ELSEVIER

Journal of Structural Geology 28 (2006) 991–997

JOURNAL OF STRUCTURAL GEOLOGY

www.elsevier.com/locate/jsg

Computerized identification of stress tensors determined from heterogeneous fault-slip data by combining the multiple inverse method and *k*-means clustering

Makoto Otsubo *, Katsushi Sato, Atsushi Yamaji

Division of Earth and Planetary Sciences, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan

Abstract

The multiple inverse method is a resampling technique that can separate stresses from heterogeneous fault-slip data. Numerous optimal stresses are determined for each extracted subset of data, and the clusters of these stresses are thought to represent significant solutions. Hitherto, the clusters have had to be visually recognized on stereonets. This study computerized the identification of the clusters by using the *k*-means clustering technique. We tested the technique using artificial datasets with known solutions. As a result, it was found that the present method detected objectively the correct solutions. In addition, the spread of each cluster was evaluated to indicate the confidence levels of the identified stresses that were represented by the cluster centers.

Keywords: Stress tensor; *k*-means; Multiple inverse method; σ-space; Stress difference; Meso-scale fault

1. Introduction

Stress tensor inversion of fault-slip data (e.g. Angelier, 1979) has been applied to many areas in the world since the early 1980s to understand paleostresses in the upper crust. Several numerical techniques have been proposed for separating stresses from heterogeneous fault-slip data (Angelier, 1994; Nemcok and Lisle, 1995; Fry, 1999; Shan et al., 2003, 2004; Yamaji, 2003a; Yamaji et al., 2006). The multiple inverse method (Yamaji, 2000b) is one of them, and has revealed stress history in active island arcs with a temporal resolution of <1 m.y. (Yamaji, 2000a, 2003b; Yamaji et al., 2003, 2005).

The multiple inverse method iteratively resamples k_f-element subsets (usually $k_f = 4$, 5 or 6) from a set of fault-slip data, and determines optimal stress tensors for the subsets. The optimal solutions are represented by reduced stress tensors. The number of subsets equals the binomial coefficient ${}_{N_f}C_{k_f} = N_f!/k_f!(N_f - k_f)!$, where N_f is the number of fault-slip data. Significant stresses are indicated by clusters on stereograms that show principal stress orientations and stress ratios. The clusters have been recognized so far by visual inspection of the stereograms. Visual recognition is more or less subjective, unless distinctive clusters are observed.

When a single cluster of reduced stress tensors appears, it is possible to determine the cluster's representative stress by taking the component-wise average of the tensors and to evaluate the spread of the reduced stress tensors about the representative one (Yamaji et al., 2005). However, the purpose of the multiple inverse method is to detect multiple stresses from heterogeneous data. Given a heterogeneous dataset, the method yields several clusters of reduced stress tensors. In this case, the representative tensors must be determined for each of the clusters.

The purpose of this study is to present a computerized technique to recognize those clusters separately and to determine the representative reduced stress tensors and the spread of tensors. To this end, we employ a technique called *k*-means clustering (MacQueen, 1967; Lloyd, 1982) for the objective division of reduced stress tensors obtained by the multiple inverse method into clusters. At the moment, the number of clusters, *k*, has to be specified by the user.

Clustering was conducted in the parameter space defined by Sato and Yamaji (2006) who reshaped that of Fry (1999). The *k*-means clustering needs a well-defined distance between the objects to be classified. The stress difference defined by Orife and Lisle (2003) is a useful distance between reduced stress tensors. The parameter space is suitable for our purpose because the Euclidean distance between points in the parameter

* Corresponding author. Tel.: +81 75 753 4150; fax: +81 75 753 4189.
E-mail address: otsubo_m@kueps.kyoto-u.ac.jp (M. Otsubo).

0191-8141/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsg.2006.03.008

space equals the stress difference between the stresses that are represented by the points.

The present technique was tested by artificial datasets. It was shown that the resolution of the visual identification of clusters was sometimes insufficient, and that the present technique detected correct stresses from artificial data that were generated with known stresses.

2. Recognition of clusters of stress tensors

2.1. k-means clustering

The core of the present method is the application of *k*-means clustering (MacQueen, 1967; Lloyd, 1982) to stress tensors. Clustering is the process of organizing objects into groups. Any two objects from a group should be similar in some way, while those belonging to different groups should be dissimilar. The *k*-means clustering explores the partition of objects that minimizes the sum of squared distances between them and cluster centers.

