因子分析和分量分析

——地质学中的多变量統计方法之一

地质科学研究院地质矿产所方法组

目　　录

一、引　　言

二、因子分析与分量分析模型

三、主因子解

四、方差最大正交旋转

五、Promax斜旋转

六、因子的估计（或因子得分的计算）

七、地质实例

附录：程序

1957年，Krumbein将一种心理学的研究方法——因子分析引进沉积学，后来经过Imbrie的大力倡导，获得了迅速的进展，目前已遍及到各个地质学科，成为地质人员最常用的統计方法之一。

地质工作的基本问题之一是对大量的地质观测資料进行解释，从而建立一个有机的成因系統，以此作为找矿勘探工作的理论依据。过去，这种解释工作和成因探讨完全是由人的思考来进行的。但是，由于每个地质人员的专长、经历和思维方式不同，往往造成地质解释和成因观点的很大出入和争论。产生这种矛盾的一个重要原因，是人的思考无法掌握大量数据间的错综复杂的关系，因而每个人只好从不同的角度对这些数据加以删简和強调，这就給成因分析带来了不同程度的片面性。

因子分析，正是一种帮助我们对大量地质观测数据进行分析和解释的統计方法。它将许多彼此间具有错综复杂关系的地质现象归结为数量较少的几个因子，每一个因子意味着地质变量间的一种基本结合关系，它往往指示出某种地质上的成因联系。对这些因子进行地质解释就构成了因子分析的主要内容。

因子分析在地质上一般用于以下几个目的：

一、对观测到的大量地质现象进行综合和归纳，剔除原始地质观察中的繁琐和重复成分，用更簡炼的形式来描述地质对象；特别是通过因子分析，能将非正交的原始变量空间转换为一个正交的因子空间，这对进行其他統计分析是很有利的。

二、对因子进行解释，可以探索各种地质现象的成因联系。例如。在沉积学中，因子往往意味着物质来源、水动力条件、生物生活环境等；在研究内生成矿作用时，因子可能具有岩浆活动阶段和矿化阶段的含义；在研究岩矿的物质组分时，因子可以指示元素的共生组合关系，为探索元素赋存状态提供依据。

因子分析的效果与地质研究的水平有密切的关系。地质工作愈广泛、深入，对因子的解释就愈正确合理，因子分析就能收取较大的效果。如果仅根据片断的材料和主观愿望硬加解释，则可能导致十分荒谬的结论。

本方法是在许多兄弟单位的热情支持下编写完成的，我所二室超基性岩及铂矿组的同志給了许多帮助，提供了计算实例数据，并用他们的研究成果纠正了一些我们在因子解释上的错误认识。青海地质二队的同志给我们的工作給予了很大的赞助，提供了许多进行因子分析的实际資料。计委地质总局一五〇工程指挥部、地质科学院水文所、北京工业大学计算站均提供了相当多的计算时间，使我们能在较短期间通过一套分量分析的ＤＪＳ－６机计算机程序，并试算了一定数量的问题。在此一并表示深切的感谢。

一、引　　言

由C. Spearman提出的因子分析（Factor Analgsis）是用来研究一组变量之间的相关性的，或者说，是用来研究相关矩阵（或协方差矩阵）内部结构的。为了研究相关矩阵（或协方差矩阵）内部结构，我们还可以利用由K. Pearson提出的分量分析。

无论是因子分析或者是分量分析，它们都是用若干新的变量，这些新的变量是原始变量在通常的或者是最小二乘法意义上的线性组合，称为因子或分量，对包含在原始变量之间的相关性进行合理地解释，并用这些因子或者分量，它们的数目往往比原始变量的数目要少，取代原始变量，而总的关于原始变量的相关信息却损失无几。因此，特别适用于复杂的地质学。正因为这样，自从一九六二年J. Imbrie和E. G. Purdy把它们在地质学中加以发展之后，很快地就成为在地质学中传播最快，应用最广的多变量统计方法之一。

