非负矩阵分解方法在水系沉积物地球化学数据处理中应用

张生元[1,2], 黄 锐[2], 徐德义[2], 成秋明[2]

1. 石家庄经济学院资源与环境工程研究所, 河北石家庄 050031
2. 中国地质大学地质过程与矿产资源国家重点实验室, 湖北武汉 430074

摘要: 鉴于水系沉积物地球化学数据可以表示为非负矩阵, 这使得利用非负矩阵分解(NMF)方法处理该类数据成为可能. 介绍了非负矩阵分解方法的基本原理和方法, 讨论了基于非负矩阵分解方法处理水系沉积物地球化学数据的可能和效果. 以个旧水系沉积物地球化学数据为例, 运用 NMF 方法和主成分分析(PCA)方法对其进行异常分析, 并对这两种方法的处理结果进行了比较, 发现 NMF 方法对于处理水系沉积物地球化学数据是一种有效的方法. 尽管这两种方法各自有其优越性, 但就本实例数据而言, NMF 方法优于 PCA 方法.

关键词: 非负矩阵分解; 主成分分析; 地球化学数据.
中图分类号: P628 文章编号: 1000-2383(2009)02-0347-06 收稿日期: 2009-01-16

Application of Non-Negative Matrix Factorization in Stream Sediment Geochemical Data Processing

ZHANG Sheng-yuan[1,2], HUANG Rui[2], XU De-yi[2], CHENG Qiu-ming[2]

1. *Institute of Natural Resource and Environment Engineering, Shijiazhuang University of Economics, Shijiazhuang 050031, China*
2. *State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan 430074, China*

Abstract: This paper explores the possibility of applying non-negative matrix factorization (NMF) to process stream sediment geochemical data for mineral exploration. The brief introduction of principle of NMF is followed by detailed comparison of the results obtained by NMF and principal component analysis (PCA) applied to a dataset of 813 samples with six trace elements from Gejiu mineral district, Yunnan, China. It is shown that the NMF is not only suitable for processing geochemical data which are usually of positive values but also provides superior results than that by PCA in the case study introduced in the paper. The example indicates that NMF might become a useful method for processing other types of geochemical data.

Key words: non-negative matrix factorization; principal component analysis; geochemical data processing.

矩阵分解是实现大规模数据处理与分析的一种有效工具. 把矩阵分解为形式比较简单或具有某种特性的一些矩阵的乘积或叠加, 在矩阵理论的研究与应用中都是十分重要的. 在数值分析领域, 利用矩阵分解可以将计算规模较大的复杂问题转化为一系列计算规模较小的简单子问题来求解; 在应用统计学领域, 通过矩阵分解得到原数据矩阵的低秩逼近, 更易于发现数据的内在结构特征; 在机器学习和模式识别方面, 矩阵的低秩逼近可以大大降低数据特征的维数, 节省存储和计算资源; 在自适应滤波中, 常常利用矩阵分解的特殊形式来减少计算的复杂性、提高滤波器的性能等. 在地球化学数据处理中常用 PCA 方法寻找地球化学元素的有用组合以提取地球化学元素的组合异常, 为地质找矿提供依据.

从数学计算的角度来看, 矩阵分解结果中存在负值是允许的, 但负值元素在实际问题中往往是难

基金项目: 国家自然科学基金重点项目(No. 40638041); 地质调查项目(No. 1212010633901 10); 地质过程与矿产资源国家重点实验室开放课题(No. GPMR200803); 国家 863 项目(No. 2006AA06Z115); 地质过程与矿产资源国家重点实验室科技部专项经费资助.
作者简介: 张生元(1961-), 男, 博士, 教授, 主要从事矿产资源定量评价方法、科研开发和教学工作. E-mail: zhangsh3002@126.com

以解释的.例如图像数据中的灰度值一般不能有负值,物质成分的含量也总是非负数值等.因此,研究矩阵的非负分解方法是一项很有意义的工作.

