多元统计方法

方开泰 编

中国科学院应用数学研究所
概率统计咨询服务部 印
1984.7

编 者 的 话

近年来，许多部门经常要求我们举办多元分析方法的学习讨班。为了给这一类活动提供适当的教材，特地选编了部分常用的方法。其中包括：多变量样本的图分析法，聚类分析，回归分析、判别分析，主分量分析和因子分析。由于这些材料来源不同，符号体例均无法统一，希读者谅解。

方开泰

1984·6·18

目　录

多变量样本的图分析法（一）

中国科学院应用数学研究所　方开泰

图形是帮助人们思维和判断的重要工具。当样本只有两个特性（变量或指标）时，可以用通常的直角坐标在平面上点图。当样本有三个变量时，虽然可以在三维的笛卡儿坐标里点图，但也是很不方便的。当变量数大于三时，用通常的方法已不能点图了。在多元分析中，样本的变量数一般均大于三。探讨多变量的点图法是长期来一直为人们所关注的研究课题，这里介绍一些有关的方法，特别是近十年来发展的一些方法。

作者非常感谢刘璋温副教授，因为他提供了许多有用的文献，并仔细校阅和修改了此文。

一、雷达图（radar chart）

设有n个样本$X_1, X_2, \cdots, X_n$，各有p个变量$x_1, x_2, \cdots, x_p$，用x_{ij}（$i=1, 2, \cdots, n$；$j=1, 2, \cdots, p$）表示第i个样本的第j个变量的值。通常数据$\{x_{ij}\}$列成表1·1的形式。

表1·1　数据矩阵

样本＼变量	x_1	x_2	…	x_j	…	x_p
X_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
X_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
⋮	⋮	⋮	…	⋮		⋮
X_i	x_{i1}	x_{i2}	…	ij	…	x_{ip}
⋮	⋮	⋮	…	⋮	…	⋮
X_n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}

雷达图也称蜘蛛网图。其作图步骤如下：划一个圆。由p个点把圆周分成相等的p个部分。由圆心连接这p个点，得p个辐射状的半径。这p个半径就作为p个变量的坐标轴。根据各变量波动的大小，将相应的坐标轴标上刻度。将每个变量的值点在各自的坐标轴上。依次连接起来得一p边形。对每个样本都这样做，可得n个p边形，形成一个雷达图。

例1、表1·2列出了邓阜仙岩体的部分化学成分。试用雷达图来表示它们。

运用上述作雷达图的步骤得图1·1。从图上看到：X_1 与 X_2；X_3，X_4 与 X_5；X_6 与 X_7 图形分别相似，可以分成三类。这正好是三种类型的花岗岩。

表1·2　邓阜仙岩体化学成分（部分）

岩　　体	SiO_2	TiO_2	FeO	CaO	K_2O
X_1：中粗粒斑状黑云母花岗岩	75·20	0·14	1·86	0·91	5·21
X_2：中粗粒斑状黑云母花岗岩	75·15	0·16	2·11	0·74	4·93
X_3：中粒二云母花岗岩	72·19	0·13	1·52	0·69	4·65
X_4：中粒二云母花岗岩	72·35	0·13	1·37	0·83	4·87
X_5：中粒二云母花岗岩	72·74	0·10	1·41	0·72	4·99
X_6：细粒白云母花岗岩	73·29	0·033	1·07	0·17	3·15
X_7：细粒白云母花岗岩	73·72	0·033	0·77	0·28	2·78