因子分析和分量分析因为都可以用来研究相关矩阵（或协方差矩阵）的内部结构，又具有某种程度的类似之处，所以在实际应用的许多场合之下，都不加区分地、笼统地把它们称为因子分析。为了区分狭义的《因子分析》与广义的《因子分析》，我们将在广义的《因子分析》四个字底下加一横线。关于因子分析和分量分析的不同之处，因为本文主要地是取其共同的，所以就不打算作为主要内容之一详细介绍了。在这方面可以参阅〔3〕。但是，在本文叙述的内容范围之内，适当的地方也略微作些分析。应当指出的是，由于在因子分析中，个

别关键性的问题尚未成功地解决，所以在实际上使用的大部分是分量分析。

正如前面指明的，因子分析可以用来研究一组变量之间的相关性，然而，J。Imbrie 等人却借助因子分析研究一组样品之间的相互关系，用获得的因子解释这样关系，并在此基础上对样品进行分类。为了区别这两者，就把前者称为 R 型因子分析，后者称为 Q 型因子分析。

因子分析的内容较多，主要内容有：因子分析模型（解析的和几何的）；主因子解或主分量解，以后統称为主因子解，它是导出其它一切解的基础；因子轴的旋转（正交的和斜的），它是导出其它解的手段；因子的估计，即因子得分的计算等等。以下将对上述内容分节加以介绍。然后給出地质上应用的一个例子。作为本文的附录，将給出分量分析的一个用 D J S－6 算法语言编写的计算机程序。

二、因子分析与分量分析模型

首先介绍因子分析模型，分量分析模型在某种意义上可以作为因子分析模型的一个特殊情况。不过，应当指出，正如〔3〕中所述的，因子分析和分量分析绝非纯粹是一般与特殊的关系。

假设原始变量 X_1，X_2，…，X_p。由于它们既有相关的一面，又有独立的一面，所以可以把各个变量所包含的統计信息分为两部分。由于各个变量的統计信息来源于方差，所以，也就意味着可以把方差分为两部分（见第三节）。一部分反映变量之间的相关；另一部分仅与变量本身的变化有关。因子分析的基本任务就在于把这两部分信息

分离开来。从这个意义上说，因子分析似乎过滤器，对包含在变量 X_1，X_2，…，X_p 中的统计信息起着过滤作用。

在因子分析中，分别用一些新的变量，即所谓的公共因子和单一因子表示上述的两种信息。这些公共因子是互不相关的，所以从数目上通常要比相关的变量 X_1，X_2，…，X_p 少。

把以上所说概括起来，就构成因子分析的基本假设：相关变量 X_1，X_2，…，X_p 可以用较少数目的互不相关的公共因子 F_1，F_2，…，F_k $(k<p)$，以及既与公共因子不相关又本身之间互不相关的单一因子 U_1，U_2，…，U_p 表示。这个假设可以用数学的公式确切地表达如下：

$$
\begin{aligned}
X_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1k}F_k + u_1U_1 \\
X_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2k}F_k + u_2U_2 \\
&- - - - - - - - - - - \\
X_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pk}F_k + u_pU_p
\end{aligned}
\qquad (2\cdot1)
$$

在 $(2\cdot1)$ 中可以看出，每一个公共因子都与所有的变量 X_i $(i=1, 2, \cdots, p)$ 有关，相反地，每一个单一因子仅与相对应的一个变量有关。把它们写成矩阵的形式，那就是：

$$X = AF + uU \qquad (2\cdot2)$$

其中

$$X = (X_1, X_2, \cdots, X_p)', \quad F = (F_1, F_2, \cdots, F_k)'$$
$$U = (U_1, U_2, \cdots, U_p),$$

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pk} \end{pmatrix}, \quad u=\begin{pmatrix} u_1 & 0 & \cdots & 0 \\ 0 & u_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_p \end{pmatrix},$$

