第二版

气象科研与预报中的多元分析方法

施能　编著

气象出版社

气象科研与预报中的多元分析方法

气象科研与预报中的多元分析方法

（第二版）

施 能 编著

气象出版社

内 容 简 介

本书较详细地叙述了多元统计分析的基本理论及其在气象科研及预报中的应用。内容包括概率论基本原理、统计检验、选择最大信息的预报因子、回归分析、二分类预报、考虑经济效益的决策、主成分分析、气象场的经验正交展开、气象场的奇异值分解、判别分析、聚类分析及奇异谱分析等。各种方法都附有原始数据资料以及详尽的计算步骤和结果。本书可供气象台预报员、科研工作者参考，亦可作为高等院校气象统计预报课程的教材或参考书。对于从事农业气象预报、水文、海洋、地震预报、地质数学、生物统计、市场预测、医学统计的科技工作者和有关院校的师生也有一定参考价值。

图书在版编目(CIP)数据

气象科研与预报中的多元分析方法(第二版)/施能编著. —2 版.
—北京：气象出版社，2002. 2

ISBN 7—5029—3280—1

Ⅰ.气…　Ⅱ.施…　Ⅲ.①气象资料—多元分析
②天气预报—多元分析　Ⅳ.P456.8

中国版本图书馆 CIP 数据核字(2001)第 079145 号

气象科研与预报中的多元分析方法(第二版)

施 能 编著

责任编辑：林雨晨　终审：黄润恒

封面设计：李洪杰　责任技编：王丽梅　责任校对：吾人

*

气象出版社出版

(北京市海淀区中关村南大街 46 号　邮政编码：100081)

北京怀柔奥龙印刷厂印刷

新华书店北京发行所发行　全国各地新华书店经销

*

开本：850mm×1168mm　1/32　印张：7.875　字数：210 千字

2002 年 2 月第二版　2002 年 2 月第一次印刷

印数：1～3000　　定价：16.00 元

再版前言

值此再版的机会对原版中的一些错误作了改正，精简了一些作者认为应用较少的内容，增加了第九章奇异谱分析。奇异谱分析是近十年发展起来的数字信号处理技术，它是从含有噪声的有限长观测序列中提取与分析周期的方法，虽然分析的对象是一维时间序列，但是在方法与原理上与多元分析关系密切，所以列入本书。

本书的特点是照顾两头，即以实用性为主和兼顾前沿性。它既是一本专著，又适合作为教材使用。本书已经在南京气象学院大气科学专业使用11届。在本书的使用和修改过程中，广泛征求了老师、同行的意见，对他们的热情指正在此致以衷心的感谢！

施 能

2001年7月于南京气象学院

符号说明

名　　称	样　本	总　体
变量 x 的平均值及均方差	$\overline{x},s_x$	μ,σ
变量 x 均方差的无偏估计	s_x^*	σ
变量 x 在两类的均值	$\overline{x(A)},\overline{x(\overline{A})}$	$\mu(A),\mu(\overline{A})$
x 在第 g 组的均值	$\overline{x(g)}$	$\mu(g)$
x 在第 g 组的均方差	$s_x(g)$	$\sigma(g)$
x 与 y 的协方差	s_{xy}	σ_{xy}
x 与 y 的相关系数	r_{xy} 或 r	ρ_{xy} 或 ρ
马哈拉诺比斯平方距离	D_m^2	Δ_m^2
概率密度函数	/	$f_A(x)$、$f_{\overline{A}}(x)$
均值向量及协方差矩阵	$\overline{\boldsymbol{x}},\boldsymbol{S}$	$\boldsymbol{\mu},\boldsymbol{\Sigma}$
离差矩阵	$\boldsymbol{SS}$	/
第 g 组的协方差矩阵	$\boldsymbol{S}_g$	$\boldsymbol{\Sigma}_g$
两类协方差矩阵	$\boldsymbol{S}(A),\boldsymbol{S}(\overline{A})$	$\boldsymbol{\Sigma}(A),\boldsymbol{\Sigma}(\overline{A})$

