April 9, 2016

GArcmB Software Package User's Guide

Atsushi YAMAJI

Division of Earth and Planetary Sciences,
Kyoto University, Japan

Software for fiting mixed Bingham distribution to 3D orientation data

Copyright

Disclaimer of warranty

This program package is free.

Free of charge software is provided on an “AS IS” basis, without warranty of any kind, including limitation the warranties of merchantability, fitness for a particular purpose and non-infringement. The entire risk as to the quality and performance of the software is borne by you. Should the software prove defective, you assume the entire cost of any service and repair.

Acknowledgments

I am grateful to our coworkers and the researchers who tested and/or used the present method. I thank K. Sato for the assistance to compile the software, and for discussions. This program was developed under the financial support from JSPS (15H02141).

1 Introduction

GArcmB is the software for the fuzzy clustering of 3D orientations and for the paleostress analysis of dilational fractures such as dikes and mineral veins. From the poles of those geological planar structures, the paleo stresses and fluid pressures at the times of fracture dilation are estimated by the present method [7, 8].

Fuzzy clustering does not divide data into distinct clusters. Intead, a datum can belong to more than one cluster; and the probability of a datum to belong to a cluster is evaluated by the software [2]. The probability is called the membership of the datum to the cluster.

The software searches for the mixed Bingham distribution that best fits a set of orientation data through the real-coded genetic algorithm. A mixed Bingham distribution is the linear combination of Bingham distributions (Fig. 1). The coefficients of the combination, $\varpi^1, \ldots, \varpi^K$, are called mixing coefficients[1], where K denotes the number of clusters (Table 1). A Bingham distribution is so a flexible statistical distribution that it can describe circular, elliptical and girdle clusters (Fig. 2). A Bingham distribution has the minimum, intermediate and maximum concentration axes that meet at right angles with each other[2]. The concentration parameters, κ_1 and κ_2, represent the shape and spread of a cluster. That is, κ_1 denotes the spread of a cluster along the great circle defined by the minimum and maximum concentration axes, whereas κ_2 denotes that along the great circle defined by the maximum and intermediate concentration axes (Fig. 2). Note that the parameters are defined to be negative in sign, and satisfy $\kappa_1 \leq \kappa_2 \leq 0$.

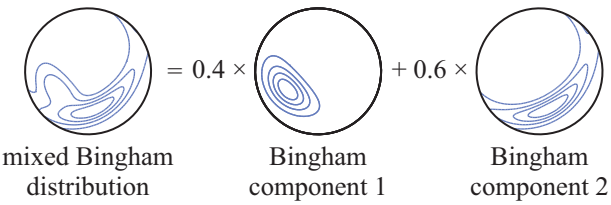

Figure 1: Equal-area projections illustrating the probability densities of a mixed Bingham distribution and its Bingham components with the mixing coefficients, 0.4 and 0.6.

[1]The textbook by Bishop explains the basics of statistical concepts and techniques including statistical mixture models, log-likelihood function, and BIC, Bayesian information criterion [2].

[2]Bingham distribution was named after C. Bingham [1], and is concisely introduced by the book and article by Borradaile and Love [3, 5]. For the details of directional statistics consult the textbooks [4, 6].

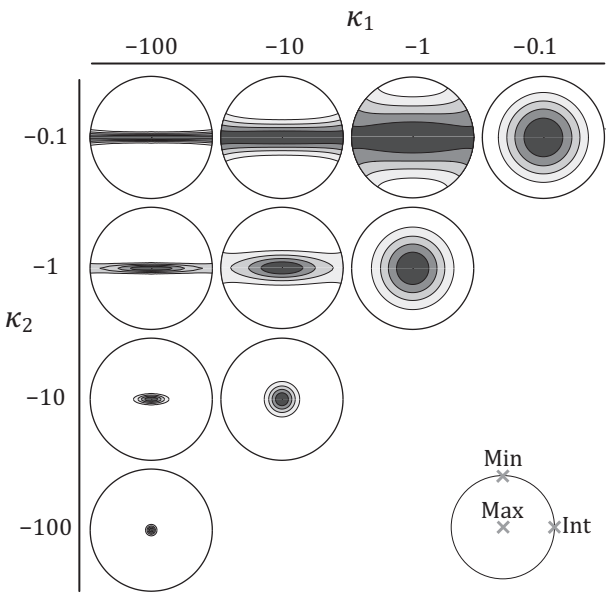

Figure 2: Equal-area projections showing the probability distributions of Bingham distributions with various concentration parameters, κ_1 and κ_2. A Bingham distribution has the maximum, intermediate and minimum concentration axes that meet each other at right angles. When a Bingham distribution is fitted to the poles to dilational fractures, the axes indicate the σ_3-, σ_2- and σ_1-axes, respectively. And, the stress ratio, $\Phi = (\sigma_2 - \sigma_3)/(\sigma_1 - \sigma_3)$ is equal to κ_2/κ_1.