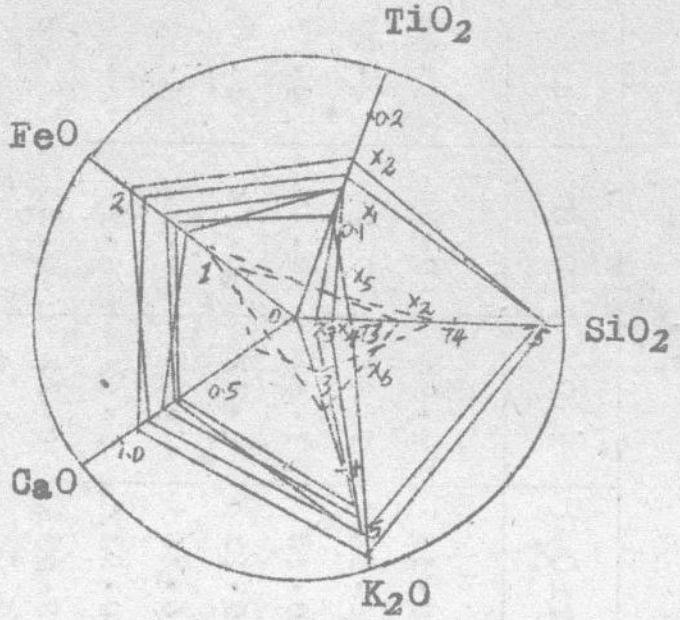

图1·1 雷达图

在雷达图上，每个样本对应一个P边形，当P较大或样本较多时，图上线段太多，看起来很不方便，这时可以每个样本画一张图，如图1·2。如果样本的分类已很清楚，这时可以将同类的计算均值，用每类的均值来作雷达图，这时图像既清晰，每类特征又很明显。例如对例1，我们求得三类的均值列表1·3，用三组均值作成雷达图1·3，我们看到图像清晰，三类特点一目了然。

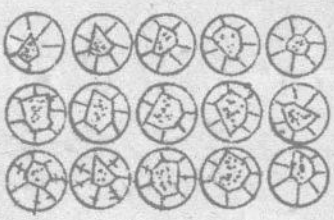

图1·2

表1·3 类均值

类	SiO_2	TiO_2	FeO	CaO	K_2O
I: $\{X_1, X_2\}$	75·18	0·15	1·99	0·83	5·07
II: $\{X_3, X_4, X_5\}$	72·43	0·12	1·43	0·75	4·84
III: $\{X_6, X_7\}$	73·51	0·033	0·92	0·23	2·97

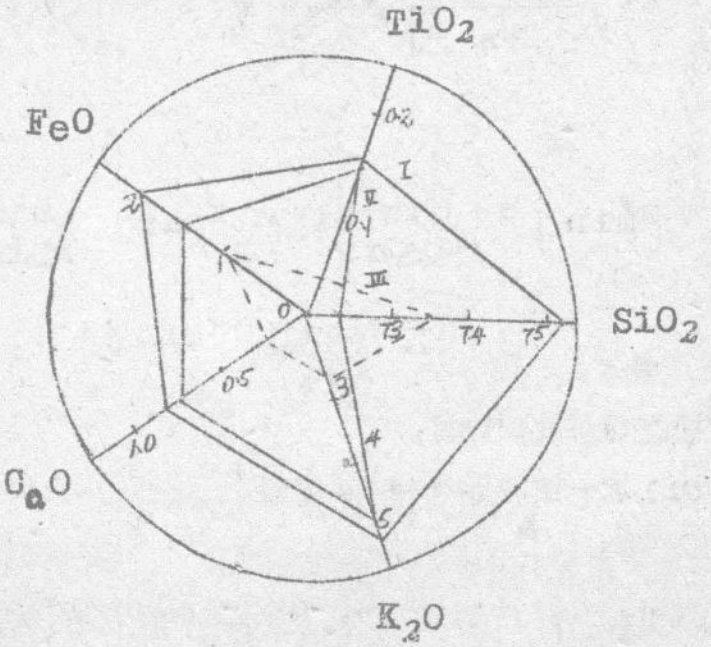

图1·3 均值的雷达图

二、星座图(constellation graph)

上节我们看到，当n很大时，用雷达图来表达每一个样本是不方便的，这时可以用星座图来描述。所谓星座图，就是将样本

点在一个半圆内。一个样本用一颗星来表示。同类的样本组成一个星座。不同类样本组成不同的星座。很象天文学上表示星座的图象。

星座图的作图步骤如下：

(i)先将数据$\{x_{ij}\}$作一线性变换。使变换后的数据$\{\varepsilon_{ij}\}$落到$[0,\pi]$内。常取

$$\varepsilon_{ij}=\frac{x_{ij}-x_{minj}}{R_j}\pi, \qquad (1)$$

其中

$$x_{minj}=\min_{1\le i\le n} x_{ij},\quad x_{maxj}=\max_{1\le i\le n} x_{ij},$$

$$R_j=x_{maxj}-x_{minj}.$$

这就是极差标准化的办法。

(ii)取一组权$\{\omega_j\}$。使得

$$\omega_j\ge 0,\ j=1,2,\cdots,p,\ \sum_{j=1}^{p}\omega_j=1.$$

重要的变量相应的权可取得大一点。或者也可以取等权。即

$$\omega_j=\omega_2=\cdots=\omega_p=1/p.$$

(iii) 画一个半径为1的上半圆及半圆底边的直径。使每个样本对应一个星。这些星就落在这个半圆内。比如为求X_1对应的星。先以圆心O为圆心。ω_1为半径划一上半圆。在这半圆周上对应弧度为ε_{11}的点记作O_1。接着以O_1为心、ω_2为半

