维普资讯 http://www.cqvip.com

文章编号：1001—1749(2008)01—0078—03

R－Q型因子分析与对应分析

曾道明[1,2]，纪宏金[1]，高 文[1]，刘 强[1]
（1.吉林大学 地球探测科学与技术学院，长春 130026；
2.中国地质科学院 地球物理地球化学勘查研究所，廊坊 065000）

摘 要：R－Q型因子分析是一种多元统计方法，但也是一种被忽视的方法。这里以简单方式证明了R－Q型因子分析对偶性，并给出一种方便的R－Q型因子得分计算公式。利用地球化学数据解释的实例，对比了该方法与对应分析的应用效果。结果表明，二种方法都能有效地识别数据中的样品与变量类型及其对应关系，并且二种方法都具有独特的优点。这些方法为快速、有效地解释勘查地球化学数据起到重要作用。

关键词：多元统计分析；因子分析；对应分析；地球化学勘查；数据解释
中图分类号：P 632 **文献标识码**：A

0 前言

法国数学家 Benzécri[1] 于上个世纪七十年代提出的对应分析，巧妙地给出了R型因子分析和Q型因子分析的对偶关系，使得变量点和样品点可投影于同一因子空间，且样品点的类型能方便地被邻近的变量点来解释。这种思路吸引了数学家们的长期关注[2,3]。

对应分析的出现，推动了数字分类法的发展与应用[4-8]。我们知道，各种统计模式识别方法的实质是数字分类，地球化学数据解释的核心也是分类，只有根据样品类型和变量类型的对应关系，才能更有效、快速地识别样品类型[9]。

实际上，对应分析仅属于R－Q型因子分析的一种。周蒂[10]曾提出另一种R－Q型因子分析的方法，并且详细介绍了该方法的思路和功能特点，认为该方法得出的结果更合理。为叙述方便，这里将前者仍称为对应分析，而后者称为R－Q型因子分析。长期以来，R－Q型因子分析很少被讨论。

为进一步探讨R－Q型因子分析，作者以简单形式证明了该方法的对偶性，给出一种方便的R－Q型因子得分计算公式，并以 Howarth[11] 给出的一个数据样本为例，比较了上述二种方法在勘查地球化学数据解释中的应用效果。

1 R－Q型因子分析

设 $\boldsymbol{X}=(x_{ij})_{n\times m}$ 为 n 行 m 列的标准化数据矩阵，其中 x_{ij} 为样品 i 在变量 j 的取值。若令

$$\boldsymbol{W}=n^{-1/2}\boldsymbol{X} \text{ 或 } \boldsymbol{X}=n^{1/2}\boldsymbol{W} \qquad (1)$$

则变量间的相关系数矩阵可表示为 $\boldsymbol{R}=\boldsymbol{W}^T\boldsymbol{W}$，样品间的相似矩阵可表示为 $\boldsymbol{H}=\boldsymbol{W}\boldsymbol{W}^T$，其中 $\boldsymbol{W}^T$ 为 $\boldsymbol{W}$ 的转置矩阵，$\boldsymbol{H}$ 确定了 n 个样品间的欧氏距离[10]。于是，从矩阵 $\boldsymbol{R}$ 出发可作R型分析，从矩阵 $\boldsymbol{H}$ 出发可作Q型分析。

R型和Q型分析的对偶性是R－Q型因子分析的基础。设矩阵 $\boldsymbol{R}$ 的秩为 r，其 r 个非零特征值为 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r>0$，将该特征值和相应的单位特征向量以矩阵形式分别记为

$$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_r) \qquad (2)$$

$$\boldsymbol{U}=(u_1,u_2,\cdots,u_r)=(u_{ij})_{m\times r} \qquad (3)$$

其中 $\boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{I}$，$\boldsymbol{I}$ 为 r 阶单位矩阵。若将R型因子负载矩阵 $\boldsymbol{A}_R$ 和因子得分矩阵 $\boldsymbol{F}_R$ 分别定义为