Table 1: List of symbols.

BIC	Bayesian information criterion
g	generation (iteration)
K	number of clusters
L	log-likelihood function indicating the goodness of fit of a mixed Bingham distribution
N	number of data
κ_1, κ_2	concentration parameters
ϖ^k	the mixing coefficient of the kth Bingham component
σ_1, σ_2, σ_3	principal stresses ($\sigma_1 \geq \sigma_2 \geq \sigma_3$)
Φ	stress ratio

The orientations are fuzzily partitioned into K clusters by the present software, where K is specified by the user before computation. The software calculates the Bayesian information criterion (BIC) of the resultant partition. The appropriate K value can be determined for a given data set using the BIC values for the K values, $1, 2, \ldots, 5$.

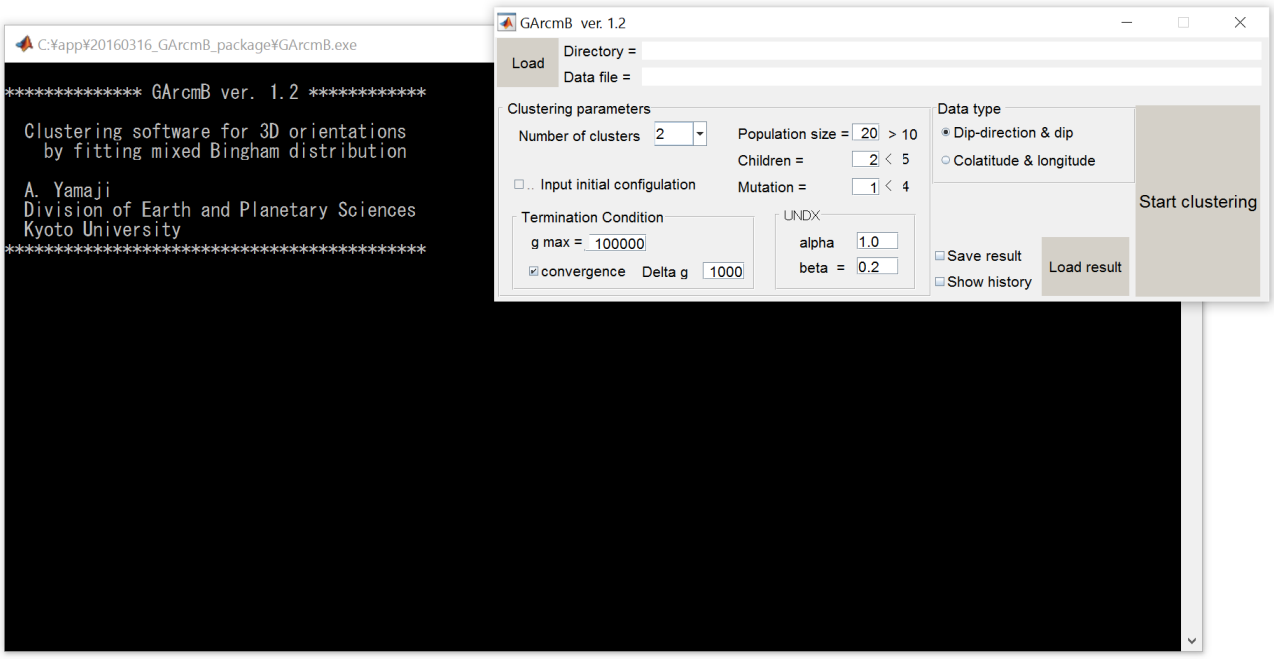

Figure 3: The console window of MATLAB (lower left) and the control panel of the present software (upper right).

It is happy for the author if the users of this program cite one or both of the papers [7, 8] in their articles.

2 Installation

The present software was developed in MATLAB R2015b, and was complied to make executable files for the users who do not have the licence of MATLAB. It is expected that the users have built the environment prior to the use of the present software. The MATLAB Runtime available from MathWorks builds the environment in your computer.

If you use Windows, file name extensions, e.g., '.exe' and '.zip,' should be visible. Consult the Help of your operating system to make them visible in Windows Explorer.

Instructions for installation

1) Operating system Make sure of the operating system of your computer. The present program package is compatible only with 64-bit operating systems (Windows, Mac and Linux). The package for 32-bit ones is not provided.

2) MATLAB Runtime Visit the home page of MathWorks, `http://www.mathworks.com/products/compiler/mcr/index.html?s_tid=gn_loc_drop` to download a version of the MATLAB Runtime appropriate to your computer and to MATLAB R2015b. MathWorks has the web pages for providing the Runtimes not only in English, but also in several languages.

