大 气 科 学
1993 年增刊 SCIENTIA ATMOSPHERICA SINICA Sup., 1993

稳健自然聚类斜交因子分析方法的研究和应用

马 振 锋 (成都中心气象台,成都 610072)
王 柏 钧 (成都气象学院,成都 610041)

提 要

从传统斜交因子分析方法着手,提出了稳健自然聚类斜交因子分析法 (RNCOFA). 该方法结构简单,易于解释,具有较好稳健性. 考虑到气象学中有很多问题便于用因子分析方法获得解决,所以将 RNCOFA 方法所得的部分因子解用于我国月降水场及月平均温度场的预报,并与传统五种正交或斜交因子分析方法作了比较,应用取得了很好的效果.

关键词: 稳健自然聚类斜交因子分析;斜交因子解;统计天气预报.

一、引 言

斜交因子分析是因子分析方法中的一个重要部分. 在地球科学的研究中,往往利用它来作为一种分析解释和推断的手段. 由于因子解的不确定性,理论上存在着无穷多个不同的因子解. 这在客观上提供了寻找更符合实际意义的旋转解的可能. 但是如何寻找符合客观实际解释的因子解,一直是因子分析中很为棘手的问题. 虽然因子分析的正交旋转已可提供种类繁多的因子解,但大量研究表明,在地球科学和客观实际中,由于各种复杂的内在联系,正交因子解应该是较特殊的情况,而比较普遍的应是斜交因子相关解. 所以如何在斜交因子分析中寻找易于解释的解答,一直为人们所关注. 我们在本文中构造的 RNCOFA 方法,具有结构简单,易于解释,特别符合原始气象数据结构特征等特点. 对于它的构思,我们曾在文献[1]中简要作过介绍,本文则是这个方法的进一步丰富和试用. 它有以下几个特点:

(1) 在 RNCOFA 方法中除采用自然聚类以寻求斜交因子解来达到易于解释的目的外,还结合时序分析 ARIMA(p,d,q) 模型对原始场作了预报.

(2) 在应用时,对相同资料,同时用 PCA、Varimax、Quartimax、Promax 和 Harris-Kaiser 方法作了分析和预报,不论从图形分析解释来看,还是从预报客观评分指标来看,说明 RNCOFA 方法是可行的.

(3) 采用了 N 选择规则. 通常在因子分析中,究竟应选取几个因子(即因子数 k 的确定),往往采用直观的 Massy 准则,即采用方差贡献累积百分比的大小,通过特征值来加以确定,这显然具有人为的随意性,为了较客观、科学地找出确定因子数目的标准,避免主

1993 年 5 月 13 日收到,6 月 11 日收到修改稿.

观人为性，我们在方法的应用中，采用了蒙特卡洛模拟和优势方差相结合的N准则[2].

(4) 鉴于在地球科学的数据采集中，离群点 (outlier) 等因素是经常出现的，往往影响了方法的效果. 为了增强方法的稳健性 (Robustness)，我们引用了多元截尾法 (Multivariate Trimming，简记为 MVT)的思想，采用马哈拉诺比斯距离，对少量异点进行截除，经过迭代求得稳健相关阵. 以此相关阵为基础，进行 RNCOFA 分析.

二、RNCOFA 的主要思想

1. 稳健 MVT 法[3]

因子分析的基础是样本相关阵，而因子解的稳健估计可归结为相关的稳健估计. 我们采用 MVT 法，截去若干具有最大马氏距离的点来估计稳健相关阵，而这些点要通过迭代过程来确定. 具体步骤是:

(1) 用最小二乘法求出平均值向量 $\vec{M}$ 和协方差矩阵 S 的初估值.

(2) 由下式计算各点的马氏距离:

$$d_i = (\vec{x}_i - \vec{M})^T S^{-1} (\vec{x}_i - \vec{M}) \tag{1}$$

这里 d_i 是第 i 个样品点的马氏距离，$\vec{x}_i$ 为第 i 个样品点的数据向量.

(3) 将马氏距离最大的 $[\alpha n]$ 个点暂时剔除，然后由余下的点计算新的 $\vec{M}$ 和 S.