Our objects are the reduced stress tensors. The following parameter space enables us to define the distance and cluster center for the clustering of the tensors. The reduced stress tensors have one-to-one correspondence with points on the five-dimensional unit hypersphere (Sato and Yamaji, 2006). The Euclidean distance in this five-dimensional parameter space equals the stress difference between corresponding reduced stress tensors (Sato and Yamaji, 2006). This dissimilarity between the tensors ranges from zero to two (Orife and Lisle, 2003). Therefore, we deal with points on the hypersphere in the same light as unit vectors whose initial points are fixed at the origin.

Let N be the number of points on the hypersphere and $\vec{x}^{(i)}$ be the point indicating the ith stress tensor, where $i = 1, 2, \ldots, N$. Our task is to automate the identification of clusters of those points. Suppose that the points are divided into k clusters, and that the cth cluster has N_c points. Therefore, we have

$$N = \sum_{c=1}^{k} N_{c^*}$$

The cluster center is defined as follows. We deal with the cluster center as the unit vector to indicate a reduced stress tensor. Let $\vec{u}_c$ be the center of cth cluster, and $\vec{x}_c^{(i)}$ be the ith vector in the cth cluster, where $i = 1, 2, \ldots, N_c$. The center is given by the equation

$$\vec{\mu}_c = \frac{1}{v}\left[\vec{x}_c^{(1)} + \cdots + \vec{x}_c^{(N_c)}\right] \quad (1)$$

where v is the normalizing factor, $v = \left|\vec{x}_c^{(1)} + \cdots + \vec{x}_c^{(Nc)}\right|$, and has the role to put the end point of $\vec{u}_c$ on the hypersphere.

The practical procedure of the *k*-means clustering technique for stress tensors is as follows:

(1) Input the number of clusters k ($k > 1$). Distribute the initial cluster centers $\vec{u}_c$ ($c = 1, 2, \ldots, k$) randomly.

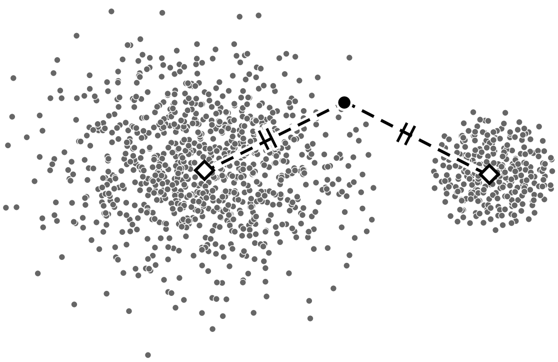

Fig. 1. Schematic picture showing the role of normalized distance. Closed circle is plotted at the same Euclidean distance from the cluster centers (open diamonds), but the point at the closed circle is more probably a member of the large cluster than of the small one. This study uses the distance of a point from a cluster center normalized by the size of the cluster.

(2) Calculate the distance between each vector $\vec{x}$ and cluster centers $\vec{u}_c$, and link the vectors to their nearest cluster centers.

(3) Update the positions of centers using Eq. (1) with linked vectors, $\vec{x}_c^{(1)}, \ldots, \vec{x}_c^{(N_c)}$.

(4) Steps (2) and (3) are repeated until the linkages no longer change.

To cope with various cluster sizes, the present study uses the normalized distance of the vector $\vec{x}$ from the cth cluster center

$$D_{\mathrm{w}}(\vec{x}, \vec{\mu}_c) = \frac{|\vec{x} - \vec{\mu}_c|}{s_c} \quad (2)$$

in step (2), instead of the distance $|\vec{x} - \vec{\mu}_c|$ itself, where s_c is a normalizing factor representing spread of the cth cluster

$$s_c^2 = \frac{1}{N_c - 1}\sum_{i=1}^{N_c}\left|\vec{x}_c^{(i)} - \vec{\mu}_c\right|^2 \quad (3)$$

where $\vec{x}_c^{(i)}$ is the ith member of the cth cluster. D_{w} indicates the deviation of a member from $\vec{u}_c$ with respect to the size of the cluster that is represented by s_c. The deviation of a member in a small cluster is evaluated to be greater than that of a member in a large cluster, even if the Euclidean distances are the same (Fig. 1).