早在 20 世纪 70 年代就已经有数学家针对非负矩阵做了一些相关的研究工作,但没有引起过多的关注.到 20 世纪 90 年代,科学家已进行了与非负矩阵分解类似研究,如,正矩阵分解(PMF)(Juvela *et al*., 1994, 1996)应用于环境处理和卫星遥控(Paatero and Tapper, 1994; Paatero, 1997).1999 年 Lee 和 Seung 两位科学家在著名杂志《Nature》上提出"非负矩阵分解"(non-negative matrix factorization, NMF)的概念和算法(Lee and Seung, 1999)用于寻找一组基函数表示非负矩阵的方法.这种新的矩阵分解算法要求矩阵中所有元素均为非负的.在以往的矩阵分解方法中,原始矩阵 $\boldsymbol{V}$ 被近似分解为低秩的 $\boldsymbol{V}\approx\boldsymbol{WH}$ 形式,这些方法的共同特点是分解因子 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 中的元素可正可负,即使原始矩阵元素是全正的,传统的算法也无法保证分解后矩阵的非负性,且正负相互抵消不利于数据特征的提取和解释.而 NMF 则是在初始矩阵及分解矩阵中所有元素均为非负数约束条件之下的矩阵分解方法,这种非负性条件符合许多实际问题的要求,更便于对分析结果的解释.

NMF 方法用于人脸表情识别比传统方法的识别效果更好,根据非负矩阵分解的特点,将其用于文本和日志分析领域,设计有针对性的分类和聚类算法,达到了较好的效果(Lee and Seung, 1999; Xu *et al*., 2003);利用非负矩阵分解算法进行盲信号分离,证明 NMF 方法在分离统计独立信源、统计相关信源以及高斯分布信源上非常有效(魏乐, 2004).在诸多应用中 NMF 能用于发现数据库中的图像特征,便于快速自动识别;能用于发现文档的语义相关度,便于信息自动索引和提取;能用于在 DNA 序列分析中识别基因等.NMF 最成功的一类应用是在图像处理和分析领域(Guillamet *et al*., 2001, 2003; Feng *et al*., 2002).由于 NMF 在众多领域有较好的应用效果,本文将其引入到水系沉积物地球化学数据处理中.

1 非负矩阵分解理论

NMF 的基本思想(Lee and Seung, 1999)为任意给定一个非负的 $m\times n$ 矩阵 $\boldsymbol{V}$,需将其分解为左右两个非负矩阵的乘积,即要找出非负的 $m\times r$ 矩阵 $\boldsymbol{W}$ 和非负的 $r\times n$ 矩阵 $\boldsymbol{H}$(通常 $r(m+n)<mn$,即 $r<mn/(m+n)$,使得 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的总数据量比 $\boldsymbol{V}$ 小),从而满足:

$$\boldsymbol{V}\approx\boldsymbol{WH}. \tag{1}$$

由于分解前后的矩阵中仅包含非负元素,因此原矩阵 $\boldsymbol{V}$ 中的任一列向量可以解释为左矩阵 $\boldsymbol{W}$ 中所有 r 个列向量(称为基向量)的加权和,而权重系数为右矩阵 $\boldsymbol{H}$ 中所对应列向量的元素.这说明 $\boldsymbol{W}$ 组成了一个基矩阵(基向量组),可以通过它的列向量的线性组合来近似表示 $\boldsymbol{V}$,这相当于用相对较少的基向量来表示大量的数据向量,因此只有在基向量覆盖了 $\boldsymbol{V}$ 中隐含的数据结构时,才能获得令人满意的近似表达效果(Donoho and Stodden, 2004).非负性是对矩阵分解非常有效的约束条件,这一约束导致了对原始数据基于部分的表示,NMF 算法所得到的非负基向量组 $\boldsymbol{W}$ 具有一定的聚类性和稀疏性,从而对原始数据的特征及结构具有较强的表达能力.这种基于基向量组合的表示形式具有很直观的语义解释,它反映了人类思维中"局部构成整体"的概念(Lee and Seung, 1999).有关非负矩阵分解的算法请参阅 Lee and Seung(1999, 2000).