基金项目：中国地质调查局地质调查项目(200110200021)
收稿日期：2006－11－08 改回日期：2007－06－22

维普资讯 http://www.cqvip.com

$$A_R = U\Lambda^{1/2} \quad (4)$$

$$F_R = WU \quad (5)$$

则因子负载、因子得分有相同的方差贡献,满足

$$A_R{}^T A_R = \Lambda^{1/2} U^T U \Lambda^{1/2} = \Lambda \quad (6)$$

$$F_R{}^T F_R = U^T W^T W U = U^T U \Lambda U^T U = \Lambda \quad (7)$$

我们知道,矩阵 H 与矩阵 R 有相同的特征值,其相应的单位特征向量为

$$V = (v_1, v_2, \cdots, v_r) = WU\Lambda^{-1/2} = (v_{ij})_{n\times r} \quad (8)$$

满足 $V^TV = I$。若将 Q 型因子负载矩阵 A_Q 和因子得分矩阵 F_Q 分别定义为

$$A_Q = V\Lambda^{1/2} \quad (9)$$

$$F_Q = W^TV \quad (10)$$

则 R 型分析和 Q 形分析的结果之间具有以下对偶关系

$$A_Q = V\Lambda^{1/2} = WU\Lambda^{-1/2}\Lambda^{1/2} = WU = F_R \quad (11)$$

$$F_Q = W^TV = W^T WU\Lambda^{-1/2} = U\Lambda U^T U\Lambda^{-1/2} = U\Lambda^{1/2} = A_R \quad (12)$$

显然只作 R 型分析就可同时得到 Q 型分析结果。

需要说明的是,式(5)和式(10)中,F_R 及 F_Q 的定义形式是作者给出的。若采用原有定义[10],则 $F_R = V$, $F_Q = U$,这时就无法得到式(11)和式(12)表现的对偶关系,也不能得到式(6)和式(7)表现的性质。

我们知道,A_R 中的 m 行 p 列代表 m 个变量点在 p 维因子空间的坐标,F_R 中的 n 行 p 列则代表 n 个样品点在同一 p 维因子空间的坐标。由于

$$A_RA_R{}^T = U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = W^TW = R \quad (13)$$

$$F_R F_R{}^T = V\Lambda^{1/2}\Lambda^{1/2}V^T = V\Lambda V^T = WW^T = H \quad (14)$$

说明 m 个变量点间的关系完全反映在 $A_RA_R{}^T$ 中,n 个样品点的关系完全反映在 $F_RF_R{}^T$ 中。由式(6)和式(7)知,变量点与样品点在因子空间中,具有相同的离散程度。于是,通常在 A_R 和 F_R 中任取二列制作二维散点图,用以表达变量与样品间的亲疏关系。

实际上,由常用的 R 型因子分析,可以方便地实现 R－Q 型因子分析。显然,A_R 的计算在二种方法中是相同的,F_R 也可由普通 R 型因子得分 $F = XA_R\Lambda^{-1}$ 很快求出,因为结合式(1)、式(4)和式(5)可得 $F = n^{1/2}WU\Lambda^{1/2}\Lambda^{-1} = n^{1/2}F_R\Lambda^{-1/2}$,或

$$F_R = n^{-1/2}F\Lambda^{1/2} \quad (15)$$

2　R－Q 型因子分析与对应分析的实例

现在给出一个实例,以便对比 R－Q 型因子分析与对应分析的效果。Howarth 曾给出了 25 个水系沉积物样品中 Zn、Fe、Mn、Cd、Cu 和 Pb 的分析数据[11]。以 25×6 的数据矩阵为样本(采用自编的、经多年检验可靠的 FORTRAN 程序)作 R－Q 型因子分析时,求得其变量间相关矩阵 R 的六个特征值分别为 $\lambda_1 = 3.393$, $\lambda_2 = 2.114$, $\lambda_3 = 0.285$, $\lambda_4 = 0.171$, $\lambda_5 = 0.027$, $\lambda_6 = 0.010$。由于前二个公因子的方差贡献约为 92%,它们构成的二维空间,能较充分地表现出样品点与变量点间的结构关系,如图 1 所示。该结果具有以下特征:

(1)识别了样品的类型。全部 25 个样品被分为三类,即 13 号、15 号和 16 号为一类,22 号和 23 号为一类,其余 20 个样品为另一类。

(2)识别了变量的类型。六个变量点被分为二类,即 Cu、Zn 和 Cd 为一类,明显表现为亲硫元素组合;Fe、Mn 和 Pb 为另一类,主要是亲铁元素组合。

(3)给出了样品点与变量点的对偶关系。结合 Howarth 的介绍[11]可知,这些对偶关系具有一定的实际意义。样品 13 号、15 号、16 号与元素 Cu、Zn、Cd 对偶成一类,这些样品所在的小溪上游,正好存在一个铜锌矿点,因此该类代表了以 Cu、Zn 和 Cd 为标型元素的矿异常类型。22 号和 23 号样品与 Fe、Mn 和 Pb 对偶成另一类,由于样品中 Fe、Mn 和 Pb 的含量较高,应属于由次生富集作用引起的"非矿异常"。其它 20 个样品则没有显示与变量间的对偶关系,它们的基本特征是,各元素含量普遍较低,因而可被认为是"背景"类样品。

利用同一原始数据作对应分析时,五个有意义的特征值分别为 $\lambda_1 = 0.2643$, $\lambda_2 = 0.0334$, $\lambda_3 = 0.0032$, $\lambda_4 = 0.0015$, $\lambda_5 = 0.0004$。其中,前二

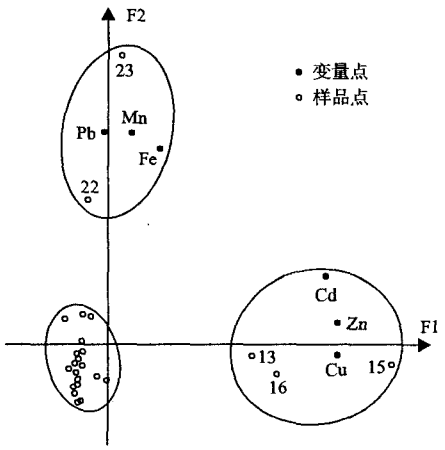


图 1　R－Q 型因子分析的结果

Fig. 1　The result of R－Q mode factor analysis

维普资讯 http://www.cqvip.com

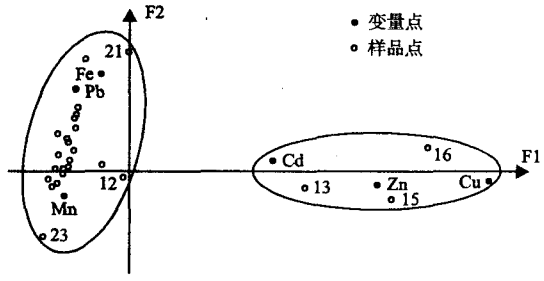


图 2 对应分析的结果

Fig. 2 The result of correspondence analysis

个公因子的方差贡献约为 98%，可充分反映样品点与变量点间的结构关系。图 2 是前二个因子空间的散点图，它将全部变量点和样品点都分为二种类型，并显示了二者的对应关系：①由 13 号、15 号、16 号样品及元素 Cu、Zn 和 Cd 组成，如上所述，它具有成矿意义。②由其它 22 个样品与元素 Fe、Mn 和 Pb 组成，属于非矿类型特征。

3 讨论与结论

首先讨论 R－Q 型因子分析的实用价值。由上例可知，R－Q 型因子分析以简单的图件将全部样品和变量分为三个类型，而根据实际资料，这些类型分别属于矿异常、非矿异常和背景。由此，R－Q型因子分析结果的可解释性和有效性得到证明。显然，该结果中包含了化探数据解释所需的核心内容，但它比常用的解释方法更加简便、快捷。

