March 25, 2025

GArcmB
Software Package
User's Guide

Atsushi YAMAJI
Division of Earth and Planetary Sciences,
Kyoto University, Japan

MATLAB codes for fiting mixed Bingham distribution to 3D orientation data

Copyright

Disclaimer of warranty

This software 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 utilized the present method. This program was developed under the financial support from JSPS (15H02141).

Contents

1 Introduction

The software, GArcmB, does the fuzzy clustering of 3D orientations by fitting mixed Bingham distribution [9] through real-coded genetic algorithm [8]. The orientations are partitioned into *K* clusters, where *K* is specified by the user before computation. In addition, 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 of different *K* values.

A Bingham distribution is so flexible that it can describe circular, elliptical and girdle distributions (Fig. 1). The distribution has the maximum, intermediate and minimum concentration axes, which meet at right angles with each other. The paired concentration parameters, κ_1 and κ_2, indicate the shape and size of a cluster. They are defined to be negative in sign and to satisfy $0 \geq \kappa_2 \geq \kappa_1$. The spread of the cluster from the maximum to intermediate concentration axis, is denoted by κ_1 whereas that from the maximum to minimum concentration axis is denoted by κ_2. The compactness of a cluster is denoted by $|\kappa_1|$ and $|\kappa_2|$. A mixed Bingham distribution is the superposition of Bingham distributions.

The present software was developed not only for the clustering but also for the geological paleostress analysis of dilational fractures, e.g., dikes and mineral veins [9]. That is, the poles of those fractures are partitioned into *K* groups, from which stress conditions and maximum fluid pressures are determined.

One of the most time consuming routines is the evaluation of the normalizing factors of Bingham distributions. Since the factors should be evaluated millions of times in a run, the present software utilizes the approximate values of the factors. Accordingly, note that the result of this software is not very accurate.

The basics of the present software is described by an article of the author [8]. The software was developed in Matlab versions 7 and 8. The compatibility with other versions of Matlab is unconfirmed except for R2024b. The program sometimes shows error messages at the end of computation in Matlab version R2024b, but no serious effect has been found.

The Optimization Toolbox of Matlab includes a genetic algorithm solver, but the present software does not require the toolbox, but uses only the basic set of Matlab. If you are not familiar with Matlab, consult the textbooks and the on-line help of Matlab.

The present method is detailed in the articles [9, 8, 10], the final manuscripts of which are available from Kyoto University Research Information Repository. Bingham distribution was named after C. Bingham [1], and is concisely introduced by the book and article by Borradaile and Love [3, 6]. Directional statistics is detailed by the textbooks [5, 7]. 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]. When the present software is used to determine stresses from dilational fractures in a rock mass, the ratio, DPI $= -3/\kappa_1$, represents the frequency of the events that high pressure fluids intruded the mass [4].