符　号	含　　义
$\boldsymbol{x}$	随机向量，即$(x_1,x_2,\cdots,x_n)$
$\boldsymbol{X}$　$\boldsymbol{X}^0$　$\boldsymbol{X}^*$	资料矩阵、中心化的、标准化的
$\boldsymbol{W},\boldsymbol{T}$	组内离差矩阵，总离差矩阵
$U(\boldsymbol{x}),W(\boldsymbol{x})$	线性判别函数，二次判别函数
$\boldsymbol{x}_i$	第 i 个样本列向量，即$(x_{1i}\quad x_{2i}\quad\cdots\quad x_{mi})^T$
$\boldsymbol{V}$	特征向量作为列组成的矩阵
${}_mZ_n$	m 个主成分组成的矩阵
$\boldsymbol{v}_i$	第 i 特征向量，即$(v_{i1}\quad v_{i2}\quad\cdots\quad v_{im})^T$
$R_{y\cdot x_1x_2\cdots x_m}$	y 与 $x_1,x_2,x_3,\cdots,x_m$ 的复相关系数
${}_m\boldsymbol{R}_m$	m 个变量的相关矩阵

目　录

第一章　气象资料及其表示方法

第一节　单个要素的气象资料

1. 数据资料

气象资料绝大多数是以数据形式给出的，例如气压、湿度、温度、降水量等等。因为这些数据是经常变化的，所以又可以用变量 $\boldsymbol{x}$ 表示某个气象要素的取值，它在 n 个时间的取值可分别表示为 $x_1, x_2, \cdots, x_n$。n 称为样本容量。数学上还可采用向量符号表示这一组数

$$\boldsymbol{x} = (x_1 x_2 x_3 \cdots x_n)^T \tag{1.1.1}$$

或写为

$$\boldsymbol{x} = (x_t)^T \qquad t = 1,2,3,\cdots,n \tag{1.1.2}$$

$\boldsymbol{x}$ 确定为 n 维空间的一个点。另一方面，在一维空间(单坐标)中，这 n 个数据又确定为 n 个点，这就是经常使用的单因子(坐标)点聚图。

2. 数据资料的统计特征

(1)平均值

n 个数据资料的平均值 $\overline{x}$ 表示为

$$\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{1.1.3}$$

它反映了该气象要素的平均状况。平均值总是处在资料数据范围之内，它反映的要素特征是不全面的，无法表示资料偏离平均值的情况。

(2)距平、方差、均方差

1)距平

某数据资料与平均值的差称为距平。表示为

$$x_t^0 = x_t - \overline{x} \qquad t = 1,2,\cdots,n \tag{1.1.4}$$

距平能够反映数据偏离平均值的状况，例如南京 1957 年 1 月份降水量是 54.9mm，同年 6 月份降水量是 120.5mm。如果不用距平来反映月降水量的多少，只能得出 1957 年 6 月份降水比 1 月份降水多的结论，这是众所周知、毫无意义的。如果各自减去它们的多年平均值，则 1957 年南京 1 月份降水距平是＋18.8mm，而 1957 年 6 月份南京降水距平是－24.2mm。这样，就认为 1957 年 1 月份南京降水是偏多的，而 6 月份降水是偏少的。距平的作用就在于将要素的值化到同一水平上进行比较。距平是有单位的，其单位与原变量单位相同。气象上经常用距平变量 x_t^0 代替原变量进行研究，由于 x_t^0 的平均值 $\overline{x_t^0}$ 等于零，所以带来许多方便，这种处理方法也称为中心化。

2)方差和均方差

对于某气象要素 x 的 n 个资料，它的方差表示为

$$s_x^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^2 \tag{1.1.5}$$

均方差表示为

$$s_x = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^2} \tag{1.1.6}$$

均方差反映了变量围绕平均值的平均变化程度(离散程度)。例如对某年某月的日平均气温求均方差，其值比较小，则表示该年该月的气温日际变化平均比较小，反之则表示气温日际变化很大，冷暖空气的活动比较频繁。

均方差反映了变量的平均变化程度，因此它实际上反映了对该变量作预报的易难程度。显然均方差小的变量比均方差大的变量容易预报。事实上，可以证明变量的不确定性(即统计熵)和该变量的均方差的对数成正比。所以均方差越大，变量不确定性越大，预报就越困难了。

计算方差或均方差时可以利用一个性质，就是变量减去某常

数后其方差，均方差不变。例如表 1.1.1 中 x 和 $x-100$ 的均方差是相同的。

表 1.1.1

t	1	2	3	4	5	6	7	8	9	均方差
x_t	100	105	106	108	110	112	114	115	99	5.40
x_t-100	0	5	6	8	10	12	14	15	−1	5.40

(3)频率分布

我们先看表 1.1.2 中的 x 与 y 的 12 个数据。计算它们的平均值和均方差都是一样的。但它们的取值特征显然有很大差别。要反映这种差别，用平均值和均方差是不够的。这就需要用到累积频率这一概念。所谓累积频率是变量小于某上限值的次数与总次数之比。表 1.1.3 是根据表 1.1.2 统计的累积频率。可以看出，两个变量的累积频率有明显差别的，并且它是上限值的函数。

