BasEMMA——Basic End-member Model Algorithm for sediment grain size

1 What is end-member inversion of sediment grain size

The grain size frequency distribution curve of sediment often has multiple peaks. Sedimentologists believe that this multi-peaks phenomenon reflects the different provenance, transportation and settlement processes of sediment. Early researchers used traditional statistical methods to analyze sediment grain size data. However, owing to the sediment grain size data is compositional data (sum to one and non-negative), traditional statistical methods are not suitable. Researchers now generally use End-Member Model Algorithm (EMMA) to process sediment grain size data. The end member inversion solves the mathematical problem shown in Equation 1.

$$D = A\,B + E \qquad (1)$$

Where, D is sediment size data, A is coefficient matrix, B is end-member matrix, and E is error.

Through the grain size analysis to sediment, D can be obtained. The purpose of end member inversion is to obtain coefficient A and end member B according to D, ignoring the error E.

2 The non-uniqueness of end-member inversion

Mathematically, Equation 1 has an infinite number of solutions. However, as mentioned above, the grain data D is the component data. The bad news is it is not suitable to the traditional statistical methods, but the good news is that the sum to one and non-negative restrictions exclude many solutions, making the possibility of acquiring the unique A and B is greatly increased. To illustrate this problem, let us first introduce what end-member space is. The end member space is the space where the end member data is located. For three end members, the end member space is a triangle (Figure 1), for four end members, the end member space is a tetrahedron, and so on. The end-member space is basically equivalent to the real space minus one dimension and restricted by abovementioned constraints.

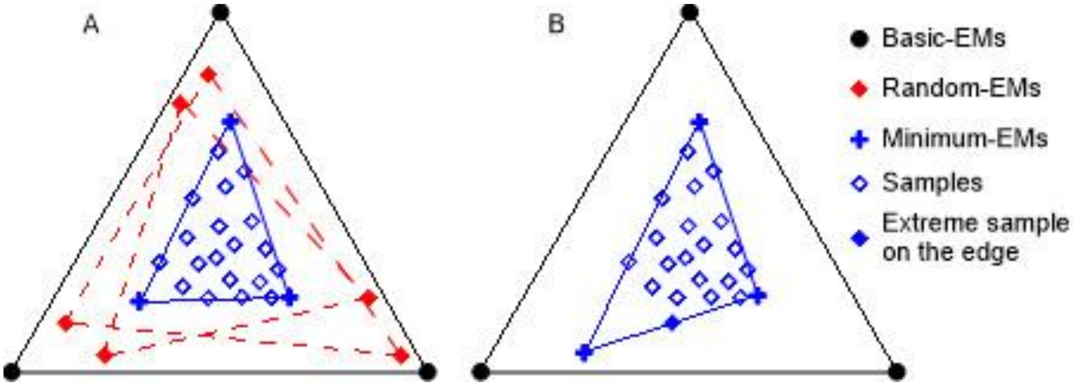

Figure 1. The non-uniqueness of the end-member inversion results (A), the Minimum end-members susceptible to errors (B)

Taking the three end-elements as an example, if the sample data are evenly distributed throughout end-member space, the inversion result of EMMA should be basically consistent with the actual end member. However, because the spatial range of the sample data is often limited and cannot cover all the typical samples in the entire sedimentary system, more often, our research object is a core at a certain location, and the sedimentary environment is relatively stable or the sedimentary characteristics is not comprehensive compared with the whole sedimentary system, the sample data will be inevitably clustered in a local area of the end-member space, resulting in the non-uniqueness of the end-member inversion results.

3 Previous EMMAs

Most of the previous EMMAs ignored the above-mentioned non-uniqueness of the end-member inversion results, and only solve the Equation 1 based on the best-fit rule of traditional statistics, and the end-members obtained are called "Random-EMs" (Figure 1A). Some researchers try to find the "Minimum-EMs" that tightly envelope all data (Figure 1B). In sedimentology, the inversed end-members by previous EMMA are still mixtures of multiple sources or transportation processes.

4 BasEMMA

We call the outermost points in the end-member space (the vertexes of the triangle in Figure 1) as the basic end-members. The linear combination of the basic end-members can synthesize all the data in the end-member space. BasEMMA is the software we developed to find the basic end-members in the end-member space. The software can be downloaded from https://github.com/ouczxd/BasEMMA. For the English description and usage of the software, see Zhang, X.D., Wang, H.M., Xu, S.M., Yang, Z.S., 2020. A basic end-member model algorithm for grain-size data of marine sediments, Estuarine, Coastal and Shelf Science 236, 106656. https://doi.org/10.1016/j.ecss.2020.106656.