3) Program package From the homepage of the present software, `http://www.kueps.kyoto-u.ac.jp/~web-bs/tsg/software/GArcmB/`, download the zipped file, `GArcmB.zip`; and extract the file in the directory where the program package is stored in your computer.

4) Test Launch the executable file, `GArcmB.exe`, to check weather the installation was successful. If the Runtime and the package are installed appropriately, the console window and the control panel in Fig. 3 should appear in the computer screen. It takes a few tens of seconds for the panel to appear. If they do not appear, check the versions of the Runtime.

3 Formats of data files

`GArcmB` reads a text file containing orientation data with either of the two data types, geological and spherical coordinates. The coordinate system used in the software is shown in Fig. 4a.

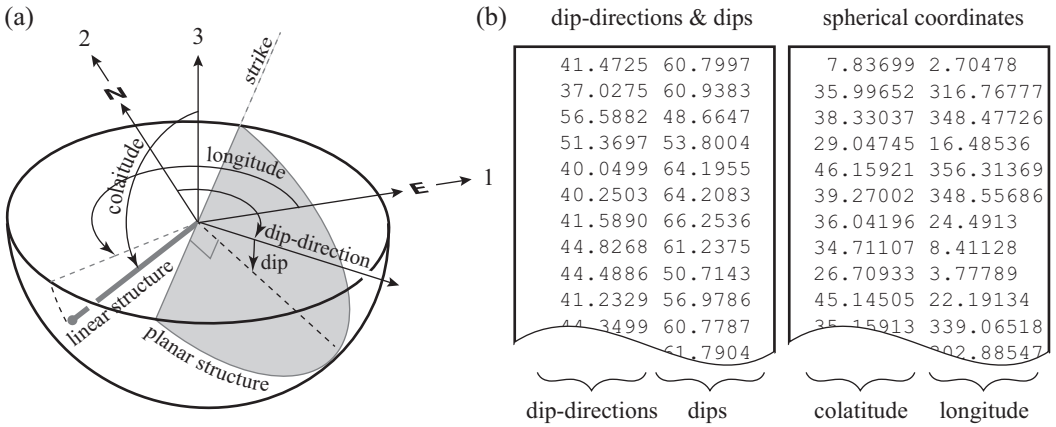

Figure 4: (a) The rectangular Cartesian coordinates used in the software. The 1- and 2-axes are oriented north- and eastward, respectively. The dip-direction of a planar structure, e.g., a fracture, is measured from the north. The orientation of a linear structure is described by spherical coordinates, longitude of which is measured from the east. Equal-angle projections of the lower-hemisphere are used. (b) Two file formats acceptable for the program. Not only real but also integer values are acceptable. Each row of the list correspond an orientation datum, and the data in a row must be separated by a tab, camma or space(s).

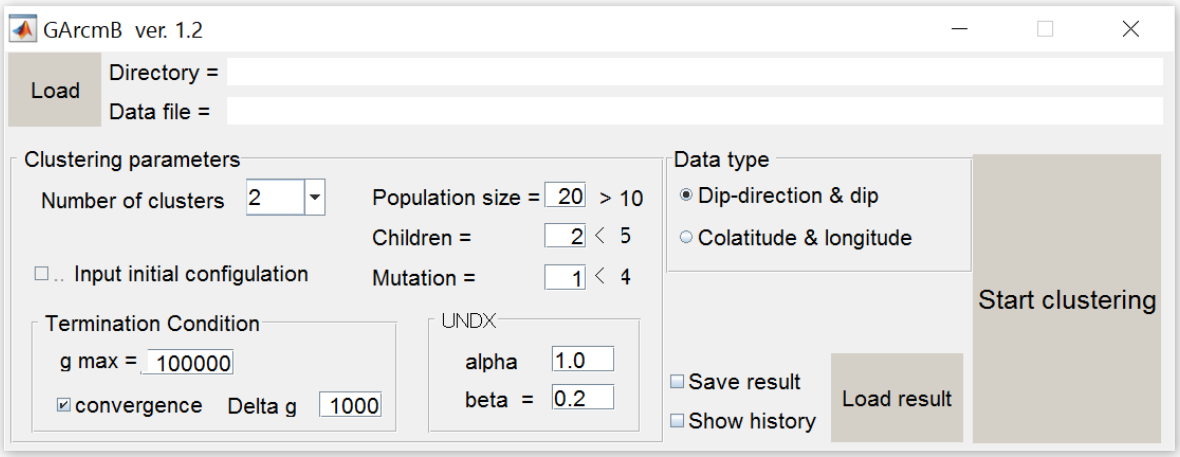

Figure 5: The main window (control panel) of the present software. Other windows pops up by pressing the button 'Start clustering.'