表 1.1.2

t	1	2	3	4	5	6	7	8	9	10	11	12	平均值	均方差
x	−6	−5	−4	−3	−2	0	0	1	2	4	6	7	0	4.04
y	−8	−5	0	0	0	0	0	0	0	0	3.15	9.85	0	4.04

3. 总体和样本

对于一批数据，它可能代表变量的全部取值，也可能只代表其中一部分。所谓总体是统计分析对象的全体，一个变量的全部可能取值组成总体，也称为母体。总体中的一部分资料组成样本。这样，总体和样本的关系就是全局和局部的关系。总体是未知的，样本是已知的。总体的特征是客观存在的，样本特征如平均值、均方差等是随样本而变化的。所以，与样本有关的变量均称为随机变量。往往需要用样本来作出关于总体的一些推断。因此，确定所讨论问题的总体，并从中抽出有代表性的样本是很重要的，否则推断是无效的。总体与样本的关系不是一成不变的，视所研究的对象和任务而

定。气象上的总体，一般都是指无限总体，一组气象资料就是无限总体中的样本。

表 1.1.3(据表 1.1.2 统计)

上限	−8	−7	−6	−5	−4	−3	−2	−1
x 的累积频率	0	0	0	1/12	2/12	3/12	4/12	5/12
y 的累积频率	0	1/12	1/12	1/12	2/12	2/12	2/12	2/12
上限	0	1	2	3	4	6	7	10
x 的累积频率	5/12	7/12	8/12	9/12	9/12	10/12	11/12	12/12
y 的累积频率	2/12	10/12	10/12	10/12	11/12	11/12	11/12	12/12

4. 分布函数

前面所讲过的累积频率是用有限资料统计的，是样本特征。在无限总体中的累积频率称为分布函量。从表 1.1.3 中看到，累积频率是上限的函数，如果我们用 x 表示上限，用 ξ 表示变量，则分布函数 $F(x)$ 定义为

$$F(x) = P(\xi < x) = \int_{-\infty}^{x} f(x)\mathrm{d}x$$

$f(x)$ 称为概率密度函数。最常见的 $f(x)$ 的函数形式是正态分布

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.1.7}$$

其中 μ 和 σ 分别是 ξ 的总体平均值(数学期望)和均方差，它是正态分布的二个参数，需用样本值进行估计。其中 μ 可以用样本平均值去估计；σ 可用 s_x 或用无偏估计量 s_x^* 估计，即

$$s_x^* = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - x)^2} \tag{1.1.8}$$

当随机变量的密度函数为(1.1.7)式时，常可简记为N(μ σ^2)。

5. 数据的标准化和正态化

(1)标准化

气象数据资料都是有单位的，为了消去单位量纲不同所造成

的影响，经常使用标准差标准化方法。

设 x_{it} 是第 i 个气象要素的第 t 个资料，则经过标准差标准化的资料为 x_{it}^*

$$x_{it}^* = \frac{x_{it} - \overline{x}_i}{s_i} \qquad t = 1,2,\cdots,n \tag{1.1.9}$$

其中 $\overline{x}_i, s_i$ 分别是第 i 个气象要素的样本平均值和均方差。经标准差标准化后的资料 x_{it}^* 的平均值为 0，均方差为 1，无单位。式(1.1.9)中，如果不除以分母，也就是用距平值代替，则称为中心化。

(2)正态化

资料必须符合正态分布，这几乎是各种统计预报模型的理论依据，也是常用的统计检验方法 (F,t,u,χ^2 检验法)的先决条件。经验表明，年、月平均温度、气压、多雨地区的月降水量通常符合正态分布。旬平均气温尚符合正态分布，但旬、候降水量不一定符合正态分布，日降水量和少雨地区的月降水量通常是偏态的。对于不服从正态分布的变量需作正态化处理，以适合各类统计预报模型，处理方法有

1)立方根(或四次方根)转换

$$x''_{it} = \sqrt[3]{x_{it}}$$

或

$$x''_{it} = \sqrt[4]{x_{it}} \qquad t = 1,2,\cdots,n \tag{1.1.10}$$

当 x_{it} 不符合正态分布时，(1.1.10) 的转换使 x''_{it} 较能符合正态分布。不少工作提出，这种转换是方便有效的。

2)双曲正切转换

$$z_{it} = \mathrm{th}\,\frac{x_{it} - \overline{x}_i}{x_i} = \mathrm{th}\,y_{it} = \frac{\mathrm{e}^{y_{it}} - \mathrm{e}^{-y_{it}}}{\mathrm{e}^{y_{it}} + \mathrm{e}^{-y_{it}}} \tag{1.1.11}$$