径作一上半圆。在这半圆周上对应弧度为ξ_{12}的点记作O_2；再以O_2为心、ω_3为半径作一上半圆。取弧度为ξ_{13}的点，记为O_3；…依次一直求到O_p。这最后的O_p就是X_1所对应的星的位置，记作z_1。由于$\sum\omega_j=1$，$\omega_j\geq 0$，星必然落在半圆内。由O点通过上述作图步骤，到达星的路线（即$OO_1O_2\cdots O_p$）称做该星的路径。通过星的位置和路径就全面的刻划了该样本的特征。对其余样本作图是类似的，详见图2·1。

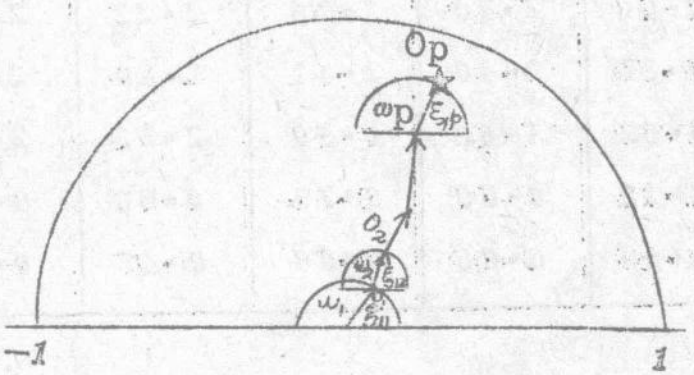

图2·1 星座图

通过作图来求路径和星的位置是不方便的，最好借助于计算来实现。不难看出X_a对应的O_1，O_2，…O_{p-1}，O_p的坐标为

$$\left(\sum_{j=1}^{l}\omega_j\cos\xi_{aj},\ \sum_{j=1}^{l}\omega_j\sin\xi_{aj}\right),\quad l=1,2,\cdots,p,$$

决定了O_1，O_2，…，O_p的位置后，不难用计算机来打图。

现在我们仍用例1来说明星座图的应用。由表1·2算得$\{\xi_{ij}\}$列于表2·1，并取

$$\omega_1=\omega_2=\omega_3=\omega_4=\omega_5=1/5。$$

由此作成图2·2。我们看到组成了三个星座。对应于三类岩体。似乎比雷达图表达得更为直观和醒目。

表2·1 ξ_{ij}值

	SiO_2	TiO_2	FeO	CaO	K_2O
X_1	3·14	2·65	2·56	3·14	3·14
X_2	3·09	3·14	3·14	2·42	2·78
X_3	0·00	2·40	1·76	2·21	2·42
X_4	0·17	2·40	1·41	2·80	2·70
X_5	0·57	1·66	1·50	2·33	2·86
X_6	1·15	0·00	0·70	0·00	0·48
X_7	1·60	0·00	0·00	0·47	0·00

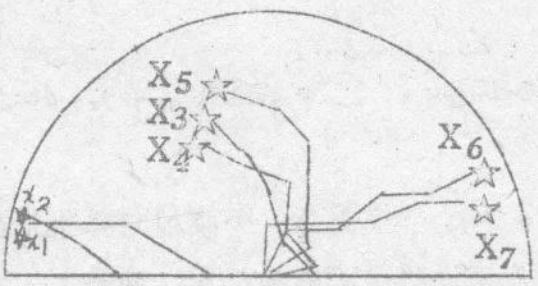

图2·2

当各类分得很清楚（如图2·2）时，只需画星的位置不一定画相应的路径，使图更加清晰明了。但有时也会出现异途同归的情况。星的路径很不一样，而星的位置很接近，这时如在图中去掉路径，就会发生误解。为了能够去掉路径，最好适当调整$\{\omega_i\}$，使得异途不能同归。

星座图最初是由Wakimoto和Taguri[3]提出来的，并在医学、行为科学、保险公司、学校教育等方面取得不少应用。笔者在国内介绍后，已在地质、气象、考古等方面得到了应用。如下例是关于考古学的。

例2、为了研究人类的进化，测了现代男人和古代雄猿的下颚和牙齿的十七个部位，有关数据来自[4]。作成星座图2·3（图中·表示一颗星），其中编号1－20的为人，21－38的为猿，人全落在半圆的右边，猿在左边，两类泾渭分明。