Geological data For the clustering of fracture orientations, the text file should be the list of the dip-directions and dips of the fractures. A raw of the list indicates a datum (Fig. 4b). The direction is indicated by the azimuth in degrees (0–360°). the direction and dip are separated by a space, tab or comma. Observe the sample file, 'fractures.txt.'

Spherical coordinates The software can read the colatitudes and longitudes as well. In this case, co-latitude and latitude in degrees are aligned from left to right in a raw of the text file (Fig. 4b). They should be separated by a space, tab or comma. See the sample file, 'spherical.txt.'

4 How to use the software

4.1 Basic operation

It is easy to use `GArcmB` with its control panel (Fig. 5).

1. Launce the executable file, `GArcmB.exe`. Then, the console window and the control panel of the software pop up on your computer screen (Fig. 3).

2. Press the button Load at the upper left of the

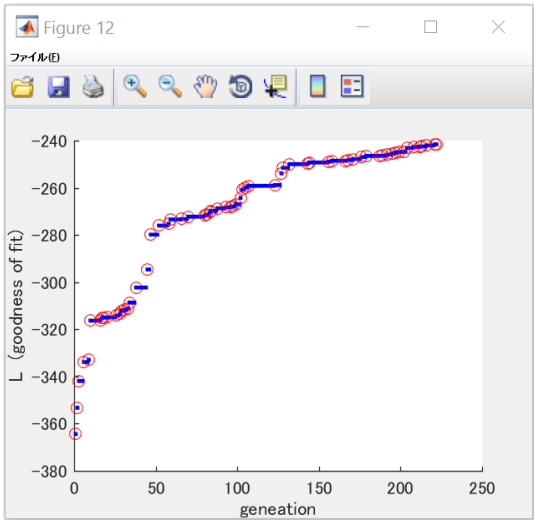

Figure 6: L versus generation (g), showing the gradual and intermittent improvement L. This window appears if the checkbox 'Show history' is clicked before the start of clustering.

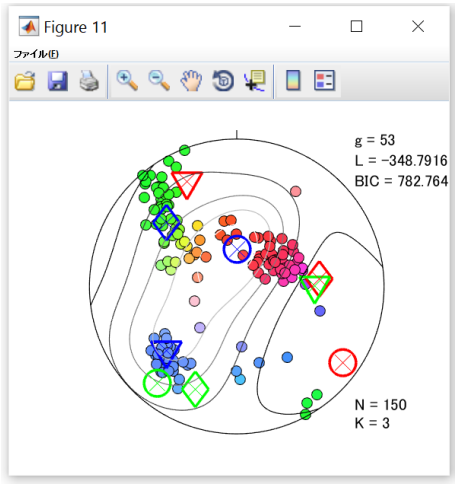

Figure 7: The lower-hemsphere, equal-area projection in which the gradual improvement of a mixed Bingham distribution and the memberships of data are displayed.

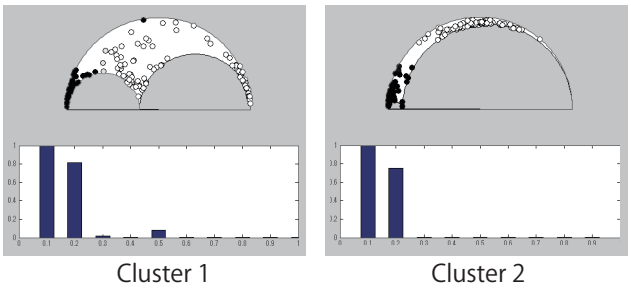

Figure 8: Mohr diagrams for the clusters and the bar graphs for the memberships of summed in the bins with the width of 0.1. See the article [7] in detail.

panel to load a data file.

3. Choose the data type, 'Dip-direction & dip' or 'Colatitude & longitude' by clicking a radio button in the panel, 'Data type.'

4. Select the number of clusters, K, from the pull-down menu on the control panel.

5. To save the result of computation when the clustering is terminated, click the checkbox 'Save result.' If you want to observe the clustering process, click the checkbox 'Show history' to plot the log-likelihood of the best individual versus generation and the goodness of fit, L, versus generation (Fig. 6). However, the time of plotting increases with g.

6. Press the big button Start clustering to begin calculation. If the number of data, N, is smaller than $6K$, the following error message is printed in the console window of MATLAB and the computation is terminated.

```
Error: insufficient number of data.
       Execution terminated.
```

Otherwise, the clustering starts. The progress can be monitord with the plot in Fig. 7.