2.2. Distribution of initial cluster centers

The results of *k*-means clustering vary, depending on the initial positions of the cluster centers. It is, therefore, necessary to evaluate and compare the results for determining the best division for the given data. For this purpose, we make a multivariate discriminant analysis. We use an extension of Fisher's (1936) linear discriminant for plural clusters. First, we quantify the dispersion of members of each cluster by the scatter matrix (Duda et al., 2001)

$$S_c = \sum_{i=1}^{N_c}\left[\vec{x}_c^{(i)} - \vec{\mu}_c\right]\left[\vec{x}_c^{(i)} - \vec{\mu}_c\right]^{\mathrm{T}} \quad (4)$$

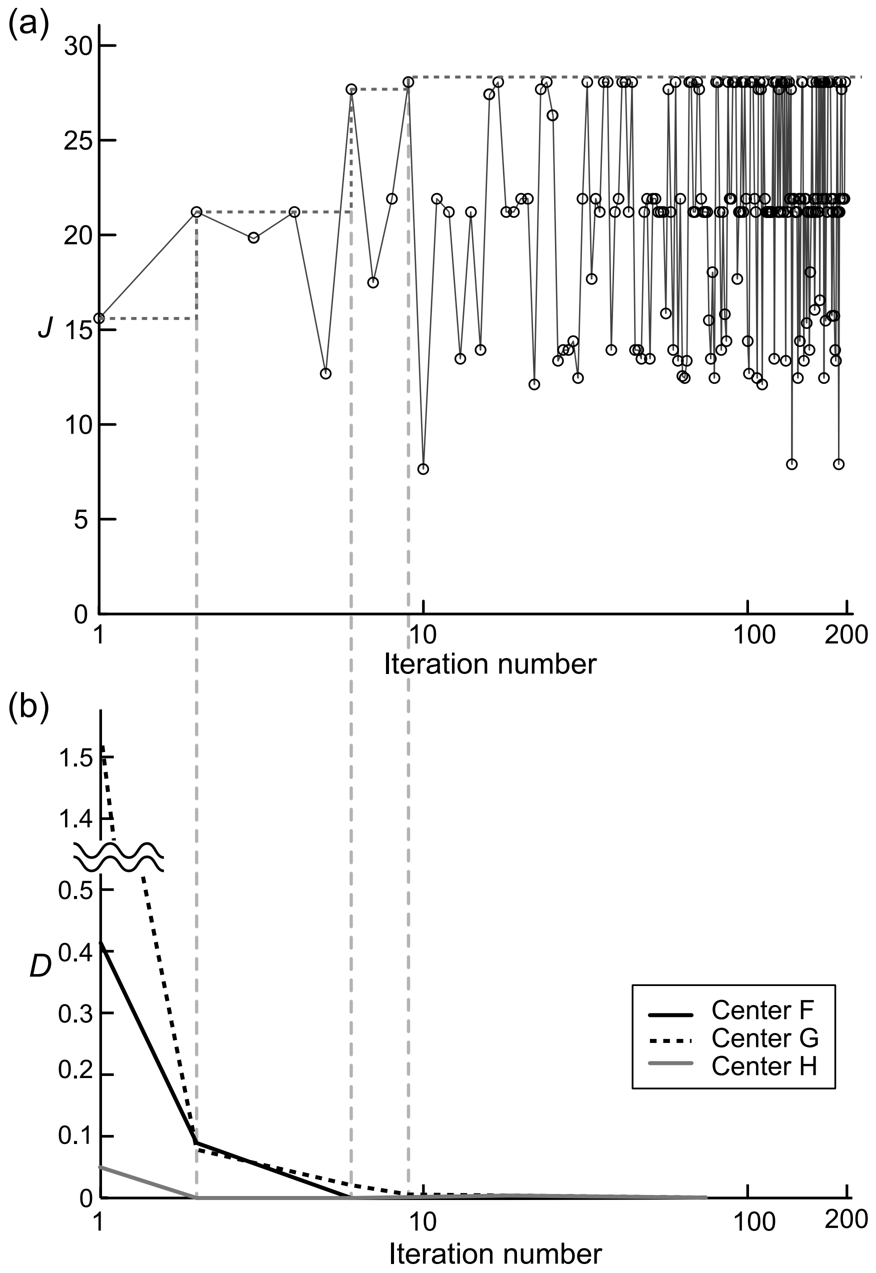

Fig. 2. (a) Variation of discrimination criterion J in 200 runs with different distributions of initial cluster centers for the reduced stress tensors shown in Fig. 3b. Dashed line depicts the improvement of J. In this example, three clusters, F, G, and H, were identified. The principal orientations of centers of the clusters are plotted in Fig. 4a. (b) Convergence in the clustering is shown by the stress difference D between the best distribution in the past trials and the find division for three clusters.