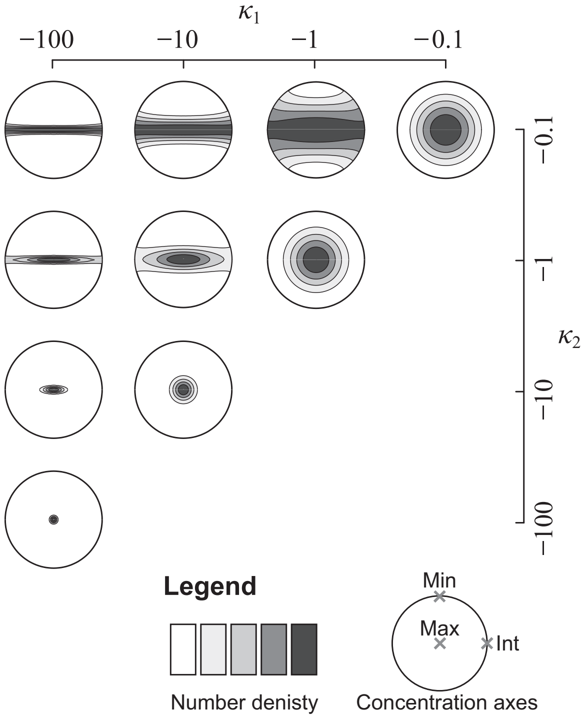
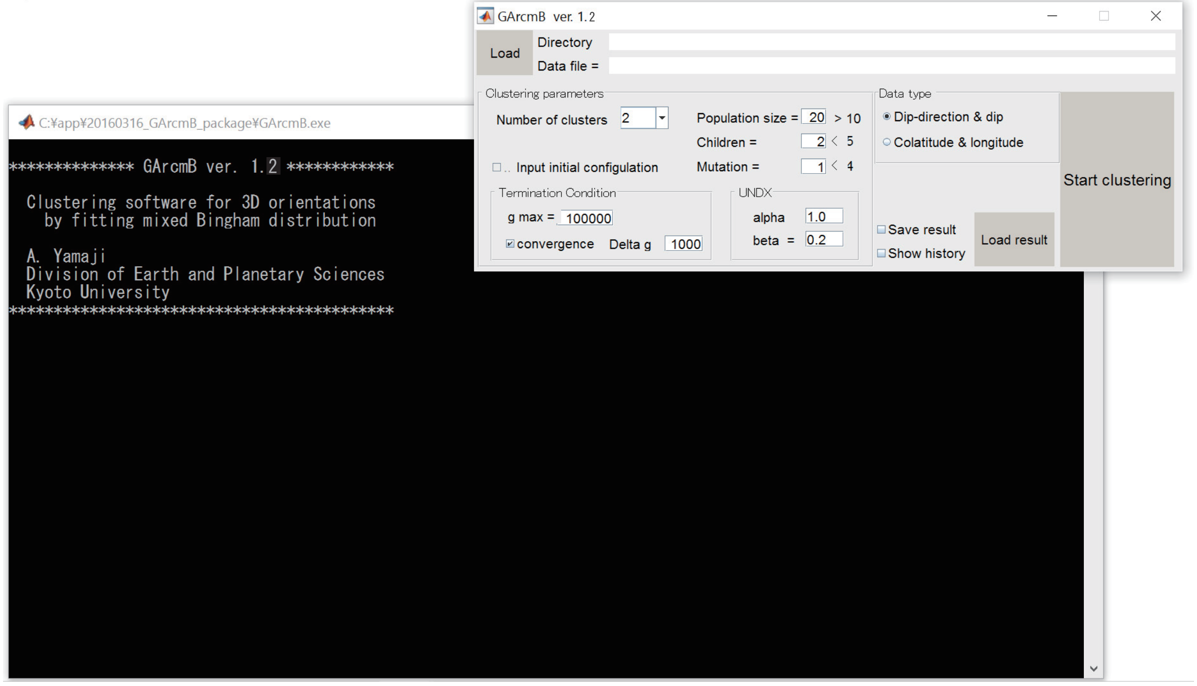


Figure 1: Sizes and shapes of Bingham distributions and the concentration parameters, κ_1 and κ_2.

It is happy for the author if the users of this program cite one or both of the papers [8, 9] in their articles.

2 Installation

The present software was developed in Matlab, and was complied to make executable files for the users who do not have Matlab. Users have to build the environment in which Matlab program runs. The Matlab Runtime available from MathWorks® builds the environment in your computer. The present software runs on your computer after the following procedure.



C:¥app¥20160316_GArcmB_package¥GArcmB.exe

```
************** GArcmB ver.  1.2 ************

  Clustering software for 3D orientations
    by fitting mixed Bingham distribution

  A. Yamaji
  Division of Earth and Planetary Sciences
  Kyoto University
*******************************************
```

GArcmB ver. 1.2

Directory

Data file =

Clustering parameters

Number of clusters 2

☐ .. Input initial configulation

Population size = 20 > 10

Children = 2 < 5

Mutation = 1 < 4

Termination Condition

g max = 100000

☑ convergence Delta g 1000

UNDX

alpha 1.0

beta = 0.2

Data type

● Dip-direction & dip

○ Colatitude & longitude

☐ Save result

☐ Show history

Figure 2: The console window of MATLAB (lower left) and the control panel of the present software (upper right).

1) Operating system Make sure of the OS of your computer. The program runs on Windows, Mac and Linux.

2) MATLAB Runtime Visit the homepage of MathWorks® to download the MATLAB Runtime version R2024b, 64-bit. Refer to the web page on how to install it on your computer. The present author obained the runtime at the website, `https://mathworks.com/products/compiler/matlab-runtime.html` on March 19, 2025.