7. If the computation comes to an end successfully, the program shows the message,

```
======== Execution completed ========
```

in the console window. And, the software finishes the clustering, and draw Mohr diagrams for fractures (Fig. 8). The window in Fig. 9 appears as well. If you do not deal with geological data, ignore the diagrams. If the checkbox, 'Save result,' was clicked before the calculation, all values in the memories used in the computation are stored in a file, e.g., `fractures_k3-161.1857.mat`, where 'fractures' is the name of the input file and `_k3` denotes that the orientation data in the file were partitioned into three clusters, and `-161.1857` is the final L vlue. The output file is automatically named, and saved in the directory containing the input file. The saved result can be reloaded by pushing the button Load result.

8. Record the final L value, which is indicated in the console window (Fig. 10), in a green cell of

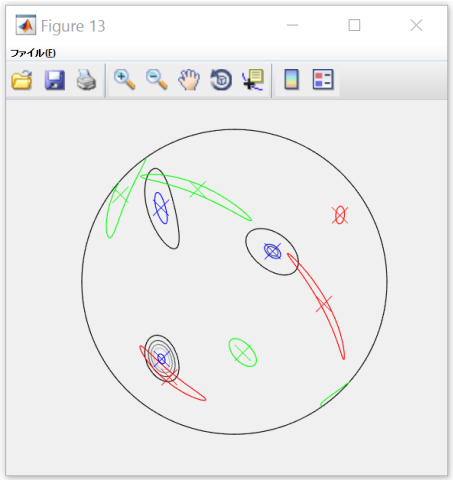

Figure 9: The window that pops up just after the termination of clustering. The concentration axes and their 95% confidence regions are shown by lower-hemisphere, equal-area projection. Contour lines indicate the probability density of the mixed Bingham distribution resulted from the clustering.

```
========== generation 59 ==========
  fractures.txt
  L = -325.744201   BIC = 736.669201
  Cluster 1
    kappa = (-10.571, -5.017)  Phi = 0.4746  varpi = 0.3481
    min:  75.5/26.8,  int: 195.9/45.1,  max: 326.5/32.9
  Cluster 2
    kappa = (-7.659, -2.649)  Phi = 0.3459  varpi = 0.3287
    min: 304.5/18.2,  int:  57.2/49.6,  max: 201.4/34.6
  Cluster 3
    kappa = (-6.709, -2.086)  Phi = 0.3110  varpi = 0.3232
```

Figure 10: The progress of the clustering is shown in the consoleg window of MATLAB. This example shows the status at the 70th generation when the file, fractures.txt was processed.

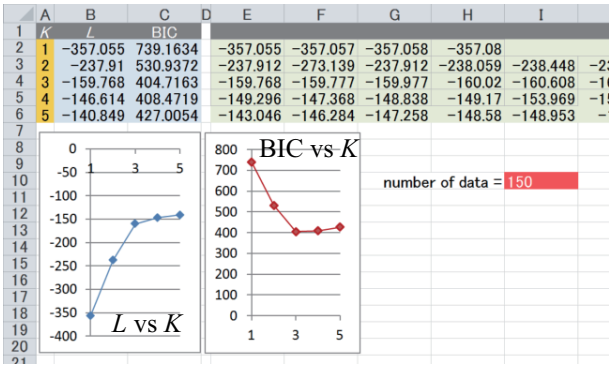

Figure 11: Example of the spread sheet that the results of computation for the same data set is summarized.

an Excel sheet (Fig. 11). If the value is greater than the L values that have been obtained from the same data set with the same K value, then, (1) copy the final L value and paste it in the the column B, and (2) record the final BIC value, which appeared in the console window, in the column C.

Run the program several times for each K value to search for the global maximum of the log-likelihood, because the results of the computation depends on the configuration at the 0th generation, which is randomly generated.

4.2 Termination condition

The iteration of the genetic algorithm is terminated when one of following conditions is met. The program quits the iteration at the 100,000th generation, the number of which is indicated as $g_{\max}$ in the control panel in Fig. 5. Another value can be set for the maximum generation before starting the clustering. On the other hand, the iteration can be stopped before the $g_{\max}$th generation when Δg generations have passed since the log-likelihood of the best individual was last updated. If you want to use this condition, The checkbox 'convergence' in the box 'Termination condition' must be checked. The value of Δg can be changed in the box.

If you want to break off the computation, close the console window.

4.3 Parameters for the genetic algorithm

You can tune the parameters of the genetic algorithm by changing the values in the box 'Clustering parameters' on the control panel in Fig. 5. 'Population size' denotes the number of individuals involved in the genetic algorithm, and 'Mutation' is the number of individuals generated in an iteration. The box 'UNDX' has the α and β values, which are used in the crossover routine of the algorithm. Two children are born in an iteration. Consult the article [7] for details of the parameters.