The eigenvalues of this matrix indicate the extent of a cluster. We have k clusters, so that

$$S_{\mathrm{W}} = \sum_{c=1}^{k} S_c = \sum_{c=1}^{k}\sum_{i=1}^{N_c}\left[\vec{x}_c^{(i)} - \vec{\mu}_c\right]\left[\vec{x}_c^{(i)} - \vec{\mu}_c\right]^{\mathrm{T}} \qquad (5)$$

is a representative spread of the entire clusters. On the other hand, the symmetric matrix

$$S_{\mathrm{B}} = \sum_{c=1}^{k} N_c\left[\vec{\mu}_c^{(i)} - \vec{\mu}_{\mathrm{all}}\right]\left[\vec{\mu}_c^{(i)} - \vec{\mu}_{\mathrm{all}}\right]^{\mathrm{T}} \qquad (6)$$

denotes the dispersion of the cluster centers, where $\vec{\mu}_{\mathrm{all}} = \frac{1}{N}\sum_{i=1}^{N}\vec{x}^{(i)}$. The eigenvalues of S_{B} indicate the dispersion of the centers. For this reason, S_{B} is called the between-cluster scatter matrix. In contrast, S_{W} is called the within-cluster scatter matrix (Duda et al., 2001).

We expect that each cluster should be concentrated after the iteration, and that cluster centers are clearly separated from each other. If there are overlapping clusters, they should be merged into a larger cluster. The compactness is expressed by the inverse of the within-cluster scatter matrix S_{W}^{-1}, and the separation of clusters is represented by S_{B}. Therefore, the two

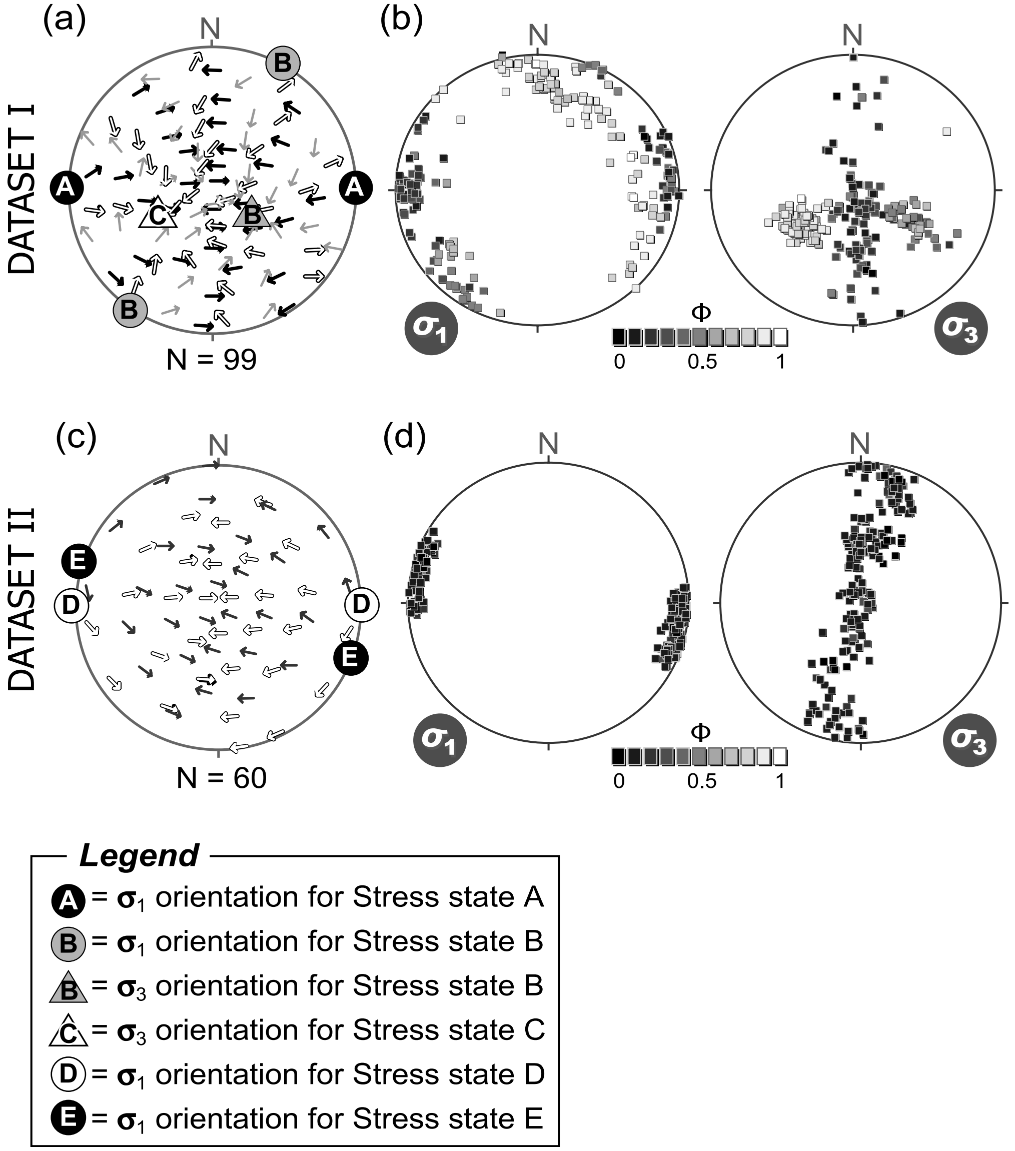
(a)
(b)
DATASET I
N
A
B
C
N = 99
σ1
σ3
Φ
0
0.5
1
(c)
(d)
DATASET II
D
E
N = 60
Legend
A = σ1 orientation for Stress state A
B = σ1 orientation for Stress state B
B = σ3 orientation for Stress state B
C = σ3 orientation for Stress state C
D = σ1 orientation for Stress state D
E = σ1 orientation for Stress state E

