BasEMMA——沉积物粒度端元反演软件

1 什么是沉积物粒度端元反演

沉积物的粒度频率分布曲线往往具有多个峰，沉积学家认为这种多峰的现象反映了沉积物不同的物源、输运和沉降过程。早期的研究者采用传统统计方法对沉积物粒度数据进行分析，但沉积物粒度数据是成分数据（总和为1，非负），传统统计方法存在不适格的问题。现在的研究者普遍采用端元反演软件（End-Member Model Algorithm，简称EMMA）处理沉积物粒度数据。端元反演解决的是公式1所示的数学问题。

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

其中，D是沉积物粒度数据，A 是系数矩阵，B是端元矩阵，E是误差。

根据沉积物粒度分析可以得出D，端元反演的目的是根据D得出系数A和端元B，并估计误差E。

2 端元反演的多解性

在数学上，公式1有无穷多的解。但是，如上所述，粒度数据D是成分数据，不利的是成分数据使用传统统计方法处理存在问题，但有利的是其总和为1并且非负的限制排除了许多解，使得根据D得出唯一的A和B的可能性大大增加。为说明这个问题，先介绍一下什么是端元空间。端元空间是端元数据所在的空间，对于三端元来说，端元空间是一个三角形（图1），对于四端元来说端元空间是一个四面体，以此类推。端元空间基本相当于降了一维并受了限制（位于0-1之间）的现实空间。

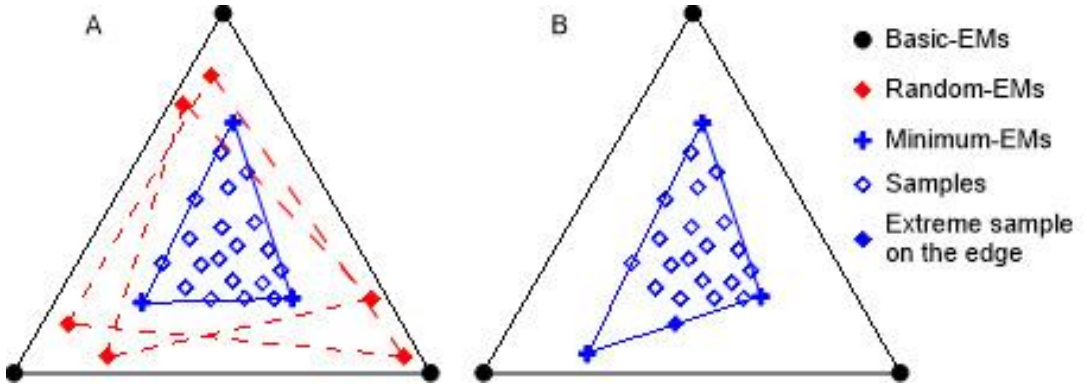

图1. 沉积物粒度端元反演结果的不唯一性(A)，易受误差影响的“最小端元”(B)

以三端元为例，样品数据分布在如图1所示的端元空间（三角形）中。如果样品数据均匀分布在整个空间，EMMA的反演结果应与实际端元基本一致。但是，由于我们取得的样品的空间范围往往有限，不能涵盖整个沉积体系所包含的所有区域，更经常的是我们的研究对象是某一位置的岩芯，其沉积环境相对于整个沉积体系来说较为稳定或者沉积特征不够全面。反映在沉积物粒度数据上，样品数据不可避免地会聚焦在端元空间的一个局部区域，这样利用EMMA进行反演就会出现多解性。

3 前人的 EMMA

以前大部分的EMMA忽略上述多解性问题，仅依据传统统计学的最佳拟合度指标反演公式1，其得出的端元我们称为“随机端元”（Random-EMs）（图1A），有的研究者试图寻找紧密包络所有数据的“最小端元”（Minimum-EMs）（图1B）。上述做法的结果是反演出的端元离实际的端元比较远，在沉积学上可以理解为其反演出的端元仍然是多种物源或者输运过程的混合物，在粒度频率分布上表现为多峰或者峰态较小（粒度分布较广）的情况。另外，“最小端元”易受误差的影响（图1B），进而导致对反演结果的误解。

4 BasEMMA（Basic End-Member Model Algorithm）

我们称在端元空间中最靠外的点（图1中三角洲的顶点）为基本端元。利用基本端元的线性组合可以合成端元空间中所有的数据。BasEMMA（Basic End-Member Model Algorithm）就是我们开发的用来寻找端元空间中基本端元的软件。软件下载地址为：https://github.com/ouczxd/BasEMMA。软件的英文说明和用法详见：