此地，“$'$”表示转置。在（2·1）中将规定诸公共因子和单一因子的数学期望为零，方差为1。并且将假设原始变量的数学期望为零，方差为1。否则，只要把它们标准化，即令

$$X_i^*=\frac{X_i-MX_i}{\sigma X_i} \quad (i=1,2,\cdots,p),$$

并用它们代替 X_i 就是。此地，MX_i 和 σX_i 分别表示 X_i 的数学期望和标准差。因为 X_i^* 与 X_i 具有相同的相关矩阵，所以，用 X_i^* 代替 X_i 将不影响因子分析的结果。公共因子的系数 a_{ij}（$i=1$，2，…，p；$j=1$，2，…，k）是待定的，称为第 i 个变量在第 j 个公共因子上的因子载荷，矩阵 A 就称为因子载荷矩阵。单一因子的系数 u_i（$i=1$，2，…，p）仅是使每一个原始变量的方差为1的补充值。因子分析的基本问题之一就是利用原始变量之间的相关决定诸因子载荷。

在转入这个问题之前，先看一下与（2·1）或者（2·2）有关的一些量的统计意义，这对于因子分析结果的解释是十分重要的。

(1) 诸因子载荷的统计意义。为此，我们考虑第 i 个变量与第 j 个公共因子的相关。为了求得它们的相关系数 $r_{X_iF_j}$，用 F_j 去乘

(2·1)中第 i 个式子的两边，于是有：

$$X_iF_j = a_{i1}F_1F_j + a_{i2}F_2F_j + \cdots + a_{ik}F_kF_j + u_iU_iF_j,$$

X_iF_j 的数学期望

$$M(X_iF_j) = a_{i1}M(F_1F_j) + \cdots + a_{ik}M(F_kF_j) + u_iM(U_iF_j). \tag{2·3}$$

在我们对公共因子和单一因子的数学期望，方差所作的规定，以及对原始变量的数学期望，方差所作的假设之下，(2·3)式中各个数学期望依次是 X_i 与 F_j，F_q 与 F_j($q=1, 2, \cdots, k$)的相关系数($q \neq j$)和方差($q=j$)。根据因子分析的基本假设，(2·3)式右端除第 j 项为 a_{ij} 之外，其余各项均为零，亦即

$$r_{X_iF_j} = M(X_iF_j) = a_{ij}$$

因此，a_{ij} 就是第 i 个变量与第 j 个公共因子的相关系数。

(2) 因子载荷矩阵 A 中各行元素的平方和

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \quad (i=1,2,\cdots,p) \tag{2.4}$$

的统计意义。根据(2·1)第 i 个式子变量 X_i 的方差

$$\begin{aligned} \sigma^2 X_i &= a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2 + u_i^2 \\ &= \sum_{j=1}^{k} a_{ij}^2 + u_i^2 = h_i^2 + u_i^2, \end{aligned} \tag{2.5}$$

即变量 X_i 的方差被分成了两部分。第一部分与(2·5)右端第一项相对应，它就是由(2·4)所给出的因子载荷矩阵 A 中第 i 行元素的平

方和，是变量 X_i 与其余变量发生相关的源泉。第二部分与 (2·5) 右端第二项相对应，是使变量 X_i 的方差为 1 的补充值，仅与变量 X_i 本身的变化有关。我们把前者称为变量 X_i 的公共因子方差，后者称为变量 X_i 的单一因子方差。

(3) 因子载荷矩阵 A 中各列元素的平方和

$$S_j = \sum_{i=1}^{p} a_{ij}^2 \quad (j=1,2,\cdots,k) \tag{2.7}$$

的统计意义。与变量 X_i 的公共因子方差 h_i^2 恰恰相反，h_i^2 是诸公共因子对同一变量 X_i 所提供的方差之总和，S_j 则是同一公共因子 F_j 对诸变量所提供的方差之总和，它是衡量公共因子相对重要性的指标，称为公共因子 F_j 的方差贡献。