4.4 Progress of clustering

The progress of clustering can be seen in the console window (Fig. 10) and the equal-area projection in another window (Fig. 7). Temporal values of L and BIC are displayed in the windows. In the example in Fig. 10, three clusters were fitted to the

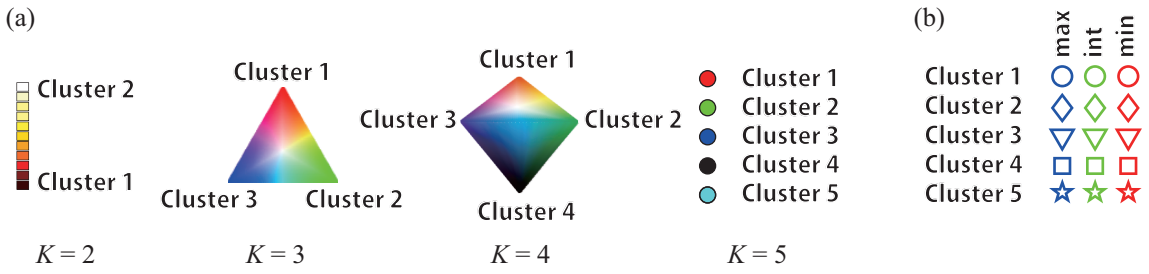

Figure 12: Colors and symbols used in equal-ara projection (Fig. 7). (a) Color schemes for indicating the memberships of data points. (b) Symbols for indicating concentration axes.

sample data in 'fractures.txt.' `kappa` denotes the concentration parameters, κ_1 and κ_2, of the Bingham distribution fitted to a cluster. `Phi` is the stress ratio, $\Phi = (\sigma_2 - \sigma_3)/(\sigma_1 - \sigma_3) = \kappa_1/\kappa_2$ [9], and varpi is the mixing coefficient, ϖ.

The memberships of data points and the concentration axes of clusters are indicated by colors and symbols, respectively, in the equal-area projection (Fig. 7). Figure 12 shows the legend for the colors and symbols. Differences in the memberships are depicted by color gradations for the cases of $K = 2$, 3 and 4. The gradation for $K = 2$ is defined as the colormap, 'hot,' in the basic set of MATLAB. The color schemes for $K = 3$ and 4 are depicted by the ternary and quaternary diagram in Fig. 12a, the EPS file of which are included in this program package. In case of $K = 5$, there is no way to depict the intermediate values of the memberships. Accordingly, the data points are drawn with the colors specific to the clusters. That is, if a data point has the memberships belonging to Clusters 1 through 5 are 0.1, 0.1, 0.2, 0.2 and 0.4, respectively, the data point is drawn with the color indicating Cluster 5 because the data point has the maximum membership to Cluster 5.

The plots produced by the present software can be made up with drawing software, e.g., Illustrator and Canvas, for publication (Fig. 13). You can use the color maps contained in the files, 'colors.eps,' for this purpose.

4.5 Animation

The animation files available from the homepage of this software were created by means of the executable files, 'GArcmB.exe' and 'animator.exe.' The clustering process displayed in the equal-area projection in Fig. 7 is stored in the intermediate file, 'animation.mat,' when the clustering is normally terminated. The executable file, 'animator.exe,' makes a movie file from the intermediate file. The movie has the name, 'GArcmB.avi.' The intermediate and movie files are saved in the directory where the present program package is stored. The files, 'animation.mat' and 'GArcmB.avi,' are overwritten, if they exist in the directory.

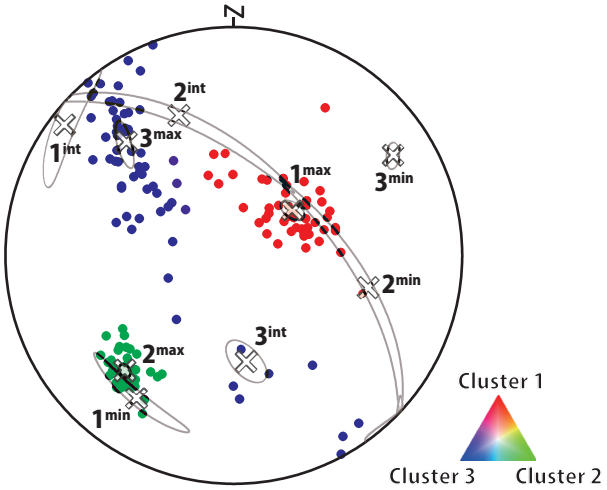

Figure 13: Lower-hemisphere, equal-area projections showing the optimal partition of the sample data 'fractures.txt.' The maximum, intermediate and minimum concentration axes of the kth cluster is denoted by crosses. The labels attached to the crosses distinguish the clusters and their concentration axes such that $1^{\max}$ denotes the maximum concentration axis of Cluster 1. The 95% confidence ellipse of each axes is shown as well. The membership of each data point is indicated by a color in the ternary plot.