3) Program package From the homepage, `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`, to check weather the installation was successful. If the Runtime and the package are installed correctly, the console window and the control panel in Fig. 2 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 and the package.

3 Formats of data files

The present software 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. 3.

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 has the dip-direction and dip of a fracture (Fig. 4a). 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, colatitude 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.'

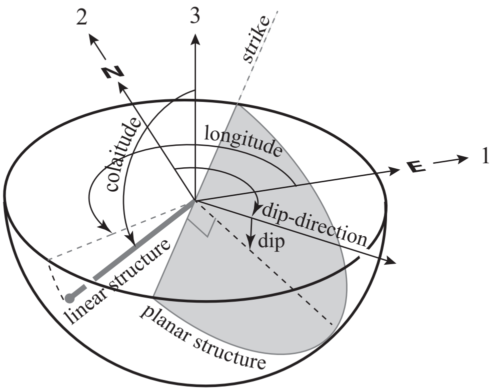
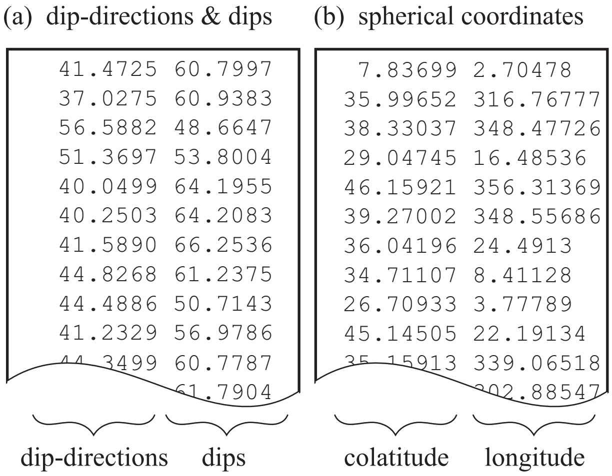


Figure 3: The rectangular Cartesian coordinates used in the software. The 1- and 2-axes are oriented north- and eastward, respectively. The dip-direction of a planar structure, e.g., a fracture, is measured from the north. The orientation of a linear structure is described by spherical coordinates, longitude of which is measured from the east. Equal-angle projections of the lower-hemisphere are used.


(a) dip-directions & dips

```
41.4725 60.7997
37.0275 60.9383
56.5882 48.6647
51.3697 53.8004
40.0499 64.1955
40.2503 64.2083
41.5890 66.2536
44.8268 61.2375
44.4886 50.7143
41.2329 56.9786
44.3499 60.7787
        61.7904
```

(b) spherical coordinates

```
 7.83699 2.70478
35.99652 316.76777
38.33037 348.47726
29.04745 16.48536
46.15921 356.31369
39.27002 348.55686
36.04196 24.4913
34.71107 8.41128
26.70933 3.77789
45.14505 22.19134
35.15913 339.06518
          02.88547
```

Figure 4: 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).

4 How to use the software

4.1 Basic operation

Since the present software has graphic user interface, it is easy to use the software.

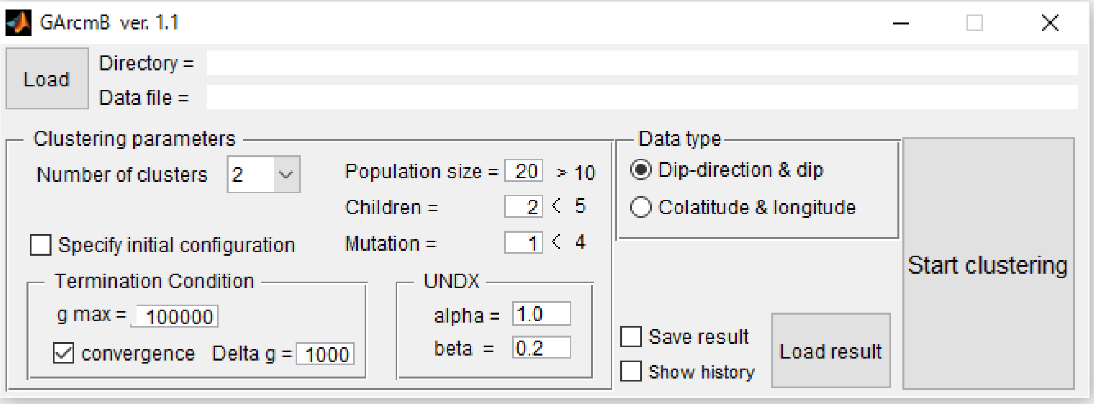
1. Double-click on the icon of the file, `GArcmB`, to launch MATLAB. Then, the console window and the control panel of the software pop up on your computer screen (Fig. 2).
2. Press the button Load at the upper left of the panel to load a data file.
3. Choose the data type, 'Dip-direction & dip' or 'Colatitude & longitude' by clicking a radio button in the panel 'Data type.'
4. Select the number of clusters, K.
5. To save the workspace 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 log-likelihoods of individuals of the population versus generation (Fig. 6). However, the plotting spend a lengthy time.
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 Command Window of MATLAB and the computation is terminated.