对分量分析而言，其模型是

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1p}F_p \\ X_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2p}F_p \\ &\quad - - - - - - - - - - \\ X_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pp}F_p \end{aligned} \tag{2·1}^*$$

写成矩阵的形式，那就是：$X = AF$。 $(2·2)^*$

其中 F_1，F_2，…，F_p 称为分量，但是，为了与因子分析统一起见，也称它们为因子。

从形式上看，$(2·1)^*$ 与 (2·1) 不同之处仅在于：

(1) 在 $(2·1)^*$ 中，没有单一因子，所有的因子都是公共的。因此，

变量的公共因子方差就等于变量的方差。

(2) 在(2·1)* 中公共因子的数目不是少于，而是等于原始变量的数目。

容易看出，上述的与因子分析模型(2·1)有关的一些定义，概念，解释和结论，除了单一因子及其有关的部分之外，对分量分析也是一样适用。

不过，预先说明一点，正如将要在第三节中看到的那样，在分量分析中，由于P个因子的方差贡献有大有小，这就给出了忽略方差贡献小的因子，因为它们不会提供更多的统计信息，从而减少所考虑的变量数目之可能。但是，与因子分析不同，为了精确地复制变量之间的相关，却必需要求所有的P个因子。

(2·1) 和 (2·1)* 是因子分析和分量分析的解释模型，现在介绍一下它们的几何模型，以便帮助理解。特别是在第五节中讨论相关因子时，帮助则更大。

在因子分析中，可以把互不相关的，即两两间的相关系数为零，各自方差为1的k个公共因子和P个单一因子想象为(k+P)个相互垂直的，即两两间夹角的余弦为零的单位向量。以它们为坐标轴，今后，我们将称这些坐标轴为因子轴，就构成了(k+P)维空间的一个直角坐标系。于是，根据因子分析的解释模型(2·1)，变量Xi可以用在这个坐标系下以a_{i1}, a_{i2}, …, a_{ik}, 0, …, 0, ui, 0, …, 0为坐标的向量Pi表示。显然，这个向量Pi的长度就等于

变量 X_i 的方差，即等于 1，与各因子轴的夹角的余弦就等于相对应的坐标，即变量 X_i 与诸因子的相关关系数。另外，分别表示变量 X_i 和 X_j 的向量 P_i 和 P_j 的夹角之余弦，此时即为它们的内积

$a_{i1}a_{j1}+a_{i2}a_{j2}+\cdots+a_{ik}a_{jk}$ 恰好是两变量 X_i 和 X_j 的相关系数 r_{ij}。

因为我们主要关心的是公共因子，所以通常考虑原始变量在公共因子所定义的空间—— 公共因子空间中的投影。

以两维的，即两个公共因子的简单情况为例，则可将（2·8）图解如下（见图 2·1）。其中 P_i^* 是表示投影变量 X_i^* 的向量。

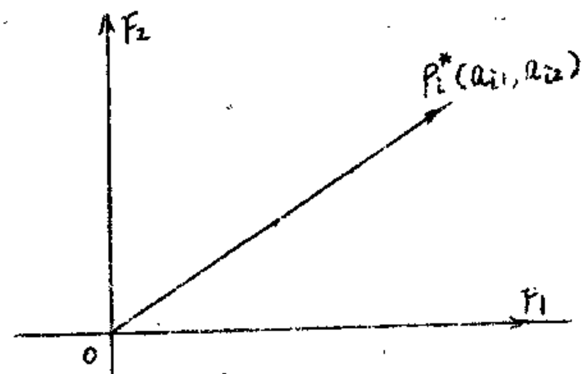

图 2·1

请注意，这时 a_{ij} 不是 P_i^* 与 F_j 之间夹角的余弦，而是 P_i^* 与 F_j 之间的内积，所以不再是 X_i^* 与 F_j 之间的相关系数。如果注意到两向量之间夹角的余弦与内积的差别，相关系数与协方差的差别，那末不难知道 a_{ij} 就是 X_i^* 和 F_j 的协方差。同样地，表示 X_i^* 与 X_j^* 的向量 P_i^* 与 P_j^* 的内积是 X_i^* 和 X_j^* 的协方差。