5 Tips to find the best partition

A data set can be partitioned into various groups, because $L(\boldsymbol{X})$ is a multimodal function of partition, $\boldsymbol{X}$. The function denotes the goodness of fit of a mixed Bingham distribution that is represented by $\boldsymbol{X}$. The

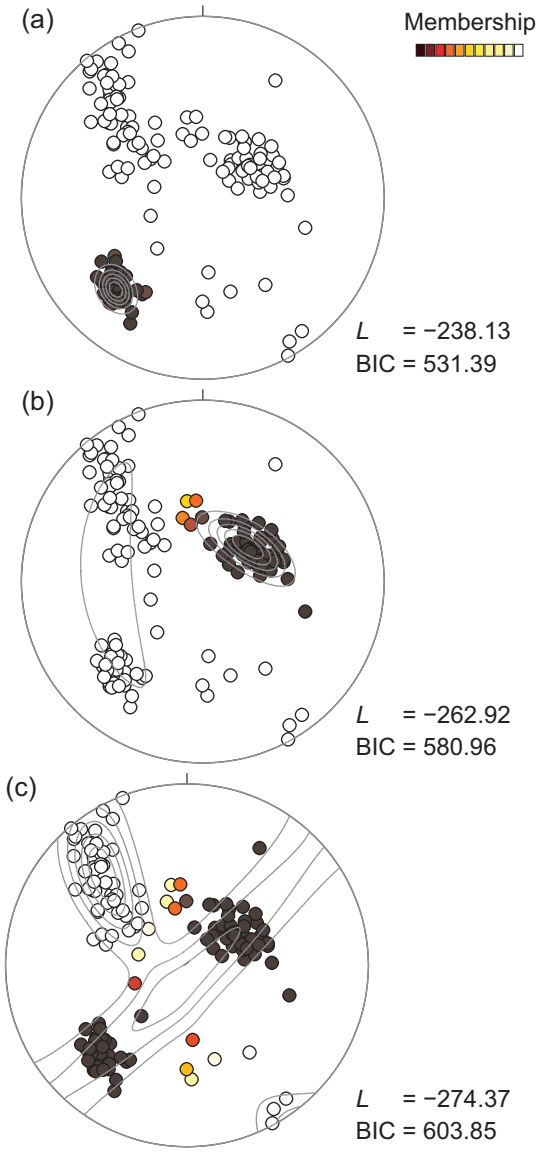

Figure 14: Equal-area projections showing the three examples of the partitions of the sample data into three clusters. Contour lines denote the iso-density lines of the mixed Bingham distribution that was fitted to the data set. (a) The optimal partition with the maximum L and minimum BIC values. The dense cluster in the SW quadrant was separated from remaining data points. (b, c) Partitions corresponding to the local maxima of $L(\boldsymbol{X})$. The dense cluster in the NE quadrant was separated from remaining ones in (b), whereas the dense clusters in the NE and SW quadrants combined into a group.

best partition has the BIC value smaller than any other partitions. And, the function has the maximum L value for a prescribed K value. Suppose the function has the global maximum, $L(\boldsymbol{X}^{\max})$. Then, the vector, $\boldsymbol{X}^{\max}$, denotes the mixed Bingham distribution that best fits the given data set.

It is not straightforward to find the best one, because $\boldsymbol{X}$ is represented by a vector in a high dimensional space. That is, $\boldsymbol{X}$ is a $5K$-dimensional vector. For example, data are partitioned into three clusters, the global maximum should be searched for in a 15-dimensional space.

It is a difficult point in the fitting of a mixed Bingham distribution comes from the fact that the various combinations of Bingham distributions can be fitted to a given data set. Among them, the user have to determine the mixed Bingham distribution that has the maximum value of the logarithmic likelihood function, $L(\boldsymbol{X})$, that evaluates the goodness of fit of a mixed Bingham distribution to a given data set [7]. However, this function has numerous maximum points: The point where the function has the global maximum should be found by the computation. Figure 14 shows examples. Namely, the figure shows three groupings of the sample data set with different L values: The data was partitioned into two groups 11 times. Figure 14a shows the best partition with the highest L value among the 11 trials. The dense cluster in the SW quadrant was separated from others in the best partition. Figs. 14b and c show the partitions corresponding to two local maxima of the function. It is clear that few data points had intermediate memberships for the case of the best partition (Fig. 14a), whereas there were data points with intermediate memberships in Figs. 14b and c.