Fig. 3. Tangent-lineation diagrams (Twiss and Moores, 1992, p. 206) showing the artificial fault-slip datasets I (a) and II (c). Lower-hemisphere, equal-angle projection. Dataset I recorded three Stress states A, B and C of which principal orientations are also indicated in (a). Solid, gray, and open arrows in (a) indicate faults activated by the stress states A, B, and C, respectively. Open and solid arrows in (c) indicate faults activated by Stress states D and E, respectively. Paired stereograms in (b) and (d) show the results of the multiple inverse method (version 4) (Yamaji, 2000b) applied to Datasets I and II. Left and right stereograms in each subfigure indicate the σ_1- and σ_3-orientations by lower-hemisphere, equal-area projections, in which the clusters of square symbols represent significant stresses. Stress ratio Φ is indicated by a gray scale. The method was applied with the combination number $k_f=4$ for the datasets. The enhance factor is chosen at $e=8$ for the datasets. See Yamaji (2000b) for the details of those parameters.

matrices are combined into the matrix $S_W^{-1}S_B$. This is a 5×5 square matrix, because we deal with points in the five-dimensional parameter space.

To evaluate and compare the results of clustering, a scalar measure instead of the square matrix is required. The measure is expected not to be affected by coordinate rotations in the parameter space. A square matrix has various invariants. Among those, the trace

$$J = \mathrm{trace}\left(S_W^{-1}S_B\right) \tag{7}$$

is the simplest discrimination criterion. A good division has a large J.

So as to search for the best clustering, we performed the clustering procedure 200 times with a constant number of clusters, k with 200 different configurations of initial cluster enters, which were randomly distributed for each trial of the 200 runs (Fig. 2a). If a calculated J exceeds the maximum J of previous trials, the optimal result is updated. As a result, convergence of the optimal result was found in the clustering (Fig. 2b).