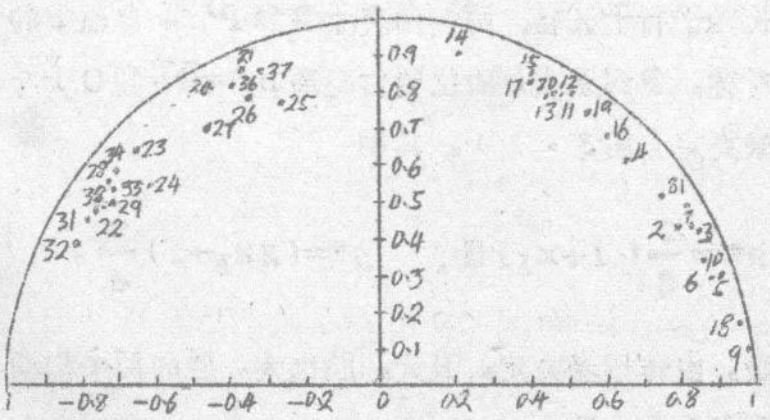

图2·3

三、脸谱图（face graph）

脸谱是一个非常形象的东西。脸谱的胖瘦、喜怒哀乐给人留下深刻的印象。京剧就在脸谱的设计上下了很大功夫。用脸谱来表达样本首先是由统计学家 Chernoff[1] 提出来的。他将样本的特性用人脸的某一部位的形状或大小来表达。一个样本用一个脸谱来表达。相似的样本，它们构成的脸谱也很相象。Chernoff 首先把脸谱图用于聚类分析之中。日本很快的引进了脸谱图，并用于多元分析的各种应用之中。

脸谱图各部分用 18 个变量（$x_1, x_2, \cdots, x_{18}$）来构成。当变量数 $P < 18$ 时，可将脸中某几个部位固定。当 $P > 18$ 时，可以设法在脸中再添加一些部位。脸谱由六部分来构成：脸的轮廓、鼻、嘴、眼、眼珠和眉。现在分别来介绍每一部分的画法。

1、脸的轮廓：它由上下两个椭圆来构成。它们的短轴均在 Y 轴上。长轴平行于 X 轴。两椭圆交于 P 和 P′，自然 P 和 P′ 关于 Y 轴对称。P 到原点 O 的位置由距离 $h^* = \overline{OP}$ 和 OP 与 X 轴夹角 θ^* 来决定（图 3·1）。其中