2 NMF 在水系沉积物地球化学数据处理中的应用

化探数据,特别是水系沉积物地球化学数据是地质找矿的重要数据源之一,通过处理化探数据获取地球化学异常是化探数据处理的目标之一.化探异常包括单元素异常和多元素组合异常.传统上获取化探元素组合异常的常用方法是主成分分析(PCA)方法.PCA 方法在方差极大和正交约束的条件下寻求化探元素组合,以此为基础提取化探元素组合异常.而 PCA 方法仅能表达数据的全局特征(Lee and Seung, 1999).由于受不同地质环境的影响,化探数据在不同的研究区域内会具有不同的组合特征,表现为局部与全局特征上的差异,这些局部特征往往与成矿或矿床分布有关.NMF 方法的基向量却具有表达数据局部结构特征的特点,这可能为化探数据处理提供了一种有效的途径.此外,化探数据都是非负的,也正好满足 NMF 方法对数据非负性的基本要求.

假设有 n 个化探样品,每个样品分析了 k 个地

球化学元素含量, 往往是含量比例如常量元素百分含量(%)或微量元素含量(ppb 或 ppm), 这些含量均为非负数值, 可以组成一个 $n\times k$ 阶的化探数据非负矩阵 $\boldsymbol{V}$, 每一列 $\boldsymbol{V}_i$ 代表一个元素, 每一行代表一个样品.

根据 NMF 理论, 可以将 k 个化探元素组成的数据矩阵 $\boldsymbol{V}$ 分解为以下的形式:

$$\boldsymbol{V}_{mn\times k}\approx\boldsymbol{W}_{mn\times r}\boldsymbol{H}_{r\times k}, \tag{2}$$

其中, $\boldsymbol{W}$ 为基向量矩阵(或向量组), $\boldsymbol{H}$ 为权值矩阵, r 为基向量的个数. 这里要求 $\boldsymbol{V}$、$\boldsymbol{W}$ 和 $\boldsymbol{H}$ 都是非负的.

由式(2)可知, 数据矩阵 $\boldsymbol{V}$ 的列向量可表示为左矩阵 $\boldsymbol{W}$ 的列向量的线性组合, 加权系数即为右矩阵 $\boldsymbol{H}$ 的对应列向量中的元素. 基向量组 $\boldsymbol{W}$ 的每个列向量都从不同角度反映了地球化学元素组合信息. 因此, 非负矩阵分解的基向量可提供原始地球化学元素组合的基本信息. 下面的应用实例中将会对这些基向量的意义进行研究.

3　实例

研究区为云南个旧锡铜多金属矿区, 总面积约 3 108 km^2, 在研究区范围内共有 813 个 1∶20 万水系沉积物地球化学样采样点, 等间距采样, 每 2 km×2 km 为一个数据点, 本文中用到了 As、Cd、Cu、Pb、Zn 和 Sn 共 6 种元素. 有关研究区内的地质概况和物探、化探以及遥感影像特征参见成秋明等(2009).

表 1　r=1, 2, 3 时 NMF 权值矩阵

Table 1　Encodings using NMF when r=1, 2, 3

r=3	0.058 1	0.989 7	0.068 1	0.026 5	0.059 9	0.090 3
	0	0.016 8	0	0.983 7	0.040 7	0.174 4
	0.494 0	0	0.307 9	0	0.261 8	0.769 8
r=2	0.042 3	0.993 0	0.058 4	0.041 9	0.052 8	0.064 7
	0.113 7	0.020 4	0.022 8	0.903 2	0.112 4	0.397 1
r=1	0.072 5	0.928 3	0.060 9	0.302 6	0.081 9	0.176 0

3.1　非负矩阵分解方法应用

在本例中选取 As、Cd、Cu、Pb、Sn 和 Zn 6 个地球化学元素的 813 个水系沉积物样品组成 813×6 的非负矩阵 $\boldsymbol{V}$, 运用 MATLAB R2008a 的非负矩阵分解函数 NMF 分别取 r=1, 2, 3 对 $\boldsymbol{V}$ 进行非负分解. 分解后的加权矩阵如表 1, 基向量之间的相关系数如表 2. 对每个基向量进行反距离插值后生成栅格图层并对其取对数, 如图 1 所示. 图 1a、1b 和 1c 分别表示 r=1, 2, 3 所得到的第一个基向量的结果(为了表达方便, 图中对结果取了自然对数). 从这 3 个图的结果可以看出, 它们的空间分布形态基本上一致, 从表 2 也可以看出它们之间的相关系数分别高达 0.996 9, 0.997 和 0.999 9; 图 1d 和图 1e 分别是 r 为 2 和 3 时的第二个基向量取对数. 从图中可以看出, 两个基向量形态非常相似, 它们之间的相关系数为 0.992 7; 图 1f 是 r 为 3 时的第 3 个基向量取对数. 可以看出, 虽然当基向量的个数不同时, 对应的基向量(比如第一基向量)会有所差异, 但总体来说变化不大, 具有相对稳定性. 为了研究每个基向量与各元素之间的关系, 我们计算了基向量与原始元