(4) 重复步骤(2)和(3)，直至达到给定的收敛水平，即前后两次迭代给出的对应协方差阵中元素的最大差值Δ的费歇变换

$$Z(\Delta) = \frac{1}{2} \ln \frac{1+\Delta}{1-\Delta} \tag{2}$$

小于某给定的很小正数. 迭代收敛时的 S 即为我们所求得的稳健协方差矩阵，对它进行适当的变换就可得出稳健相关矩阵.

2. N 准则

为了避免 Massy 准则的人为性，我们采用了以蒙特卡洛模拟为基础的优势方差准则，即N准则. 它假设原始数据来源于 p 维正态总体，即服从 $N_p(0,\Sigma)$ 分布，其中Σ为对角总体协方差矩阵. 现在假设进行 100 次独立随机抽样，每次得到 np(n 为原始资料样本数)个服从上述正态分布的数据，又假设这 np 个数据构成的随机相关阵的非零特征值为

$$I_1(\omega) \geqslant I_2(\omega) \geqslant \cdots \geqslant I_p(\omega) \qquad \omega = 1,2,\cdots,100$$

令

$$U_i(\omega) = I_i(\omega) \left[p^{-1} \sum_{i=1}^{p} I_i(\omega) \right]^{-1} \qquad i = 1,2,\cdots,p$$

且 $U_i(\omega_1) < U_i(\omega_2) < \cdots < U_i(\omega_{100})$，再令

$$\begin{cases} \sigma_i(5) = U_i(\omega_5) \\ \sigma_i(95) = U_i(\omega_{95}) \end{cases}$$

对于给定数据 Z 与之相联系的非零特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$，令

$$V_i = \lambda_i \left(p^{-1} \sum_{j=1}^{p} \lambda_j \right)^{-1}$$

那么我们得到规则 N 为

$$\begin{cases} K\text{取满足 } V_i > \sigma_i(95) \text{ 中 } i \text{ 之最大值} \\ \text{否则 } K \text{ 不存在} \end{cases} \tag{3}$$

3. 自然聚类因子分析

我们针对实际数据特点，采用适合球状点群结构的 ISODATA 和 ISOMIX 动态聚类方法[4]，用以寻求斜交因子，并通过斜交因子变换、相关、模型、结构等矩阵的关系，计算出有关斜交因子解来.

4. 时序分析的 ARIMA 模型

当斜交因子载荷求得以后，我们可以计算出因子得分 $F_i(t)$. 为此我们对它建立自回归滑动平均季节模型

$$\begin{aligned} F_i(t) = {} & \phi_1 F_i(t-1) + \phi_2 F_i(t-2) + \cdots + \phi_p F_i(t-p) \\ & + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \end{aligned} \tag{4}$$

式中 $\phi_1, \phi_2, \cdots, \phi_p$ 为 p 阶自回归参数，$\theta_1, \theta_2, \cdots, \theta_q$ 为 q 阶滑动平均参数，$a_t, a_{t-1}, \cdots, a_{t-q}$ 为白噪声序列，我们在 ARIMA(p,d,q) 模型中的 d 取 0 或 1. 由于此模型可以推导出一种等价形式，即逆转形式

$$a_t = F_i(t) - \sum_{j=1}^{\infty} l_j F_i(t-j) \qquad i = 1, 2, \cdots, k \tag{5}$$

式中 l_j 称为逆函数，它具有负指数下降的性质 $(l_0 = -1)$. 逆转形式是由随机序列的加权和表示白噪声序列，利用逆函数 l_j 与模型参数 ϕ_i 与 θ_j 间的线性关系，可以发展出确定参数初值的逆函数估计法. 这样我们可由 ARIMA 模型得到 $F_i(t)$ 的 m 步预报，进而可由因子分析模型计算出气象要素场的预报值.

三、结果与讨论

我们采用我国 160 个测站从 1951 到 1985 年月降水量和月平均温度资料，构成 420 × 160 的原始数据阵，以该阵为基础，做了 RNCOFA 分析，同时还进行了 PCA、方差极大、四次幂极大、Harris-Kaiser 和 Promax 五种正交或斜交因子分析，并用它们的因子解作了两个场的一年预报. 下面介绍一些主要结果.