$$h^* = \frac{1}{2}(1+x_1)H, \quad \theta^* = (2x_2 - 1)\frac{\pi}{4},$$

H 为一常数，由作图者决定，H 大，脸也大。脸的两个椭圆与 Y 轴分别交于 U 和 L，且

$$h \triangleq \overline{OU} = \overline{OL}, \quad h = \frac{1}{2}(1+x_3)H$$

由上述条件还不能唯一确定脸的两个椭圆，进一步规定，脸

的上半部椭圆的离心率为x_4。下半部椭圆的离心率为x_5。

我们知道，椭圆的标准方程为

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1,$$

a，b分别为它的长轴和短轴。离心率e的定义是

$$e=\sqrt{a^2-b^2}/a,$$

a，b，e中只要知道了其中任两个就可决定另一个。利用离心率，椭圆的标准方程可以表为

$$(1-e^2)x^2+y^2=b^2. \qquad (2)$$

给了上述条件后如何来作脸的两个椭圆呢？现以下半椭圆来说明之。记下半椭圆的长轴，短轴分别为a和b。记$d\cong h-b$，P点的坐标为（X^*，Y^*），显见

$$X^*=h^*\cos\theta^* \quad Y^*=h^*\sin\theta^*.$$

椭圆中心的坐标为（0，$-d$），故椭圆方程为

$$x^2(1-e^2)+(y+d)^2=b^2=(h-d)^2$$

它过P点，应有

$$X^{*2}(1-e^2)+(Y^*+d)^2=(h-d)^2,$$

解得

$$d=\frac{(h^2-Y^{*2})-(1-e^2)X^{*2}}{2(Y^*+h)}$$

e为离心率（$=x_5$），上式h，Y^*，X^*，e均为已知，故可求出

d。从而便可得到脸下半部的椭圆方程。上半部的椭圆方程求法是类似的。

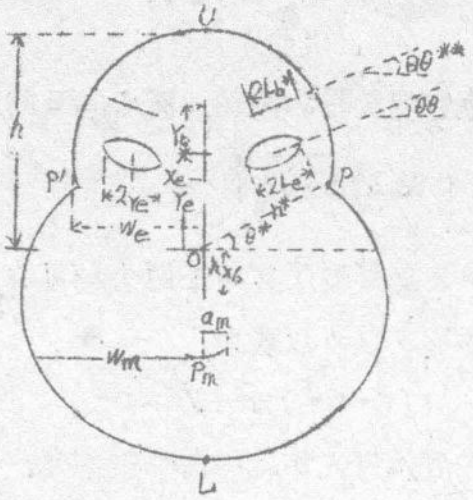

图3-1　脸谱图

2、鼻：以O为中心在Y轴上。上下各取长度hx_6 画一条粗线。

3、嘴：在O点下方。$h[x_7+(1-x_7)x_6]$ 的位置。用半径为$h/|x_8|$的圆弧来描述。规定：x_8 为正。圆弧向上；x_8 为负。圆弧向下。嘴的大小由a_m 来决定。$a_m=x_9(h/|x_8|)$。嘴的圆弧关于Y轴对称。如果嘴太大。越过脸的轮廓。这时用$x_9 w_m$ 来代替。其中w_m 为点P_m 到脸轮廓的水平距离。

4、眼：用两个椭圆来表达。椭圆的中心为：(x_e, y_e)。$(-x_e, y_e)$。其中规定

$$y_e=h[x_{10}+(1-x_{10})x_6],\quad x_e=w_e(1+2x_{11})/4,$$

其中w_e 为点$(0, y_e)$至脸轮廓的水平距离。两眼椭圆的长轴

与X轴的夹角分别为 θ 和 $\pi-\theta$。其中 $\theta=(2x_{12}-1)\pi/5$，其长轴为 Le。$Le=x_{14}\min(xe, we-xe)$。椭圆的离心率为 x_{13}。

5、眼珠：从眼的椭圆中心，沿着椭圆的长轴至 $\pm r_e(2x_{15}-1)$ 的位置，其中

$$r_e=(\cos^2\theta+\sin^2\theta/x_{13}^2)^{-1/2}Le。$$

r_e 为眼的长轴在X轴上投影的长度。在作图时要注意务使眼珠为左右对称。

6、眉：从眼的椭圆中心向上至 y_b 的高度决定了眉的中心，其长度为 $2L_b$。它与水平的夹角为 θ^{**}，其中

$$\theta^{**}=\theta+2(x_{17}-1)\pi/5，\quad y_b=2(x_{16}+0.3)Lex_{13}，$$

$$L_b=r_e(2x_{18}+1)/2。$$

表3·1 脸谱图各变量的意义

变 量	描述脸的变量	变 换 公 式	数 值 范 围	备 注
x_1	h^*	$h^*=\frac{1}{2}(1+x_1)H$	0—1	OP之长度，H为脸的大小比例数
x_2	θ^*	$\theta^*=(2x_2-1)\pi/4$	0—1	X轴与OP的角度
x_3	h	$h=\frac{1}{2}(1+x_3)H$	0—1	OU=OL之长度
x_4	x_4		0·2—0·8	脸的上半椭圆离心率
x_5	x_5		0·2—0·8	脸的下半椭圆离心率
x_6	x_6	$2hx_6$	0·1—0·7	鼻的长度
x_7	P_m	$P_m=h[x_7+(1-x_7)x_6]$	0—1	嘴的位置
x_8	x_8		—5—5	嘴的曲率（半径$=h/|x_8|$）
x_9	a_m	$a_m=x_9(h/|x_8|)$或x_9w_m	0—1	嘴的大小
x_{10}	y_e	$y_e=h[x_{10}+(1-x_{10})x_6]$	0—1	眼的位置（纵坐标）
x_{11}	x_e	$x_e=w_e(1+2x_{11})/4$	0—1	眼的位置（横坐标）
x_{12}	θ	$\theta=(2x_{12}-1)\pi/5$	0—1	眼的倾斜角
x_{13}	x_{13}		0·4—0·8	眼的椭圆离心率
x_{14}	L_e	$L_e=x_{14}\min(x_e, w_e-x_e)$	0—1	眼的长轴之长度
x_{15}	x_{15}		0—1	眼珠的位置
x_{16}	y_b	$y_b=2(x_{16}+0·3)L_ex_{13}$	0—1	眼到眉的高度
x_{17}	θ^{**}	$\theta^{**}=\theta+2(1-x_{17})\pi/5$	0—1	眉的倾斜角
x_{18}	L_b	$L_b=r_e(2x_{18}+1)/2$	0—1	眉的长度（$2L_b$）