再讨论 R－Q 型因子分析与对应分析差别。在上例的对应分析中，样品和元素被分为矿与非矿二类，并且同样具有实际意义。与 R－Q 型因子分析结果不同，对应分析没有识别出非矿异常样品和背景样品间的差异，而是将它们归为同一类型。

这二种方法的上述差别，可归于它们所采用的相似性度量不同[10]。R－Q 型因子分析中，对样品采用的是一种对绝对值敏感的距离度量，其意义显然。对应分析采用的是比例相似性，其分类结果对数据量值大小不敏感。根据 Imbrie 的 Q 型因子分析理论[12]，样品间比例相似性度量也是有一定地质依据的。在上例中，由于非矿异常和背景样品中的元素含量比例相似，故被对应分析判为同一类型，但由于这二类样品中的元素绝对含量不同而被 R－Q 型因子分析识别出来。

综上所述，可得出以下结论：

(1)作者定义的 R 型和 Q 型得分，便于 R－Q 型因子分析对偶性的证明，也便于由普通 R 型因子分析实现 R－Q 型因子分析。

(2)对应分析和 R－Q 型因子分析，都具有一定研究与应用价值。前者在快速识别矿与非矿类型时表现了优势；后者则从非矿类样品中进一步识别出了背景与非矿异常。

(3)在地球化学数据解释中，已知地质资料的引入有利于二种方法结果的合理解释。

参考文献：

[1] DAVID M, CAMPIGLIO C, DRLINGling R. Progress in R－ and Q－mode analysis: correspondence analysis and its application to the study of geological processes [J]. Can. J. Earth Sci., 1974, 11:131.

[2] TAO FENGMEI, XIA LIAXIAN. Stepwise correspondence analysis and study of physical characteristics of Chinese ethnic groups [J]. Journal of Biomathematics, 2003, 18(2):139.

[3] GROENEN P J F, VELDEN M. Inverse correspondence analysis[J]. Liner Algebra and Its Applications, 2004, 388:221.

[4] 纪宏金，仲崇学．区域化探中若干新图件的计算与制作[J]．物探化探计算技术，1991，13(2)：98.

[5] JI HONGJIN, ZHU YONGZHENG, WU XISHENG. Correspondence Cluster Analysis and Its Application in Exploration Geochemistry [J]. Geochem. Explor., 1995, (55):137.

[6] 洪方明，贾志永，龙志和．对应聚类分析与政府投资领域的确定[J]．西南民族学院学报(哲学社会科学版)，1998，19(2)：132.

[7] 贾志永，洪方明．对应聚类分析在铁路建设投融资决策中的应用[J]．西南民族学院学报(哲学社会科学版)2000，21(1)：11.

[8] 纪宏金，于元峰，时艳香，等．对应聚类分析在元素医学研究中的应用[J]．中国卫生统计，2004，21(4)：248.

[9] 纪宏金，孙丰月，陈满，等．胶东地区裸露含金构造的地球化学评价[J]．吉林大学学报(地球科学版)，2005，35(3)：308.

[10] 周蒂．对应分析的局限性及 RQ 式因子分析[J]．物探化探计算技术，1985，7(4)：332.

[11] HOWARTH R. J., SINDING-LARSEN R. Multivariate analysis. In: R. J. Howarth (Ed), Statistics and data analysis in geochemical prospecting [M]. Handbook of Exploration Geochemistry, Vol. 2. Elsevier, Amsterdam, 1983.

[12] JORESKOG K G, KLOVAN J E, REYMENT R A. Methods in Geomathematics 1, Geological Factor Analysis[M]. Elsevier, Amsterdam, 1976.

作者简介：曾道明(1982－)，男，四川德阳人，硕士，主要致力于应用地球化学研究。