BasEMMA is based on an Excel file and uses VBA programming. Data input and calculation results are stored in the Excel file, too. The software runs directly in Excel, and the intermediate results can be viewed in real time through Excel charts. We used predecessors' test data and the test data we made based on the sediment grain size data from the East China Sea to test BasEMMA. The test results show that BasEMMA can get accurate results under various conditions. We also inverted the sediment grain size data of the core 30 from the East China Sea, and gave an example of using BasEMMA to study sediment grain size data.

5 software usage

BasEMMA is an Excel file: "Appendix A BasEMMA.xlsm". The software is written in VBA, so you need to allow the work of Macros to run BasEMMA when opening the Excel file, or other similar security prompts (depending on different versions of the Microsoft Excel).

The file contains 6 worksheets. The Data worksheet is used to store your grain size data. See the sample for the data format. What needs to be explained is that the genetic algorithm is used by BasEMMA, and the genetic algorithm is an optimized search algorithm with good robustness and error-resistance, but longer operation time. To the sample data, it will take about 5-10 minutes. We recommend that the number of grain grades is 8-40 and the number of samples is 30-1000. It is recommended to delete the grain grades with all zero data. The grades with smaller content at both ends (the coarsest and finest grades) can be merged for efficiency.

The Setup worksheet is used to set the parameters of the software. The main parameters include: EM number from ? to ?. It is recommended to use the EM number from 2 to 5. Our experience is that after the EM number exceeds 5, the remaining information in the grain size data is not enough to exceed the influence of errors and is difficult to extract. If you really need to find more than 5 EMs, you need to modify 5 to the number you want, but not more than 9, because the space for storing the data is set to store up to 9 columns of coefficients. After modification, the corresponding tables for storing the inversion results need to be added. The names of the tables are "EM6", "EM7", etc. Grain-size number is the number of grain grades, Sample number is the number of samples, Maximum generation number is the iterations number of the genetic algorithm, 30 is basically enough, 100 is the standard value, 300 is the best. During the calculation, you can end the evolution (calculation) at any time by pressing the "Ctrl+Break" key. On some computers, the Break button is marked as blue Pause, and you need to press "Ctrl+Fn+Pause" at the same time. Population number is the population size in the genetic algorithm, 300 is the standard case, you can adjust it between 100-1000, the larger the population size, the longer the calculation time.

The EM* worksheet is used to store the inversion results. The cell (1, 1) stores the current generation number and the total generation number. Cell (2, 1) stores the distance between end members, the larger the better, the maximum is 1. Cell (2, 2) stores the difference between the inverted grain size data and the actual grain size data, the smaller the better, and the minimum is 0. Cell (1, 2) stores the difference between cell (2, 2) and cell (1, 2). The smaller the better, it may be a negative value. The remaining part below the 1st and 2nd columns stores the above distance and difference for each generation. Columns 3 and 4 are the determination coefficient of each grain grade, columns 5-6 are the determination coefficient of each sample, column 7 is the difference between the inversion result and the actual data, and columns 8 and 9 are each the average content of end members, the 10th column and the right * columns are the end member coefficients of each sample, the 20th column and the right columns are the frequency distribution of each end member.

Operating procedures:

(1) Input grain data

(2) Set parameters

(3) Press the "start" button in the setup Sheet to watch the inversion process, which will take about 5-10 minutes

(4) Determine the number of end-members according to Zhang et al, (2020).

(5) Set EM number from 3 to 3 to perform a more thorough inversion again if your EM number is 3.

6 Determination of the number of end elements

Generally 3-4 are enough. When determining the number of end members, the actual needs should be considered first, then the uncertainty of the inversion result should be considered, and finally the inflection point should be considered. In terms of actual needs, you can observe the determination coefficient of grain grades.

7 Explanation of end elements

Even if the goal of BasEMMA is to find the basic end members, the grain size frequency distribution curve of the end members obtained by BasEMMA sometimes has double peaks, and the secondary peak mainly appear in the very fine grain size (generally less than 2 microns). We believe it should be caused by "capture effect of coarse particulate matter on very fine particulate matter", “the effect of flocculation” and errors.

By Xiaodong Zhang (zxd@ouc.edu.cn) and translated by Ms. Yuhan Yao.

2020/9/7