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

7. If the termination condition is met, the software finishes the genetic algorithm, and draw Mohr diagrams for fractures (Fig. 11). If you do not deal with geological data, just ignore the diagrams. If the checkbox, 'Save result,' was clicked before the calculation, all values in the workspace is 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 `-100.33` is the log-likelihood of the best partition found by the calculation. This output file name is automatically generated, and the output file is saved in the directory containing the input file. The saved result can be reloaded by pushing the button Load result.
8. If the computation comes to an and successfully, the program shows the time of computation and the message

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

at the end in the command window.



Number of clusters 2

Population size = 20 > 10

Children = 2 < 5

Mutation = 1 < 4

☐ Specify initial configuration

g max = 100000

☑ convergence Delta g = 1000

alpha = 1.0

beta = 0.2

● Dip-direction & dip

○ Colatitude & longitude

☐ Save result

☐ Show history

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

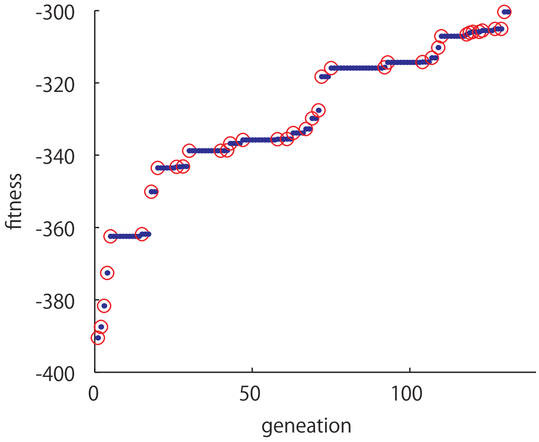


Figure 6: The diagrams showing the clustering history drawn in the graphic window ‘Figure 12,’ which appears just after the clustering is started if the checkbox ‘Show history’ was clicked before the start.

9. Record the final L value, which is indicated in the command window (Fig. 7), in a green cell of an Excel sheet (Fig. 8). 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 command 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.