表 2　NMF 基向量和 PCA 的 3 个主成分之间的相关系数

Table 2　Correlations among six basis vectors using NMF and three principal components using PCA

	As	Cd	Cu	Pb	Sn	Zn	W1	W21	W22	W31	W32	W33
Cd	0.642											
Cu	0.692	0.711										
Pb	0.617	0.831	0.588									
Sn	0.747	0.731	0.748	0.713								
Zn	0.614	0.810	0.618	0.842	0.630							
W1	0.660	0.997	0.717	0.868	0.749	0.839						
W21	0.645	1.000	0.716	0.827	0.734	0.810	0.997					
W22	0.639	0.795	0.590	0.991	0.713	0.883	0.838	0.792				
W31	0.642	1.000	0.713	0.829	0.732	0.809	0.997	1.000	0.793			
W32	0.592	0.806	0.562	0.999	0.693	0.841	0.846	0.803	0.993	0.805		
W33	0.697	0.349	0.577	0.476	0.548	0.729	0.391	0.353	0.579	0.349	0.479	
PCA1	0.824	0.911	0.834	0.886	0.877	0.870	0.931	0.913	0.889	0.911	0.867	0.644
PCA2	−0.36	0.199	−0.35	0.351	−0.26	0.368	0.216	0.191	0.345	0.196	0.380	−0.14
PCA3	0.379	−0.15	−0.39	0.099	0.038	0.023	−0.11	−0.15	0.139	−0.15	0.111	0.234