3. Test

In order to test the above clustering technique, we analyzed two artificial fault-slip data sets I and II generated with known

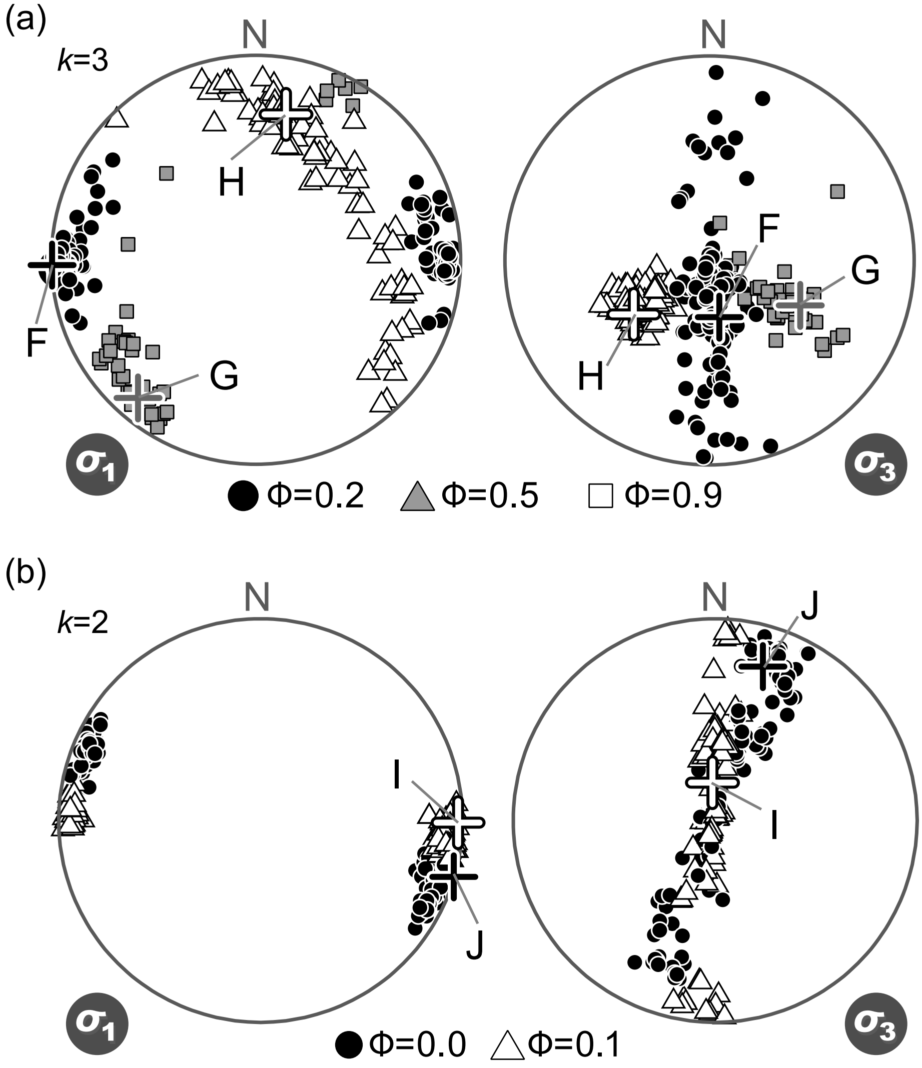

Fig. 4. Result of the present clustering technique of the reduced stress tensors deduced from Datasets I and II. Lower-hemisphere, equal-area projection. The number of clusters, *k*, is 3 and 2, respectively, for Datasets I (a) and II (b). Crosses indicate the cluster centers. Members of the same cluster have the same symbols, i.e., solid circles, gray triangles and open squares. In Dataset I, detected Stress states F, G, and H correspond to assumed Stress states A, B, and C, respectively. In Dataset II, detected Stress states I and J correspond to assumed Stress states D and E, respectively.

stresses. The fault planes were generated with random orientations. Measurement errors and variations in natural stress field were simulated by adding or subtracting random angles values to the slip directions of the faults that were predicted by the Wallace–Bott hypothesis (Wallace, 1951; Bott, 1959). The values had a normal distribution with the mean and the standard deviation at zero and 15°, respectively. The multiple inverse method (Yamaji, 2000b) was applied to these datasets to yield reduced stress tensors.