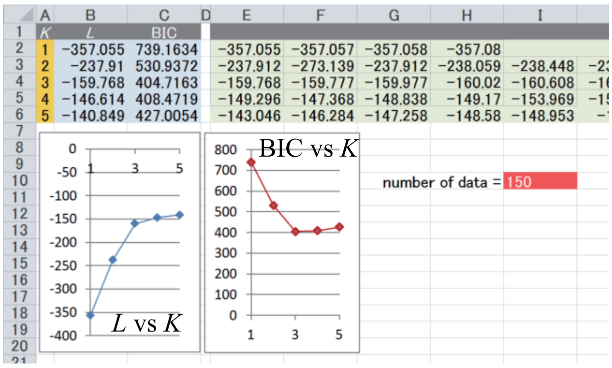
```
========== generation 59 ==========
  fractures.txt
  L = -325.744201   BIC = 736.669201
  Cluster 1
    kappa = (-10.571, -5.017)  Phi = 0.4746  varpi = 0.3481
    min:  75.5/26.8,  int: 195.9/45.1,  max: 326.5/32.9
  Cluster 2
    kappa = (-7.659, -2.649)  Phi = 0.3459  varpi = 0.3287
    min: 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 7: The progress of the clustering is shown in the command window of MATLAB. This example shows the status at the 70th generation when the file, fractures.txt was processed.

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, press Control-C in the command window.



	A	B	C	D	E	F	G	H	I
1	*K*	*L*	BIC						
2	1	−357.055	739.1634		−357.055	−357.057	−357.058	−357.08	
3	2	−237.91	530.9372		−237.912	−273.139	−237.912	−238.059	−238.448
4	3	−159.768	404.7163		−159.768	−159.777	−159.977	−160.02	−160.608
5	4	−146.614	408.4719		−149.296	−147.368	−148.838	−149.17	−153.969
6	5	−140.849	427.0054		−143.046	−146.284	−147.258	−148.58	−148.953

number of data = 150

Figure 8: Example of the spread sheet that the results of computation for the same data set is summarized.

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 [8] for details of the parameters.

4.4 Progress of clustering

The progress of clustering can be seen in the command window of MATLAB and three graphic windows, which appear shortly after the clustering launched. Fig. 7 shows the command window, where `L` indicates the log-likelihood of the temporally best individual, and `BIC` indicates the corresponding BIC value. In this example, the data were partitioned into three clusters. `kappa` denotes the concentration parameters, κ_1 and κ_2, of the Bingham distribution fitted to a cluster. `Phi` is the stress ratio, $\Phi = (\sigma_2 - \sigma_3)/(\sigma_1 - \sigma_3) = \kappa_1/\kappa_2$ [10], and varpi is the mixing coefficient, ϖ.

As soon as the button Start clustering is pressed, two graphic windows like those in Figures 11 and 13 appears. If the checkbox 'Show history' was clicked, 'Figure 12' appears as well. Fig. 6 is an example: The left panel of this figure shows the log-likelihood of the best individual versus generation. Red circles indicate the generation at which the the best log-likelihood of the population increased. Blue line in this plot shows the temporally best log-likelihood of the population. The right-panel in Fig. 6 show the log-likelihoods of individuals in the last 200 generations. Colored lines show those of the individuals.