其中 $\overline{x}_i$ 仍为第 i 个要素的平均值。z_{it} 即为经过双曲正切转换后的值。研究表明，双曲正切转换在旬降水量转换时有较大的优点。因为旬降水量可能不是正态分布，它的极端值会使所得的相关关系

不真实。所以经双曲线正切转换后，还能使所得的相关关系具有较大的稳定性和可靠性。

3)化为有序数后的正态化转换

这种转换方法是先将需作正态化变换的 x_i 的 n 个数据按数值大小排列($x_{i1} \leqslant x_{i2} \leqslant x_{i3} \cdots \leqslant x_{in}$)，再将 x_{it} 对应的序号 t 利用公式

$$\frac{t}{n+1} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_{it}^*} e^{-\frac{1}{2}x^2} dx \qquad (1.1.12)$$

变换为 x_{it}^*，x_{it}^* 即服从标准正态分布。例如 $n = 89$，$t = 30$，则 x_{i30} 经标准正态化转换后化为 x_{i30}^*，由(1.1.12) 知 $x_{i30}^* = -0.431$。这个方法既作了正态化处理，又作了标准化处理。

6. 状态资料和统计特征

(1)状态资料

对某些气象现象的观测结果是无法用数据形式表示出来的，例如天气现象雾、冰雹、霜等等，只能表示“有”、“无”、“强”、“弱”等状态。此外，对于降水量这样一些数值进行分级也可转化为非数据的状态资料，如特大暴雨、大暴雨、大雨、中雨、小雨、微迹、晴等等。为区别起见，有时我们用离散化的资料来表示这些状态，如用 0 表示不出现某类天气现象，用 1 表示出现某类天气现象。用 5,4,3,2,1,0 表示上述各降水等级。但应该注意，这些数值不一定能反映各状态之间的差异。例如，5 与 4 之差是 3 与 1 之差的一半，用这个来反映降水量级别之间的差别显然是不合适的。所以，这些数字一般不用来进行数值计算。通常，只能理解为某种“记号”，它无异于甲、乙、丙、丁表示各种状态。所以对这类状态资料，我们不再谈及它们的平均值和均方差，而只介绍频率表或分布列。

(2)频率表、分布列

对于气象要素现象的各种状态，我们列出各种状态出现的频率，那末该现象的统计特征也就最完美地表现出来了。

例如，某站台在 4～9 月的 50 年历史资料中，共出现有冰雹日 183 天，则表示 4 至 9 月的冰雹日的状态可用表 1.1.4 表示。这种

表也称为频率表，如果这种表是对总体统计的，这种表就称为分布列。

表 1.1.4

状态	有冰雹 1	无冰雹 0
频率	$\frac{183}{9150}=0.02$	$\frac{8967}{9150}=0.98$

第二节　多要素的气象资料

1. 数据矩阵

气象中经常要利用许多气象要素资料，这时可将资料写成矩阵形式。设有 m 个气象要素，每个气象要素有 n 个观测值，则任何一个数据表示为 $x_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,n)$。全部写出并排列成矩阵形式有

$$
{}_m\boldsymbol{X}_n=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}=(\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_n) \qquad (1.2.1)
$$

其中 $\boldsymbol{x}_i=(x_{1i}\ \ x_{2i}\cdots x_{mi})^T$ 为第 i 个样本的资料向量，它是由公式(1.2.1) 所表示的矩阵的第 i 列元素组成($i=1,2,\cdots,n$)。

对于(1.2.1) 式的数据矩阵我们可以从两方面去考虑，如果比较任意两行，则就是检查相应两个变量之间的关系。如果比较任意两列，这一比较将给出 m 个变量集合中的该两个样本之间的关系。通常对第一种研究变量之间相互关系的分析，称为“R 型分析”；对后一种研究样本之间关系的称为“Q 型分析”。

2. 数据的二种空间表示

1)将数据矩阵(1.2.1)看成为 m 维空间的 n 个点。这是一种空间点聚图。我们设想有 m 个正交轴，确定了一个 m 维空间。这样，任何一个样本对应 m 维空间的一个点，n 个样本就对应 m 维空间