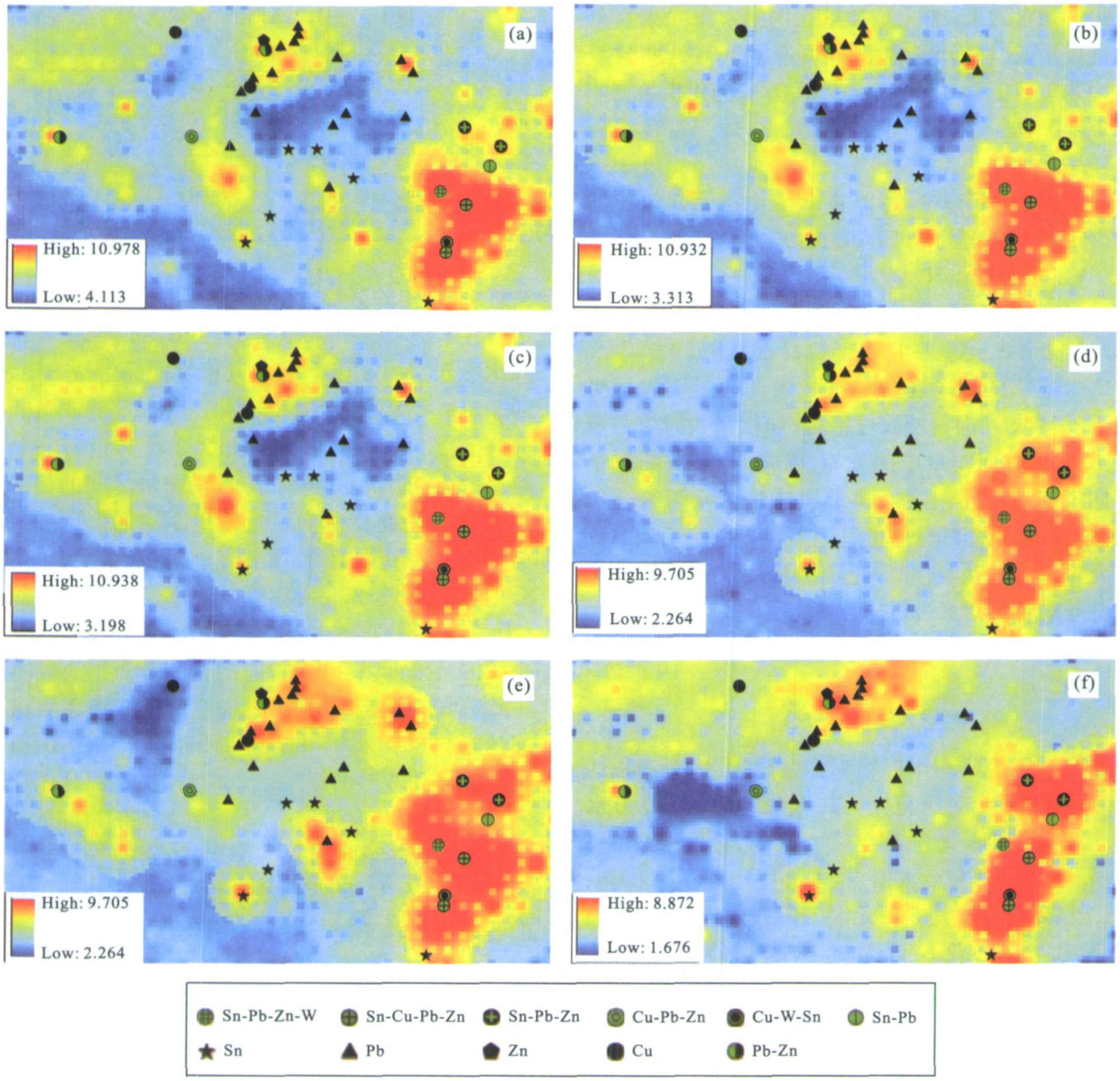

图 1 NMF 方法 $r=1, 2, 3$ 时 6 个基向量分别取对数后的 6 个图层

Fig. 1 Six basis vectors obtained using NMF when $r=1, 2, 3$

a. $r=1$ 时的第一个基向量取对数; b. $r=2$ 时的第一个基向量取对数; c. $r=3$ 时的第一个基向量取对数; d. $r=2$ 时的第二个基向量取对数; e. $r=3$ 时的第二个基向量取对数; f. $r=3$ 时的第 3 个基向量取对数

素向量之间的相关性, 当 $r=3$ 时, 从表 1 可以看出, Cd 主要由第一基向量决定, Pb 主要由第二基向量表示, 但 Cd 与 Cu、Pb、Sn、Zn 的相关系数至少大于 0.71, 所以第一基向量间接反映了元素 Cu、Pb、Sn 和 Zn 元素组合, Pb 与 Sn 和 Zn 的相关系数分别为 0.71 和 0.84, 间接反映了元素 Sn 和 Zn. 而 As、Cu、Sn 和 Zn 与第三基向量关系比较大.

3.2 主成分分析(PCA)方法应用

为了与非负矩阵分解方法进行比较, 同样选取上述 6 个元素的数据进行主成分分析, 前 3 个主成分(PCA 1, PCA 2, PCA 3)的因子载荷见图 2, 由于这 3 个主成分的因子得分包含负值, 不能直接取对数, 第一主成分的因子得分加 0.7 取对数后见图 3a, 将第二、三主成分的得分分别加 7 后取对数如图 3b 和 3c. 从图 3 和图 2 可以看出第一主成分的因子载荷都为正的, 反映 6 个元素的综合效应; 第二主成分因子载荷正值是 As-Cu-Sn 组合, 负值是 Zn-Pb-Cd 组合; 第三主成分因子载荷的正值是 Cu-Cd-Pb 组合, 负值是 As-Sn-Zn 组合.