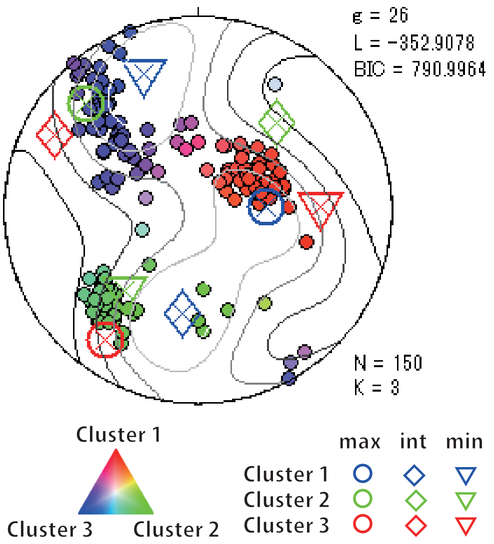


Figure 9: Lower-hemisphere, equal-area projections showing the temporal result of clustering. The colors of data points indicate the memberships of the data to Clusters 1, 2 and 3. The maximum, intermediate and minimum concentration axes of each Bingham component are indicated as well. The ternary diagram depicts the memberships. Contour lines indicating the probability density distribution of the temporal mixed Bingham distribution.

The memberships of data points and the concentration axes of clusters are indicated by colors and symbols, respectively, in the equal-area projection (Fig. 9). Figure 10 shows 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. 10a, the EPS file of which are included in this software 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 in Fig. 9 can be made up with drawing software, e.g., Illustrator and Canvas, for publication (Fig. 12). You can use the color maps contained in the

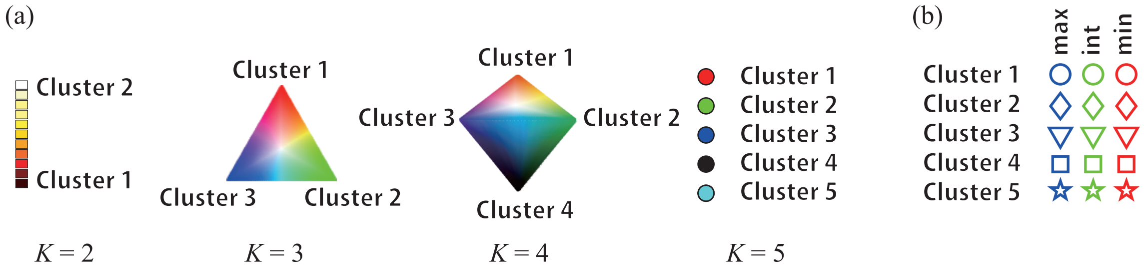


Figure 10: Colors and symbols used in equal-ara projection (Fig. 9). (a) Color schemes for indicating the memberships of data points. (b) Symbols for indicating concentration axes.

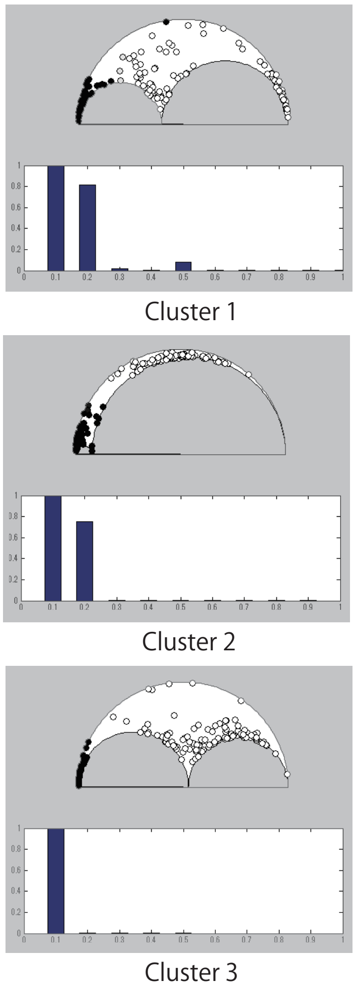


Figure 11: Mohr diagrams for the clusters and the bar graphs for the memberships of summed in the bins with the width of 0.1. See the paper [8] in detail.

Table 1: List of symbols

g	generation
K	number of clusters
L	log-likelihood function
N	number of data
ϖ	mixing coefficient
Φ	stress ratio

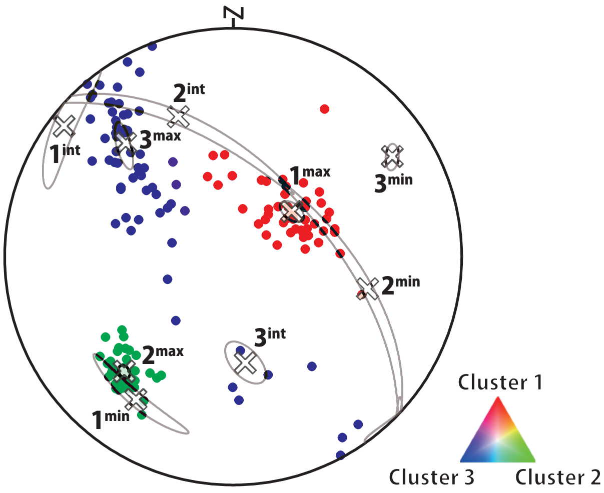


Figure 12: Lower-hemisphere, equal-area projections showing the optimal partition of the sample data 'fractures.txt.' The maximum, intermediate and minimum concentration axes of the kth cluster is denoted by crosses. The labels attached to the crosses distinguish the clusters and their concentration 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.

files, 'colors.eps,' for this purpose.

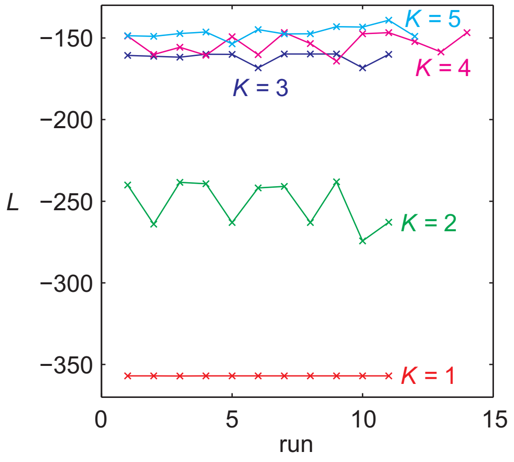


Figure 13: L versus run number for the data set in Fig. 12.

5 Tips to find the best partition

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 [8]. However, this function has numerous maximum points: The point where the function has the global maximum should be found by the computation. Figure 15 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 15a 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. 15b 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. 15a), whereas there were data points with intermediate memberships in Figs. 15b and c.