的 n 个点，其中第 j 个点的沿第 i 个轴的坐标值就是 x_{ij}。这种空间表示方法经常用在分析样本之间的关系，如寻找相似个例等。

2) 将数据矩阵(1.2.1)看成为 n 维空间的 m 个点。取 n 个正交轴，每一个轴对应一个样本。m 个变量就确定为 n 维空间的 m 个点。如果考虑数据矩阵(1.2.1)按行中心化，也就是每行各个元素减去该行的平均值，则第 i 个点的沿第 j 个坐标轴的坐标值是 $x_{ij}-\overline{x}_i$。设第 i 个点是 R_i，则 R_i 与坐标原点 O 的距离平方是

$$OR_i^2=\sum_{j=1}^{n}(x_{ij}-\overline{x}_i)^2 \qquad i=1,2,\cdots,m \qquad (1.2.2)$$

恰好是第 i 个变量的方差的 n 倍。以后我们还会看到，在这样的空间坐标轴中的任意二点与坐标原点之夹角的余弦 $\cos\theta_{ij}$ 就是变量 x_i 和 x_j 的相关系数。

$$\cos\theta_{ij}=\frac{\sum_{t=1}^{n}(x_{it}-\overline{x}_i)(x_{jt}-\overline{x}_j)}{\sqrt{\sum_{t=1}^{n}(x_{it}-\overline{x}_i)^2\cdot\sum_{t=1}^{n}(x_{jt}-\overline{x}_j)^2}} \qquad (1.2.3)$$

所以这种空间经常用来研究变量之间的关系。

3. 均值向量

因为有 m 个变量，所以可分别求出 m 个变量的样本平均值，这 m 个平均值组成一个向量就是均值向量，表示为

$$\overline{\boldsymbol{x}}=(\overline{x}_1\overline{x}_2\cdots\overline{x}_m)^T \qquad (1.2.4)$$

其中 $\overline{x}_i=\frac{1}{n}\sum_{t=1}^{n}x_{it}$。$\overline{\boldsymbol{x}}$ 经常看成 m 维空间的一个点，实际上它就是 m 维空间中的 n 个点的重心。

4. 协方差和协方差矩阵

当利用多个变量时，为了表示变量之间的相互关系，需要计算协方差。变量 x_i 和 x_j 的协方差定义为

$$s_{ij}=\frac{1}{n}\sum_{t=1}^{n}(x_{it}-\overline{x}_i)(x_{jt}-\overline{x}_j)$$

$$= \frac{1}{n}\sum_{t=1}^{n} x_{it} \cdot x_{jt} - \overline{x}_i \overline{x}_j \quad i,j = 1, 2 \cdots, m \qquad (1.2.5)$$

这 m^2 个协方差(当 $i=j$ 时为方差)组成的矩阵称为样本协方差矩阵,表示为

$$S = (s_{ij}) = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{bmatrix} \qquad (1.2.6)$$

这个矩阵是 m 阶矩阵,并且是对称矩阵,其对角线元素 s_{ii} 是第 i 个变量的方差。(1.2.5)式可用矩阵相乘直接求得。令 X^0 是中心化资料组成的矩阵

$$X^0 = \begin{bmatrix} x_{11}^0 & x_{12}^0 & \cdots & x_{1n}^0 \\ x_{21}^0 & x_{22}^0 & \cdots & x_{2n}^0 \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1}^0 & x_{m2}^0 & \cdots & x_{mn}^0 \end{bmatrix} \qquad (1.2.7)$$

则
$$_{m}S_{m} = \frac{1}{n} X^0 \cdot (X^0)^T \qquad (1.2.8)$$

其中 T 表示矩阵转置;而 $x_{ij}^0 = x_{ij} - \overline{x}_i$ 是中心化资料。

有时为方便起见,用离差积 ss_{ij} 组成矩阵,称为离差矩阵,用 SS 表示

$$ss_{ij} = \sum_{t=1}^{n} (x_{ij} - \overline{x}_i)(x_{jt} - \overline{x}_j)$$

$$SS = (ss_{ij}) \qquad (1.2.9)$$

用公式(1.2.6)计算的矩阵 S 是 m 个变量总体协方差矩阵 Σ 的估计矩阵,但不是无偏估计。为了得到 Σ 的无偏估计,应使用

$$S = \frac{1}{n-1} SS = (s_{ij}^*) \qquad (1.2.10)$$

$$s_{ij}^* = \frac{1}{n-1}\sum_{t=1}^{n} (x_{it} - \overline{x}_i)(x_{jt} - \overline{x}_j) \qquad (1.2.11)$$