1. 因子解的稳健分析

为了得到两个气象要素场的稳健相关矩阵，首先通过计算马哈拉诺比斯距离，并经过试验对比，确定了降水场和温度场应用 MVT 法的截尾比例分别是 12% 和 7%. 再经过

迭代得出了两个场的稳健相关矩阵 $R(160 \times 160)$，对比原始数据相关阵 $r(160 \times 160)$，发现两者有以下近似关系：

$$R_{ij} \approx \begin{cases} r_{ij} + 0.04 & 0.80 < r_{ij} \leqslant 1 \\ r_{ij} & |r_{ij}| \leqslant 0.80 \\ r_{ij} - 0.04 & -1 \leqslant r_{ij} < -0.80 \end{cases} \tag{6}$$

以两个场的稳健相关阵为基础，首先进行了 PCA 分析，并结合蒙特卡洛模拟 N 规则确定了降水场前 20 个主因子和温度场前 4 个主因子，它们分别解释了各自场总方差的 80.6% 和 98.5%。

2. 因子载荷的实际意义

我们把降水场的 20 个主因子和温度场的 4 个主因子作为初始解，进行 RNCOFA 等其它五种正交或斜交因子分析．图 1 给出了由 RNCOFA 方法得到的降水场前六个斜交因子载荷分布．由图 1a-c 可知，一般对应于不同因子仅有一个显著载荷区与之对应，它们顺次是华北、华南、西北、东北、长江流域及西南地区．这些对应于不同因子的区域与前人工作[5]比较，似乎更具有天气、气候意义．例如华北、西北地区可能与西风槽降水系

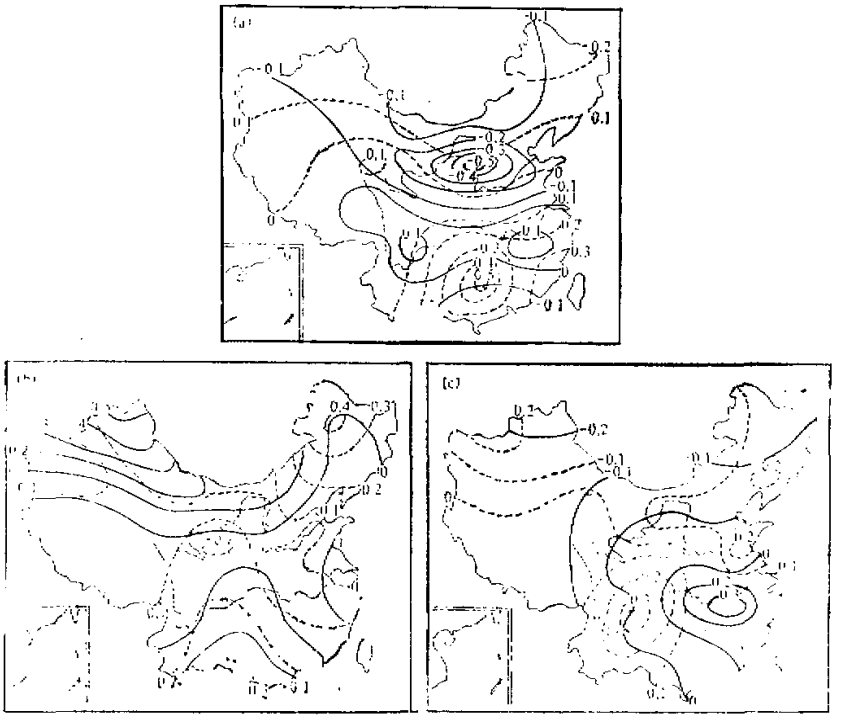

图 1　我国月降水场前六个斜交因子载荷分布

(a) 第一实线，第二虚线，(b) 第三实线，第四虚线，(c) 第五实线，第六虚线。

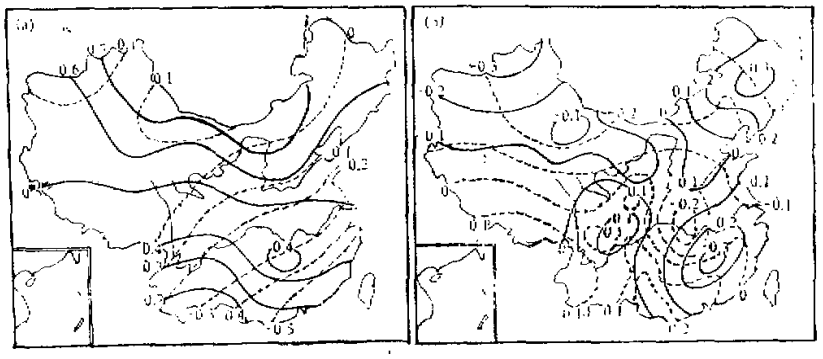

图 2 我国月平均温度场前四个斜交因子载荷分布
(a) 第一实线,第二虚线; (b) 第三实线,第四虚线.