The multi-modality of the function, $L(\boldsymbol{X})$, makes the search problem difficult. Although the present 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 some 10 times for each of the cases of $K = 1$ through 5.

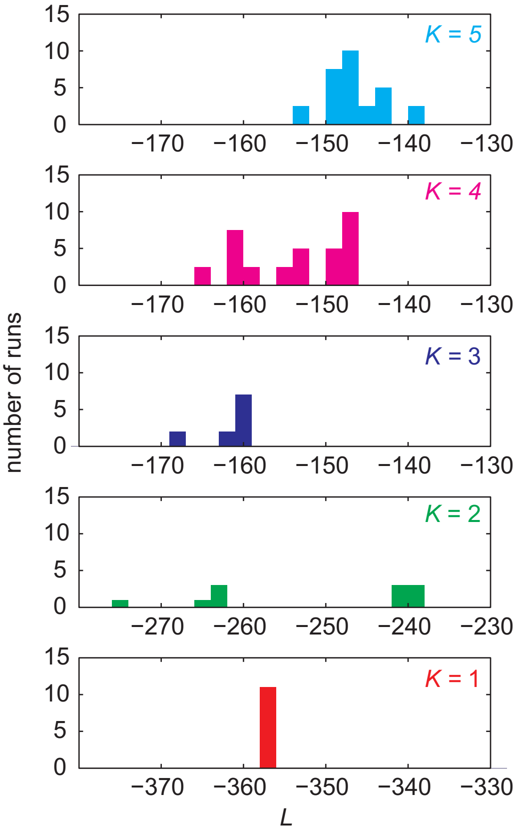


Figure 14: Histograms of the L values in Fig. 13.

Figure 16 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.* 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 [8] 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.

5.1 Tuning of the parameters of genetic algorithm

The user can change the parameter values of the genetic algorithm for the clustering by using the items in the block entitled ‘Clustering parameters’ on the control panel (Fig. 5). If the computation seems not to detect the best partition, change some of the values. Consult the paper [8] for the details of the parameters.

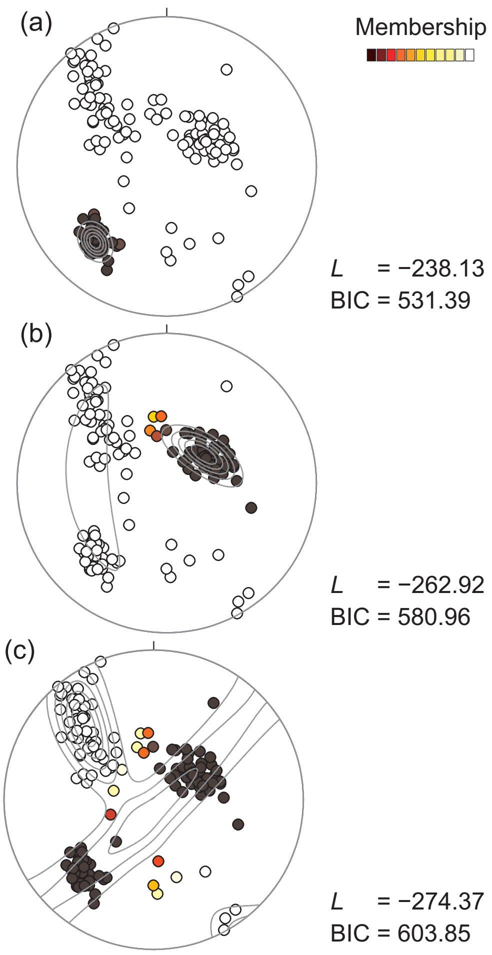


Figure 15: Equal-area projections showing the three examples of the partitions of the sample data into three clusters. Contour lines denote the iso-density lines of the mixed Bingham distribution that was fitted to the data set. (a) The optimal partition with the maximum L and minimum BIC values. The 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.

5.2 Initial configuration

By pressing the button Start clustering on the control panel (Fig. 5), the program randomly generates K Bingham distributions as the initial configuration of the genetic algorithm. However, this configuration can be specified by the user: To do this,

1. Prepare a file in which the initial configuration is described.

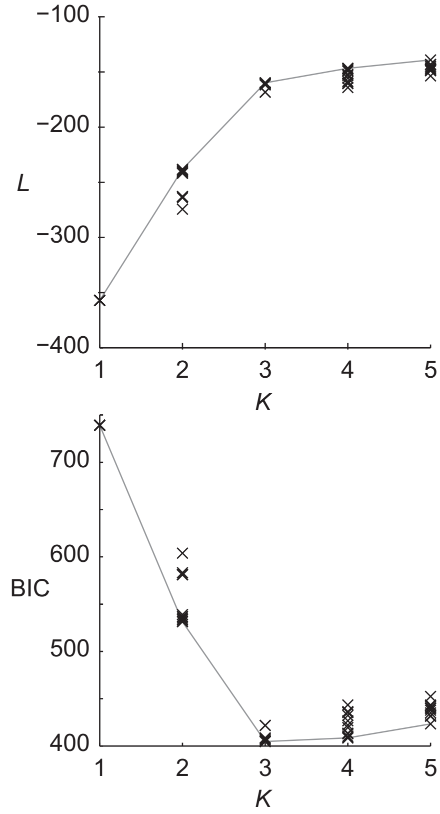


Figure 16: Diagram to examine the convergence of computations for a data set.

2. Click the checkbox input initial configuration.

3. Press the button Start clustering.

Test this procedure by using the file `init_config_fractures_k3.csv` in the program package, which the approximate configurations of three clusters that should be detected from the sample file `fractures.txt` are entered.