对分量分析。因为所有P个因子都是公共因子，所以，任意一个变量 $X_i\ (i=1,2,\cdots,p)$ 就与公共因子空间中以 $a_{i1},a_{i2},\cdots,a_{ip}$ 为坐标的向量Pi相对应。如果我们不是考虑所有的P个公共因子，而只是考虑其中方差贡献较大的一部分公共因子，那末也与在因子分析中仅考虑公共因子而不考虑单一因子的情形相彷，这时，我们就考虑诸变量在这一部分公共因子所组成的空间中的投影，至于其它问题也就不言而喻了。

三、主因子解的导出

正如上一节所提到的，因子分析的基本问题之一是用变量之间的相关决定因子载荷。为此，首先要建立两者之间的关系。

记

$$R=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}$$

其中 $r_{ij}\ (i,j=1,2,\cdots,p)$ 为变量Xi 与Xj的相关系数，当 i = j 时，$r_{ij}=1$ ，R就是诸变量 $X_i\ (i=1,2,\cdots,p)$ 的相关矩阵。

由(2·1)第j个式子可得

$$\begin{aligned} r_{ij} &= a_{i1}a_{j1}+a_{i2}a_{j2}+\cdots+a_{ik}a_{jk}, & i\neq j \\ 1 &= a_{i1}^2+a_{i2}^2+\cdots+a_{ik}^2+u_i^2, & i=j \end{aligned}$$

写成矩阵的形式，那就是

$$R=AA'+u^2$$

其中

$$
u^2 = \begin{pmatrix} u_1^2 & 0 & \cdots & 0 \\ 0 & u_2^2 & \cdots & 0 \\ - & - & - & - \\ 0 & 0 & \cdots & u_p^2 \end{pmatrix}
$$

记

$$R^* = R - u^2 \qquad (3\cdot2)$$

则可以把（3·1）改写成

$$R^* = AA' \qquad (3\cdot3)$$

由（3·2）可知，R^* 与 R 的区别仅在于对角元素，即 R^* 由 R 的对角元素依次被公共因子方差 h_i^2 取代而得。称 R^* 为约相关矩阵。然而，对分量分析 $(2\cdot1)^*$ 而言，由于没有单一因子那一项，所以代替（3·3）的是

$$R = AA' \qquad (3\cdot3)^*$$

$(3\cdot3)^*$ 与（3·3）的不同之处，除左边部分一个是相关矩阵另一个是约相关矩阵之外，右边部分的 A 也是不同的，一个是（p×p）阶的，另一个是（p×k）阶的。但是，这对以下（3·3）或 $(3\cdot3)^*$ 的求解没有影响。为了叙述方便，統一地由

$$R = AA' \qquad (3\cdot4)$$

表示，把它展开，那就是

$$r_{ij} = \sum_{\ell=1}^{m} a_{i\ell} a_{j\ell} \quad (i, j = 1, 2, \cdots p), \qquad (3.4)^*$$

其中R既表示相关矩阵又表示约相关矩阵，当R为相关矩阵时，m＝k，否则，m＝p。今后我们将始终假定约相关矩阵为半正定的，否则（3·4）将无解。

由（3·4）可知，如果A是它的解，C是任意一个（m×m）阶的正交矩阵，那末AC也必然是它的解，这是因为

$$(AC)(AC)' = (AC)(C'A') = A(CC')A' = AA'$$

的关系。这就告诉我们，（3·4）具有多解性。在不同的原则底下可以得到不同的解。用下述原则所得到的解称为主因子解（对因子分析而言）或者主分量解（对分量分析而言）。但是为了統一起见，都称为主因子解。它是导出其它解的基础。这个原则用“相关”的术语可以簡洁地陈述如下：首先根据原始的相关选出一个因子F_1，使其在各个变量的公共因子方差中所分担的部分的总和，即方差貢献为最大。然后消去这个因子的影响，而从剩余的相关中选出与F_1不相关的因子F_2，使其在各个变量的剩余的公共因子方差中所分担的部分，即方差貢献为最大。这个过程继续到各个变量的公共因子方差被分解完毕为止。