统相对应,华南地区可能与台风和西南低涡系统相对应. 值得注意的是,这六个降水模式基本给出了我国降水的空间分布型,同文献[5]给出的我国降水空间分型模式基本一致. 我们再来分析由 RNCOFA 方法给出的温度场前四个因子载荷分布 (图 2a-b). 第一斜交因子载荷反映了我国气温随纬度的变化,等值线大体上与纬圈平行,只是在江南地区有一向西南方向伸展的低值区. 第二因子载荷反映我国气温随经度的变化,等值线大体上呈东北-西南向,可能与我国地形分布有关.第三因子载荷场分布则把我国分成四块,分别为东北、西南、西北、江淮-华南地区,反映了我国四种不同的大的气候类型。第四因子载荷的分布型式与我国大陆度分布甚为一致[5].可能反映了我国气温受海陆分布的影响.

3. 降水场和温度场的预测

我们将以上正交、斜交因子分析所得的前六个(降水场)和前四个(温度场)因子载荷对应的因子得分看成多元变量的时间序列,应用 ARIMA(p,d,q) 模型进行随机建模. 因子得分的主周期由周期图和方差分析加以揭示, ARIMA(p, d, q) 模型识别和参数估计用 Pandit-Wu 方法[6] 进行. 表 1 和表 2 分别列出了由 RNCOFA 方法给出的降水场前六个和温度场前四个因子得分序列的周期和 ARIMA 模型的阶数及参数的计算结果.

表 1 降水场前六个因子得分的周期和建模阶数、参数

序号	周 期	模型	自回归阶数	滑动平均参数
1	10	ARIMA(4,1,3)	1.229,−0.692,0.728,−0.736	0.137,−0.499,0.233
2	8	ARIMA(6,1,4)	1.651,−1.358,1.412, −1.392,1.424,−0.528	0.657,−0.832, 0.807,−0.606
3	—	ARIMA(3,0,3)	1.108,−0.323,0.220	0.190,−0.264,0.013
4	12	ARIMA(5,1,3)	1.268,−0.883,−0.951, −0.914,0.976	0.153,−0.632,0.293
5	—	ARIMA(6,0,5)	1.528,−1.168,1.210, −1.216,−1.295,0.408	0.504,0.675,0.613, −0.586,0.603
6	11	ARIMA(4,1,2)	1.525,−1.065,1.195,−1.303	0.314,−0.671

为了检验预报结果，采用如下两个误差标度：

（1）距平符号百分率

表 2　温度场前四个因子得分的周期和建模阶数、参数

序号	周　期	模型	自回归阶数	滑动平均参数
1	12	ARIMA(5,1,3)	0.962,0.061,−0.168,0.209,−0.171	−0.397,0.136,−0.331
2	10	ARIMA(4,1,3)	0.985,0.004,−0.091,0.129	−0.387,0.136,−0.332
3	—	ARIMA(3,0,2)	0.965,0.041,−0.145	−0.295,0.053
4	—	ARIMA(4,0,3)	0.965,0.051,−0.158,0.0191	−0.272,0.079,−0.012

$$H_1 = R_1/p \tag{7}$$

式中 R_1 是预报的距平符号和观测值的距平符号相同的个数，$p = 160$ 是测站数。

（2）绝对差值百分率

$$H_2 = R_2/p \tag{8}$$

式中 R_2 是预报值与观测值差的绝对值小于 1℃ 的个数。降水场和温度场的预报评分列表 3、表 4。从评分结果来看，不论是降水场还是温度场，预报得分的年平均值 RNCOFA 方法是比较好的，尤其对温度场的预报 RNCOFA 方法高于其它五种方法。

表 3　1986 年我国月降水预报评分(H_1)(%)