The configurations should be described with the format shown in Table 2. The items in the list are separated by commas. See the contents of the sample file, `init_config_fractures_k3.csv`. The properties of a cluster occupies a row of the list: This example specifies the initial configurations of three clusters. The first and second items of the row are the azimuth and plunge of the minimum concentration axis whereas the third and fourth ones those of the maximum concentration axis. The fifth and sixth ones are the concentration parameters, κ_1 and κ_2 (Fig. 1). Note that these parameters are negative in sign satisfying $0 \geq \kappa_2 \geq \kappa_1$. These six values specify a Bingham distribution. The genetic algorithm starts not exactly from the Bingham distribution, but from a Bingham

Table 2: Format of the comma-delimited file that the initial configurations of clusters are entered.

Min. con. ax.		Max. con. ax.		κ_1	κ_2	$\Delta\kappa$
az	pl	az	pl			
102.3,	40.6,	223.4,	31,	−84.413,	−39.807,	0.1
57.8,	20.6,	316.7,	27.1,	−50.266,	−3.163,	0.1
210,	23.8,	58.9,	63.2,	−50.133,	−15.848,	0.1

distribution that is approximately equal to the specified one. The final item of the list, $\Delta\kappa$, indicates the difference of them. The value of this parameter should be set to satisfy $0 < \Delta\kappa \ll |\kappa_2|$.

The list in Table 2 shows the initial configuration of three clusters. What if the number of clusters is set two on the control panel (Fig. 5)? Then, the software loads only the first two lines of the list at the beginning of the computation. If the number of clusters is four, then the three lines are loaded, and the fourth cluster is randomly initialized.

The file can be created with spreadsheet software or a text editor, and should be saved as a csv (comma-separated-value) file with the file extension '.csv.'

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] Faye, G.D., Yamaji, A., Yonezu, K., Tindell, T. and Watanabe, K., 2018. Paleostress and fluid-pressure regimes inferred from the orientations of Hishikari low sulfidation epithermal gold veins in southern Japan. *Journal of Structural Geology*, **110**, 131–141.

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

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

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

[8] 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.

[9] 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.

[10] 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.

Appendix A: Changelog

Ver. 1.2.2, March 25, 2025 The seed of the random number generator used in the software was updated.

Ver. 1.2.1, March 19, 2025 The version of MATLAB Runtime was updated to R2024b.

Ver. 1.2, April 9, 2016 The first version open for the public.