Dataset I (Fig. 3a), which was used by Yamaji (2003a), had 99 faults. They were composed of three Subsets I_a, I_b and I_c to which different stresses were assigned. The faults in Subset I_a were activated by Stress A, which was an axial compression ($\sigma_3 = \sigma_2 < \sigma_1$), and had the σ_1-orientation of 090°/00°. Subset I_b was activated by Stress B which was a triaxial stress ($\Phi = 0.5$) with the σ_1- and σ_3-orientations of 030°/00° and 120°/50°, respectively. Subset I_c was activated by Stress C, which was an axial tension ($\sigma_3 < \sigma_2 = \sigma_1$) with the σ_3-orientation of 240°/50°. The three stresses were almost equally separated by stress differences of ∼1.4.

The multiple inverse method yielded three distinctive clusters of output stresses from Dataset I (Fig. 3b). Two hundred and three tensors were plotted on the stereograms. Application of the clustering to the stresses with $k = 3$ resulted in the identification of the clusters F, G, and H (Fig. 4a). The center of the cluster F had $\Phi = 0.18$ and the σ_1- and σ_3-orientations of 267°/02° and 172°/66°, respectively. The center of the cluster G had a stress ratio of 0.53 and the σ_1- and σ_3-orientations of 219°/10° and 117°/50°, respectively. The center of the cluster H had a stress ratio of 0.94 and the σ_1- and σ_3-orientations of 011°/28° and 235°/53°, respectively. Clusters F, G, and H had 81, 81, and 41 tensors, respectively. We associated Stresses F, G, and H with the assumed ones A, B, and C. Stress differences between the assumed and recognized stresses were 0.20, 0.24, and 0.15, respectively. Compared with the theoretical maximum value of stress difference at 2

(Orife and Lisle, 2003), these misfits are sufficiently small. We concluded that the present technique was successful in detecting the stresses.

Once stress tensors are classified as cluster centers, we can estimate the spread of each cluster of s_c. The recognized clusters F, G, and H had values of s_c at 0.43, 0.46, and 0.47, respectively. Cluster F had the minimum value, indicating that the stress represented by the center of Cluster F is the most reliable.

Next, we tested the resolution of our clustering technique. Dataset II consisted of two Subsets $\mathrm{II_d}$ and $\mathrm{II_e}$ to which different stresses were assigned. They are E–W- and WNW–ESE-trending axial compressions, respectively (Fig. 3c). The angle between σ_1-axes was 20° and the stress difference was 0.6. Each subset had 30 faults.

Fig. 3d shows the result of the multiple inverse method applied to Dataset II. One hundred and eighty-eight tensors were plotted on the stereograms. Two clusters should appear on each of the stereograms in the subfigure corresponding to the assumed stresses. However, it is difficult to separate the two clusters by visual inspection of the stereograms. Applying the present clustering method to the stresses that are plotted on the stereograms, we recognized Clusters I and J (Fig. 4b). The center of Cluster I had $\Phi = 0.05$ and the σ_1- and σ_3-orientations of 092°/02° and 355°/74°, respectively. The s_c was 0.58. The center of Cluster J had $\Phi = 0.03$ and the σ_1- and σ_3-orientations of 108°/01° and 17°/22°, respectively. The s_c was 0.68. Clusters I and J had 105 and 83 tensors, respectively. We associated Stresses I and J with given ones D and E. The stress differences between the detected and assumed ones were 0.09 and 0.08, respectively. These misfits are small as in the case of Dataset I.

Once stresses are recognized as the cluster centers, the fault-slip data can be sorted into homogeneous subsets according to the compatibility of the data to the stresses. A fault-slip datum is said to be compatible with a stress if the angular misfit of the theoretical slip direction, which is calculated with the Wallace–Bott hypothesis (Wallace, 1951; Bott, 1959), from the observed slip-direction is smaller than a threshold value, e.g. 10–20° (Angelier, 1979; Etchecopar et al., 1981; Nemcok and Lisle, 1995; Liesa and Lisle, 2004).

4. Outstanding problem

Deciding on the number of clusters is a difficult problem in clustering (Duda et al., 2001). The *k*-means clustering requires the number of clusters *k* known a priori (MacQueen, 1967). Although many arguments have been made in information science about this problem, it remains unsolved. The same problem exists also in clustering on a hypersphere that this study uses. Banerjee and Ghosh (2004) proposed the clustering technique in a multi-dimensional hypersphere and called this clustering 'spherical *k*-means'. They specified the number of clusters beforehand, and performed the clustering. It is the future aim of our research to develop a method of determining the optimal number of division, *k*, automatically.

