. We can use bootstrap method to determine the parameters errors in the future.

Triggered earthquakes are able to trigger other earthquakes and small earthquakes can trigger large earthquakes (Ogata 1998, Helmstetter 2003, Wang et al., 2010b). Using the probabilities in the partition matrix, fuzzy clustering may also help scientists to better understand the bias physics of earthquake triggering.

In our procedure, few aftershocks may be attributed to two even more adjacent faults when the threshold value T is too small. In fact, the threshold value T depends on both data distribution and cluster number c. This topic is interesting and will be discussed in the future.

The faults structure in the vicinity of large event is helpful to test different hypothesis on stress mechanisms in seismic hazard assessment and earthquake forecasting (Wan et al., 2002, 2004 and 2010). Moreover, imaging accurately 3-D structure of faults make it possible to understand static or dynamic earthquake triggering, as well as the mechanics of faulting in time scale up to a few decades.

Acknowledgements We gratefully acknowledge the financial support of the Teachers Scientific and Research Fund of China Earthquake Administration (20090126), the Natural Science Fund of Hebei Province (A2011408006) and the Fundamental Research Funds for the Central Universities (ZY20110101) as well as the helpful comments from anonymous referee, and the editor.

References

Babuška R, Van der Veen P J and Kaymak U (2002). Improved covariance estimation for Guastafson-Kessel clustering. In: *Proceedings of* 2002 *IEEE International Conference on Fuzzy System* Honolulu, Hawaii, 1 081–1 085.

Bensaid A M, Hall L O, Bezdek J C, Clarke L P, Silbiger M L, Arrington J A and Murtagh R F (1996). Validity-guided (Re) clustering with applications to image segmentation. *IEEE Transactions on Fuzzy System* **4**(2): 112–123.

Berkhin P (2002). Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 56.

Ekström G and England P (1989). Seismic strain rates in regions of distributed continental deformation. *J Geophys Res* **94**(B8): 10 231–10 257.

Faulkner D R, Lewis A C and Rutter E H (2003). On the internal structure and mechanics of large strike-slip fault zones: Field observations of the Car boneras fault in southeastern Spain. *Tectonophysics* **367**: 235–251.

Gustafson D E and Kessel W C (1979). Fuzzy clustering with fuzzy covariance matrix. *Proceedings of the IEEE CDC*, San Diego, California, 761–766.

Guzofski C A, Shaw J H, Lin G and Shearer P M (2007). Seismically active wedge structure beneath the Coalinga anticline, San Joaquin basin, California. *J Geophys Res* **112**: B03S05, doi:10.1029/2006JB004465.

Hardebeck J L and Shearer P M (2002). A new method for determining first-motion focal mechanisms. *Bull Seismol Soc Am* **92**: 2 264–2 276.

Helmstetter A (2003). Is earthquake triggering driven by small earthquakes? *Phys Rev Let* **91**(5): 058501.

Lin G, Shearer P M and Hauksson E (2009). LSH: an earthquake relocation catalog using southern California pick and waveform data. [2011-11-11]. http://www.rsmas.miami.edu/personal/glin/LSH.html.

Ogata Y (1998). Space-time point-process models for earthquake occurrences. *Ann Inst Stat Math* **50**: 379–402, doi:10.1023/A:1003403601725.

Ouillon G and Sornette D (2011). Segmentation of fault networks determined from spatial clustering of earthquakes. *J Geophys Res* **116**: B02306, doi:10.1029/2010JB007752.

Pal N R and Bezdek J C (1995). On cluster validity for the fuzzy C-means model. *IEEE Transactions on Fuzzy Systems* **3**(3): 370–379.

Plesch A and Shaw J H (2002). SCEC 3D community fault model for southern California. *Eos Trans AGU* **83**(47), Fall Meet. Suppl. Abstract S21A-0966.

Scholz C (2002). *The Mechanics of Earthquakes and Faulting*. Cambridge University Press, New York.

Wan Y G and Shen Z K (2010). Static coulomb failure stress changes on faults caused by the 2008 M_W7. 9 Wenchuan,

China earthquake. *Tectonophysics* **491**: 105–118.

Wan Y G, Shen Z K, Diao G L, Wang F C, Hu X L and Sheng S Z (2008). An algorithm of fault parameter determination using distribution of small earthquakes and parameters of regional stress field and its application to Tangshan earthquake sequence. *Chinese J Geophys* **51**(3): 569–583 (in Chinese with English abstract).

Wan Y G, Wu Z L and Zhou G W (2004). Focal mechanism dependence of static stress triggering of earthquakes. *Tectnophysics* **390**: 235–243.

Wan Y G, Wu Z L and Zhou G W and Huang J (2000). How to get rake angle of the earthquake fault from known strike and dip of the two nodal planes. *Seismological and Geomagnetic Observztion and Research* **21**(5): 26–30 (in Chinese with English abstract).

Wan Y G, Wu Z L, Zhou G W, Huang J and Qin L (2002). Global test of seismic static stress triggering model. *Acta Seismologica Sinica* **15**(3): 318–332.

Wang F C, Cao H R and Wan Y G (2010a). Application of linear errors-in-variables model for determination of main earthquake's fault parameters of Wenchaun earthquake. *Journal of Applied Statistics and Management* **29**(3): 381–390 (in Chinese with English abstract).

Wang F C, Wan Y G and Hu S T (2008). Application of particle swarm optimization to the estimation of mainshock fault plane parameters. *J Seism Res* **31**(2): 149–154 (in Chinese with English abstract).

Wang Q, Jackson D D and Kagan Y Y (2009). California earthquakes, 1800–2007: A unified catalog with moment magnitudes, uncertainties and focal mechanisms. *Seism Res Lett* **80**: 446–457, doi:10.1785/gssrl.80.3.446.

Wang Q, Jackson D D and Zhuang J (2010b). Missing links in earthquake clustering models. *Geophys Res Lett* **37**: L21307.

Xie X L and Beni G (1991). A validity measure for fuzzy clustering. *IEEE Trans. PAMI*, Aug. **13**(8): 841–847.