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基于Excel文件并使用VBA编程，数据输入和运算结果都存储在Excel文件中，软件在Excel中直接运行，中间结果可以通过Excel图表实时查看。我们利用前人的测试数据以及我们基于东海陆架沉积物粒度数据制作的测试数据对BasEMMA进行了测试，测试结果表明BasEMMA在各种情况下都能得出准确的结果。我们还对东海陆架30号岩芯的沉积物粒度数据进行了反演，给出了利用BasEMMA研究沉积物粒度数据的范例。

5 软件用法

BasEMMA是一个Excel文件：“Appendix A BasEMMA.xlsm”。软件采用VBA编写，因此在打开该文件时需要允许运行宏，或者类似的其他安全提示（视不同的office版本而定）。

文件包含6个工作表。Data工作表用来存放你的粒度数据，数据存放格式参见样例。需要说明的是BasEMMA是使用遗传算法进行求解的，遗传算法是一种优化搜索算法，其鲁棒性和抗差性好，但运算时间较长。样例中的200个样品28个粒级运算一次大概需要5-10分钟。我们推荐粒度数据的粒级在8-40之间，样品数量在30-1000之间。数据越多分级越细计算时间越长，但据我们的经验，超过20-30个粒级后更加细致的分级并不能对端元反演结果

带来新的改观。8个分级的情况也是可以的，当然其得出的端元的粒度频率分布不够光顺，但其系数基本不受影响。建议删去含量全为0的粒级，如果为了效率可以合并两端含量较小的粒级（最粗和最细），这样可以提高软件的效率。

Setup工作表用来设置软件的参数。主要参数包括：EM number from ？ to ？。建议采用样例中的EM number from 2 to 5进行操作，我们的经验是端元数量超过5之后，粒度数据中剩余的信息不足以超越误差的影响，难以提取。如果你确实需要超过5个以上的端元数量，那么需要修改5为你希望的数量，但最多不能超过9个，因为后面存储数据的空间设定为最多存储9列系数。修改后需要增加相应的存储反演结果的表格，表格的命名分别为“EM6”,“EM7”等。Grain-size number是粒度数据的粒级数量，Sample number是样品数量，Maximum generation number是遗传算法的代数，30代是基本够用的，100代是标准值，300代就基本不会再有较大的改进了。在计算的过程中，可以通过按“Ctrl+Break”组合键随时结束进化（计算），在有的电脑上Break按键标注为蓝色的Pause，需要同时按下“Ctrl+Fn+Pause”。Population number是遗传算法中的种群规模，300是标准情况，你可以在100-1000之间进行调整，种群规模越大计算时间越长。

EM*工作表用来存放反演结果。单元格（1，1）存放当前代数和总代数。单元格（2，1）存放端元之间的距离，越大越好，最大为1。单元格（2，2）存放反演粒度和实际粒度之间的差值，越小越好，最小为0。单元格（1，2）存放单元格（2，2）与单元格（1，2）的差，越小越好，可能为负值。第1和2列下面的其余部分存储每代的上述距离和差值。第3和4列为每一粒级的平均决定系数，第5-6列为每一样品的平均决定系数，第7列为反演结果和实际数据的差值，第8和9列为每一端元的平均含量，第10列及其右边的*列为每一样品的端元系数，第20列及其右边的*列为每一端元的频率分布。

操作流程：

（1） 输入粒度数据

（2） 设置相关参数

（3） 按setup表中的“start”按钮，观看反演过程，约需5-10分钟

（4） 根据Zhang et al.,(2020)确定端元数量。

（5） 设置EM number from 3 to 3再进行一次较为彻底的反演。

6 端元数量的确定

一般3-4个就足够了。在确定端元数量时，应首先考虑实际需要，然后考虑反演结果的不确定性，最后参考拐点。在实际需要方面，重点观察粒级决定系数，主要粒级的决定系数普遍大于0.6-0.7就可以了。

7 对端元的解释

即便BasEMMA的目标是寻找基本端元，反演得出的端元的粒度频率分布曲线有时候也会存在双峰现象，次峰主要出现在极细粒级（一般小于2微米）。我们前期的研究对该现象的解释可能不够全面。前期我们认为是“粗颗粒物质对极细颗粒物质的捕获作用”造成了这一现象，现在看来造成这种现象的原因较多，不仅包括上述捕获作用，还有絮凝作用以及误差的影响等。总体上，极细颗粒泥沙的输运可能并不对应于特定物源或者输运过程，也有可能是极细颗粒由于含量较少，其物源和输运规律被大量的其他粒级的物质所掩盖了。

张晓东 zxd@ouc.edu.cn

2020/9/2