5. Summary

The *k*-means clustering was used to objectively recognize the clusters of stress tensors that the multiple inverse method yielded. Applied to artificial heterogeneous data, the technique was capable of the assumed stresses successfully. Automatic determination of the number of clusters will require development.

Acknowledgements

The authors are very grateful to R.J. Lisle and T. Orife for their careful and constructive reviews. This work was supported by Kyoto University Active Geosphere Investigations for the 21st Century Centers of Excellence Program (KAGI21), which was approved by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, and by the Japan Society for the promotion of Science (No. 14540423).

References

Angelier, J., 1979. Determination of the mean principal directions of stresses for a given fault population. Tectonophysics 56, T17–T26.

Angelier, J., 1994. Fault slip analysis and paleostress construction. In: Hancock, P. (Ed.), Continental Deformation. Pergamon Press, London, pp. 53–100.

Banerjee, A., Ghosh, J., 2004. Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres. IEEE Transactions on Neural Networks 15, 702–719.

Bott, M.H.P., 1959. The mechanics of oblique slip faulting. Geological Magazine 96, 109–117.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, 2nd ed. John Wiley and Sons, New York.

Etchecopar, A., Vasseur, G., Daignieres, M., 1981. An inverse problem in microtectonics for the determination of stress tensors from fault striation analysis. Journal of Structural Geology 3, 51–65.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.

Fry, N., 1999. Striated faults: visual appreciation of their constraint on possible paleostress tensors. Journal of Structural Geology 21, 7–21.

Liesa, C.L., Lisle, R.J., 2004. Reliability of methods to separate stress tensors from heterogeneous fault-slip data. Journal of Structural Geology 26, 559–572.

Lloyd, S.P., 1982. Least-squares quantization in Pcm. IEEE Transactions on Information Theory 28, 129–137.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability 1, 281–296.

Nemcok, M., Lisle, R.J., 1995. A stress inversion procedure for polyphase fault/slip data sets. Journal of Structural Geology 17, 1445–1453.

Orife, T., Lisle, R.J., 2003. Numerical processing of palaeostress results. Journal of Structural Geology 25, 949–957.

Sato, K., Yamaji, A., 2006. Embedding stress difference in parameter space for stress tensor inversion. Journal of Structural Geology, this issue, doi.:10.1016/i.jsg.2006.03.004

Shan, Y.H., Suen, H.B., Lin, G., 2003. Separation of polyphase fault/slip data: an objective-function algorithm based on hard division. Journal of Structural Geology 25, 829–840.

Shan, Y.H., Li, Z., Lin, G., 2004. A stress inversion procedure for automatic recognition of polyphase fault/slip data sets. Journal of Structural Geology 26, 919–925.

Twiss, R., Moores, E., 1992. Structural Geology. Freeman, New York.

Wallace, R.E., 1951. Geometry of shearing stress and relation to faulting. Journal of Geology 59, 118–130.

Yamaji, A., 2000a. The multiple inverse method applied to meso-scale faults in mid-Quaternary fore-arc sediments near the triple trench junction off central Japan. Journal of Structural Geology 22, 429–440.

Yamaji, A., 2000b. The multiple inverse method: a new technique to separate stresses from heterogeneous fault-slip data. Journal of Structural Geology 22, 441–452.

Yamaji, A., 2003a. Are the solutions of stress inversion correct? Visualization of their reliability and the separation of stresses from heterogeneous fault-slip data. Journal of Structural Geology 25, 241–252.

Yamaji, A., 2003b. Slab rollback suggested by latest Miocene to Pliocene forearc stress and migration of volcanic front in southern Kyushu, northern Ryukyu Arc. Tectonophysics 364, 9–24.

Yamaji, A., Sakai, T., Arai, K., Okamura, Y., 2003. Unstable forearc stress in the eastern Nankai subduction zone for the last 2 million years. Tectonophysics 369, 103–120.

Yamaji, A., Tomita, S., Otsubo, M., 2005. Bedding tilt test for palaeostress analysis. Journal of Structural Geology 27, 161–170.

Yamaji, A., Otsubo, M., Sato, K., 2006. Paleostress analysis using the Hough transform for separating stresses from heterogeneous fault-slip data. Journal of Structural Geology, this issue, doi.:10.1016/i.jsg.2006.03.016