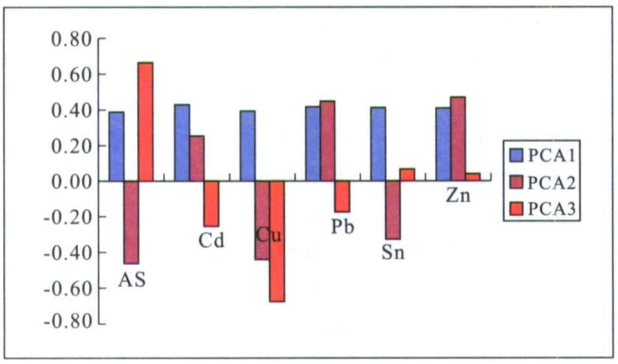

图 2　主成分分析的 3 个主成分的因子载荷

Fig. 2　Factor loading on three principal components

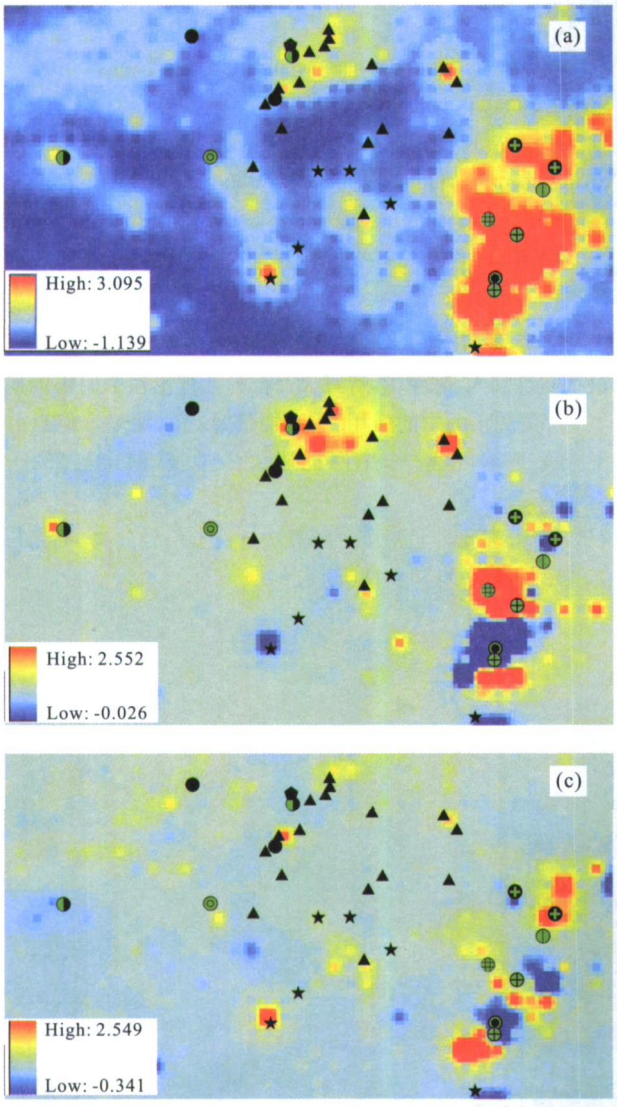

图 3　主成分分析得到的 3 个主成分的因子得分

Fig. 3　Scores of three principal components

a. log(PCA1+0.7); b. log(PCA2+7); c. log(PCA3+7)

3.3　非负矩阵分解与主成分分析比较

(1)NMF 和 PCA 都是矩阵分解方法, 都服从等式(1)分解模型, 主要差别在于约束条件不同, NMF 模型的约束条件是 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 都是非负矩阵, 而对于 PCA 方法不要求 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 是非负矩阵, 但 $\boldsymbol{W}$ 是列正交的, $\boldsymbol{H}$ 是行正交的.

(2)Lee and Seung(1999)通过对人脸图像进行分析认为, NMF 方法的基向量具有稀疏特征, $\boldsymbol{H}$ 含有较多的 0 元素, 这一结论从本例的结果也得到了验证, 特别是基向量具有大量的重复取值. 这一特征说明 NMF 具有较好的反映局部的特性, 而 PCA 方法仅能反映数据的全局特征.