方　法	1	2	3	4	5	6	7	8	9	10	11	12	平均
RNCOFA	60.0	58.8	51.3	50.3	56.3	56.8	54.4	51.9	46.9	48.1	43.8	41.9	51.7
PCA	57.5	58.1	58.8	53.1	52.5	58.8	50.6	48.8	48.1	47.5	45.6	44.4	52.0
QUA.	55.6	54.3	57.5	50.6	51.3	52.5	52.5	50.6	50.0	46.8	47.5	48.1	51.5
VAR.	56.3	52.5	54.4	50.0	57.5	56.3	51.2	49.4	48.8	47.5	46.2	49.3	51.6
H-K.	59.4	58.1	59.3	56.2	54.4	50.6	52.5	50.0	47.5	43.8	45.6	40.6	51.5
PRO.	54.4	55.0	51.9	48.8	46.3	50.6	51.3	49.4	48.8	49.4	47.5	47.5	50.3

表 4　1986 年我国月平均温度预报评分(H_1)(%)

方　法	1	2	3	4	5	6	7	8	9	10	11	12	平均
RNCOFA	79.4	82.5	81.9	76.3	75.6	76.3	75.6	70.6	71.3	70.0	69.4	68.8	74.8
PCA	71.9	71.3	68.8	70.6	68.8	69.4	71.3	66.9	67.5	70.8	63.8	64.4	68.7
QUA.	78.1	78.1	80.0	80.0	77.5	71.3	73.1	70.6	68.1	69.4	68.8	63.7	73.2
VAR.	75.6	75.0	77.5	76.3	71.3	70.0	71.9	67.5	65.6	65.0	63.1	63.1	70.2
H-K.	75.0	73.8	74.4	72.5	71.9	73.8	70.0	66.3	64.4	61.9	61.3	61.9	68.9
PRO.	76.9	84.4	82.5	81.9	77.5	78.1	75.0	73.8	68.1	66.3	63.1	62.5	74.2

四、结　论

我们从不同于传统斜旋转因子分析的角度，构造了一种新的斜交因子分析法(RNCO-

FA)，除了具有内容新颖、结构简单、解释明确和灵活实用的特点外，而且具有较好的稳健性，是对经典因子分析方法的某种改进．尤其对受奇异点影响的非正态、非平稳的气象数据进行因子分析，将会具有较好的可靠性与合理性．

在 RNCOFA 的气象应用中发现，斜交因子模型的空间分布具有天气、气候意义．从因子解与 ARIMA 模型结合做预报的准确率来看，也接近或超过目前的长期业务预报水平．

总之，无论从分析解释的角度，还是从预报角度，我们均把 RNCOFA 方法与其它五种正交或斜交因子分析进行了对比，说明 RNCOFA 方法是可行的．但因篇幅所限，不便一一列举．当然，RNCOFA 方法在理论上还很不完善，需要进一步改进．

参 考 文 献

[1] 王柏钧，1986，关于因子分析方法的研究和探讨，成都气象学院科技，7(9)，7—13．

[2] Mobley, C.D., 1988, Principal component analysis in meteorology and oceanography. Elsevier: AONT Press, 517—539.

[3] Devlin, S.J., 1981, Robust estimation of dispersion matrices and principal components, *J. Amer. Statist. Assoc*, **76**, 354—362.

[4] 王柏钧、程积康，1982，多元分析，地质出版社，184—193．

[5] 盛承禹，1986，中国气候总论，科学出版社，371—403．

[6] Pandit, S. M. Wu., 1983, Time series and system analysis with application, New York, John Wiley and Sons, 19—213.

A Study of the Robust Nature Cluster Oblique Factor Analysis Method

Ma Zhenfeng

(*Chengdu Weather Centre, Chengdu* 610072)

Wang Baijun

(*Chengdu Institute of Meteorology, Chengdu* 610041)

Abstract

Based on the traditional oblique factor analysis, the robust nature cluster oblique factor analysis (RNCOFA) is presented in this paper. Solution from RNCOFA possesses the simple structure and robustness, and is also very explicit and easy to be explained. Applications of this method to meteorology is also discussed. The solution from RNCOFA is used to forecast the monthly precipitation and monthly mean surface temperature over China. Compared with the traditional orthogonal or oblique method, such as Varimax, Quartimax, Promax and Harris-Kaiser, the result of RNCOFA is satisfactory.

Key words: RNCOFA; Oblique factor solution; Statistical weather forecast.