The multi-modality of the function, $L(\boldsymbol{X})$, makes the search problem difficult. Although the present

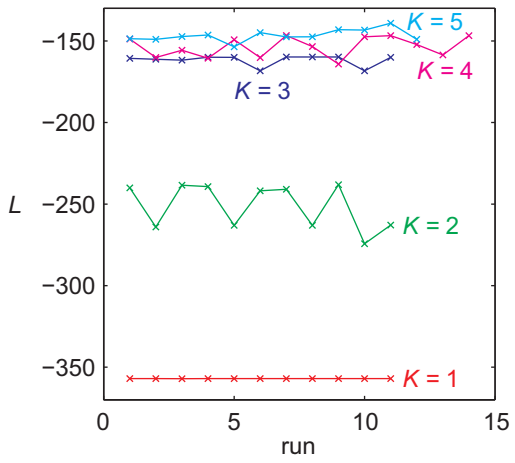

Figure 15: The final L values after clustering for the data set 'fractures.txt.' In case of $K = 2$, L reached the values at −238 and −262 repeatedly, which corresponds to the maxima of the function, $L(\boldsymbol{X})$, shown in Figs. 14a and b. The minimum L at about −275 in the case of $K = 2$ corresponds to the configuration shown in Fig. 14c.

method effectively to do so, *the user have to launch the program with the same number of clusters for several times* to find the best partition of a given data set. Consider that the data are partitioned into K groups. The data set was processed with the program ten times for each of the cases of $K = 1$ through 5. Figure 15 indicates that the graph of L versus K converged until the 5th or 6th computation for the case of the sample data. *The user should make sure of such convergence of the graph*. Fig. 16 shows L and BIC versus K for the data set, 'fracture.txt.'

It depends on data how many times the program have to be launched for the same data set. The genetic algorithm employed in the present method [7] so robust that the program can detect the best partition for the case of $K = 3$, though there are local maxima even in this case. The fact is that they are artificial data, which were made by combining three data sets, each of which was generated from a Bingham distribution. The fluctuations in the L value of this case were smaller than those of other cases.

Appendix A: Version history

What's new in version 1.2 April 9, 2016

- New executable file for animation.
- Improves the seed of random number generator.

Version 1.1 April 1, 2016
The software opened for the public for the first time.

References

[1] Bingham, C., 1974. An antipodally symmetric distribution on the sphere. *Annals of Statistics*, **2**, 1201–1225.

[2] Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

[3] Borradaile, G.J., 2003. *Statistics of Earth Science Data: Their Distribution in Time, Space and Orientation*. Springer, Berlin.

[4] Fisher, N.I., Lewis, T. and Embleton, B.J.J., 1993. *Statistical Analysis of Spherical Data*. Cambridge University Press, Cambridge.

[5] Love, J.J., 2007. Bingham statistics. In: Gubbins, D., Herrero-Bervira, E. (Eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*. Springer, Dordrecht, pp. 45–47.

[6] Mardia, K.V. and Jupp, P.E., 1999. *Directional Statistics*. Wiley, Chichester.

[7] Yamaji, A., 2016. Genetic algorithm for fitting a mixed Bingham distribution to 3D orientations: a tool for the statistical and paleostress analyses of fracture orientations. *Island Arc*, **25**, 72–83.

[8] Yamaji, A. and Sato, K., 2011. Clustering of fracture orientations using a mixed Bingham distribution and its application to paleostress analysis from dike or vein orientations. *Journal of Structural Geology*, **33**, 1148-1157.

[9] Yamaji, A., Sato, K. and Tonai, S., 2010. Stochastic modeling for the stress inversion of vein orientations: Paleostress analysis of Pliocene epithermal veins in southwestern Kyushu, Japan. *Journal of Structural Geology*, **32**, 1137–1146.

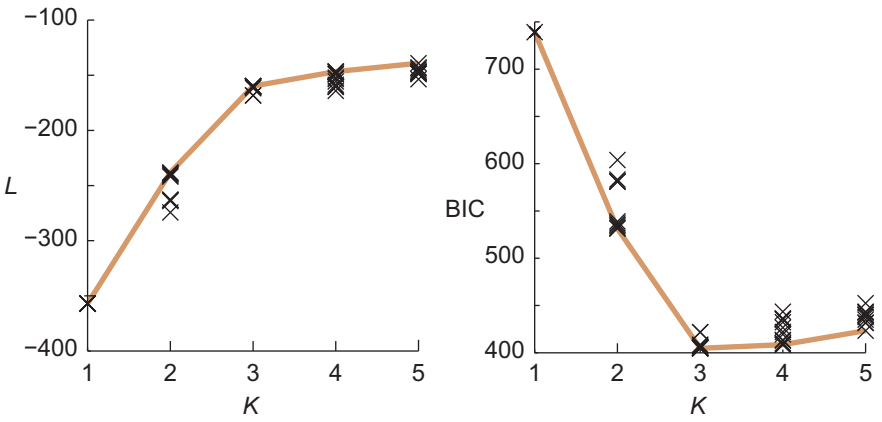

Figure 16: L and BIC values for the data set 'fracture.txt.'