(3)这里只对 PCA 的 3 个主成分和 $r=3$ 时的 3 个基向量进行比较. 从图 1c 和图 3a 可以看出 PCA1 和 NMF 第一基向量都较好地反映了研究区内的多金属矿产相关的化探组合元素异常分布, 二者的相关系数为 0.911, 具有较高的相关性; 从图 3b 可以看出, PCA2 较好地反映了 Pb 和 Zn 相关异常, 但对于包含 Cu-Pb-Zn 或者 Sn-Pb-Zn 反映较差. 从图 1e 看出 NMF 第二个基向量对多种矿产都有较好的反映, 并且其分布具有较好聚类性; 从图 3c 可以看出, PCA3 仅对少数几个与 Cu 有关的多金属矿产反映较好(负值), 对大部分矿产反映较差. 从图 1F 看出 NMF 第 3 个基向量对多金属矿产和 Pb 矿产反映较好. 就本实例而言, 通过上述比较可以看出 NMF 能更好地反映多金属矿产的分布特征, 优于主成分分析方法.

4　结论

非负矩阵分解方法为水系沉积物地球化学数据处理提供了一种较好的方法; 就本实例而言, 非负矩阵分解方法优于传统的主成分分析方法; 与主成分分析方法仅能反映数据的全局特征相比, 非负矩阵分解方法能较好地反映多元素分布的局部特征. 尽管本实例是针对水系沉积物地球化学数据而介绍的, 但该方法对于其他数据处理同样具有参考意义, 也可用于其他的地球化学数据处理, 建议在其他地球化学数据处理中进行试验.

References

Cheng, Q. M., Zhao, P. D., Chen, J. G., et al., 2009. Application of singularity theory in prediction of tin and copper mineral deposits in Gejiu district, Yunnan, China: Weak information extraction and mixing information decomposition. *Earth Science—Journal of China University*

of Geosciences, 34(2): 232—242 (in Chinese with English abstract).

Donoho, D., Stodden, V., 2004. When does non-negative matrix factorization give a correct decomposition into parts? In: Thrun, S., Saul, L., Scholkopf, B., eds., Advances in neural information processing systems 16. MIT Press, Cambridge, MA.

Feng, T., Li, S., Shum, H., et al., 2002. Local non-negative matrix factorization as a visual representation. In: Proceedings of the 2nd international conference on development and learning. IEEE, Cambridge. U. K., 178—183. DOI: 10. 1109/ DEVLRN. 2002. 1011835.

Guillamet, D., Bressan, M., Vitria, J., 2001. A weighted non-negative matrix factorization for local representations. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition V1, Kauai, HI, 942— 947. DOI: 10. 1109/CVPR. 2001. 990629.

Guillamet, D., Vitria, J., Schiele, B., 2003. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognition Letters*, 24(14): 2447—2454.

Juvela, M., Lehtinen, K., Paatero, P., 1994. The use of positive matrix factorization in the analysis of molecular line spectra from the thumbprint nebula. In: Clemens, D. P., Barvainis, R., eds., Clouds, cores, and low mass star. *ASP Conference Series*, 65: 176—180.

Juvela, M., Lehtinen, K., Paatero, P., 1996. The use of positive matrix factorization in the analysis of molecular line spectra. *Mon. Not. R. Astron. Soc.*, 280: 616—626.

Lee, D. O., Seung, H. S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401: 788—791.

Lee, D., Seung, H., 2000. Algorithms for non-negative matrix factorization. In: Leen, T., Dietterich, T., Tresp, V., eds., Advances in neural information processing systems. MIT Press, Cambridge, MA, 556—562.

Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1): 23—35.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2): 111—126.

Wei, L., 2004. Blind sources separation based on non-negative matrix factorization. *Electronics Optics & Control*, 11(2): 38—41, 53 (in Chinese with English abstract).

Xu, W., Liu, X., Gong, Y., 2003. Document-clustering based on non-negative matrix factorization. In: Proceedings of SIGIR' 03, July 28-August 1. 267—273, Toronto, CA.

附中文参考文献

成秋明, 赵鹏大, 陈建国, 等, 2009. 奇异性理论在个旧锡铜矿产资源预测中的应用: 成矿弱信息提取和复合信息分解. 地球科学——中国地质大学学报, 34(2): 232—242.

魏乐, 2004. 基于非负矩阵分解算法进行盲信号分离. 电光与控制, 11(2): 38—41, 53.