按照上述原则，求主因子解的第一步就是选出第1个因子F_1，使它的方差貢献S_1（见（2·7））在条件（3·4）*下为最大。应用拉格朗日乘数法，令

$$2f = S_1 - \sum_{i,j=1}^{p} \mu_{ij} r_{ij} = S_1 - \sum_{i,j=1}^{p} \sum_{g=1}^{m} \mu_{ij} a_{ig} a_{jg} \qquad (3.5)$$

其中 $\mu_{ij}=\mu_{ji}$ 为拉格朗日乘数。由上式可知，如果T达到最大，那末，S_1必然要达到最大。所以要求S_1最大，只要求T最大就行。根据数学分析中的极值原理，T最大的必要且充分条件是T对（$p\times m$）个变量a_{ij}中的任意一个偏导数为零，亦即

$$\frac{\partial T}{\partial a_{ij}}=\delta_{1q}a_{i1}-\sum_{j=1}^{p}\mu_{ij}a_{jq}=0\quad(i=1,2,\cdots p;\ q=1,2,\cdots,m)\tag{3.6}$$

在这里

$$\delta_{1q}=\begin{cases}1 & q=1\\ 0 & q\neq 1\end{cases}.$$

以a_{ij}乘（3·6）式的两边，并对i求和，有

$$\delta_{1q}\sum_{i=1}^{p}a_{i1}^{2}-\sum_{i=1}^{p}\sum_{j=1}^{p}\mu_{ij}a_{i1}a_{jq}=0\ .\tag{3.7}$$

最后一项能予以简化，因由（3·6）（令q＝1，注意到 $\mu_{ij}=\mu_{ji}$，且改变下标）：

$$\sum_{i=1}^{p}\mu_{ij}a_{i1}=a_{j1}\ ,$$

（3·7）变为

$$\delta_{1q}S_1-\sum_{j=1}^{p}a_{j1}a_{jq}=0\ .\tag{3.8}$$

以a_{iq}乘（3·6）的两边，且为q求和，有

$$a_{i1}S_1-\sum_{j=1}^{p}a_{j1}\Big(\sum_{q=1}^{m}a_{iq}a_{jq}\Big)=0\quad(i=1,2,\cdots p)\ ,$$

或者应用条件（3·4）*，有

$$\sum_{j=1}^{p} r_{ij} a_{j1} - S_1 a_{i1} = 0 \quad (i=1,2,\cdots,p) .$$

写成矩阵的形式，那就是

$$(r_{i1}, r_{i2}, \cdots, r_{ip}) \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{pmatrix} - S_1 a_{i1} = 0 \quad (i=1,2,\cdots,p) ,$$

或者

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ - & - & - & - \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{pmatrix} - S_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} .$$

记

$$a_1 = (a_{11}, a_{21}, \cdots, a_{p1})' , \quad 0^* = (0, 0, \cdots, 0)' ,$$

并且用E表示（p×p）阶的单位矩阵，则上式可以简化成

$$(R - S_1 E) a_1 = 0 \qquad 或者 \qquad R a_1 = S_1 a_1 . \qquad (3 \cdot 9)$$

这是个关于a_{i1}（$i=1,2,\cdots,p$）的齐次线性方程组，有非零解的充分且必要条件是，要使S_1满足条件

$$|R - \lambda E| = 0 .$$

以上是矩阵R的特征方程。因为还要求S_1为最大，所以S_1应等于R的最大特征值λ_1。因为a_1既要满足（3·9），又要满足（2·7），所以它可由R的与λ_1相对应的任一特征向量$\alpha_1 = (\alpha_{11}, \alpha_{21}, \cdots, \alpha_{p1})'$