

多元分析资料汇编

(VI)

中国科学院应用数学研究所

北京 友谊宾馆

1980.5

目 录

多指标样本的点图法	方开泰, 刘璋温	1
一、雷达图		1
二、星座图		4
三、脸谱图		9
四、连结向量图		14
五、三角多项式图		20
六、降维作图法		21
七、其它方法		23
参考文献		23
非参数回归与稳健回归	陈希孺	25
4.1 引言		25
4.2 权函数方法		29
4.3 若干应用		48
4.4 理论证明		57
4.5 稳健回归		79
参考文献		96
线性回归分析	方开泰	99
第一章 随机向量		100
§ 1.1 记号		100
§ 1.2 期望和协差阵的运算		101
§ 1.3 二次型的均值和方差		104
§ 1.4 随机变量的独立性		111
§ 1.5 χ^2 分布		112
第二章 多元正态分布		113
§ 2.1 定义		113

§ 2.2	正态分布的一些性质	116
§ 2.3	正态变量的二次型	120
第三章	线性回归的估计和分布理论	125
§ 3.1	最小二乘估计	125
§ 3.2	最小二乘估计的性质	129
§ 3.3	σ^2 的估计	130
§ 3.4	分布理论	133
§ 3.5	广义最小二乘	135
§ 3.6	增加回归因子的回归	138
§ 3.7	增加试验的回归	142
§ 3.8	有约束回归	145
第四章	假设检验	149
§ 4.1	F 检验	149
§ 4.2	全相关系数	159
§ 4.3	H 的典型形	163
第五章	选择“最优的”变因子集	165
§ 5.1	引言	165
§ 5.2	产生一切可能的回归	166
§ 5.3	产生的方法	168
§ 5.4	优良准则	175
§ 5.5	只产生比较好的回归	194
一类带约束的回归	方开泰, 王东谦, 吴国富	206
一. 问题		206
二. 主要结果		207
三. 计算方法		217
四. 应用		219
参考文献		225

多指标样本的点图法

方开泰 刘璋温

图形是帮助人们思维和判断的重要工具，当样本只有两个指标（变量）时，可以用通常的直角坐标在平面上点图，当样本有三个指标时，虽然可以在三维的笛卡儿坐标里点图，但也是很很不方便的。当样本的指标大于三时，用通常的方法已不能点图了，而在多元分析中，一般指标数均大于三，探讨多指标的点图法是长期来一直为人们所关注的研究课题，这里介绍一些有关方法，特别是近十年来发见的一些方法。

雷达图 (radar chart)

设样本有 p 个变量 $x_1, x_2, \dots, x_p$ ，有 n 个样本 $x_1, x_2, \dots, x_n$ ，用 x_{ij} ($i=1, 2, \dots, n; j=1, 2, \dots, p$) 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。

雷达图的作图步骤是：以 O 为原点，适当的长度画一个圆，然后将圆周分成相等的 p 等分，过圆心 O 连接它们得到 p 个辐射状的半径，这 p 个半径就作为 p 个变量的坐标轴，根据数据波动的大小，将每根坐标轴上刻度。

现在可以将每个样品表示在图上了，顺序连接样本在每根坐标轴上相应的坐标，得 p 边形，这个 p 边形的形状就表达了这个样品的特性。

表 1.1 数据矩阵

指标 样本	x_1	x_2	$\dots$	x_j	$\dots$	x_p
x_1	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
x_2	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_n	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}

例 1：表 1.2 列出了邓阜仙岩体的 7 种化学成分，试用雷达图来表示它们。

表 1.2 邓阜仙岩体化学成分 (百分)

岩 体	SiO ₂	TiO ₂	FeO	CaO	K ₂ O
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 : - - - - -	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 : - - - - -	72.35	0.13	1.37	0.83	4.87
X ₅ : - - - - -	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 : 细粒二云母花岗岩	73.72	0.033	0.77	0.28	2.78

运用上述作雷达图的步骤，得图 1.1。从图上看到，X₁ 与 X₂，X₃，X₄ 与 X₅，X₆ 与 X₇ 图形相似，可以分成三类，这正好是三种类型的花岗岩。由于雷达图上每个样本都有 p 个边（此例 $p=5$ ），当样品很多时，图上线段太多，看起来很不方便，这时可以将同类的计算均值，用每类的均值来作雷达图，这时图象既清晰，每类特征又很明显。

例如对例 1，我们求得三类的均值列于表 1.3，用三组均值作成雷达图 1.2，我们看到图象清晰，三类的特点一目了然。

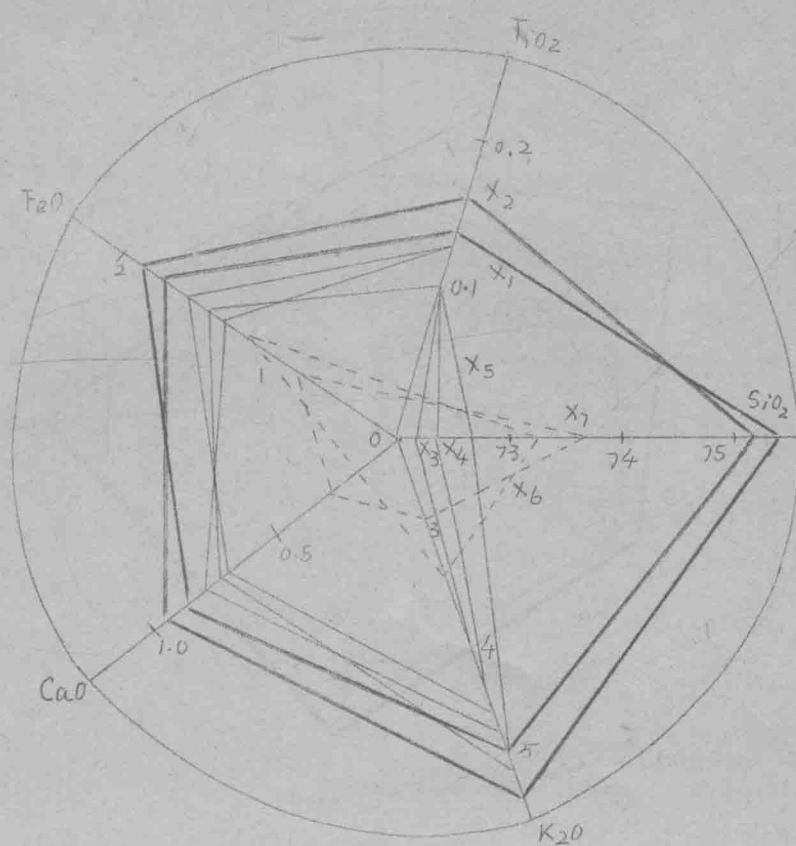


图 1.1 雷达图

表 1.3 类均值

类	SiO ₂	TiO ₂	FeO	CaO	K ₂ O
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

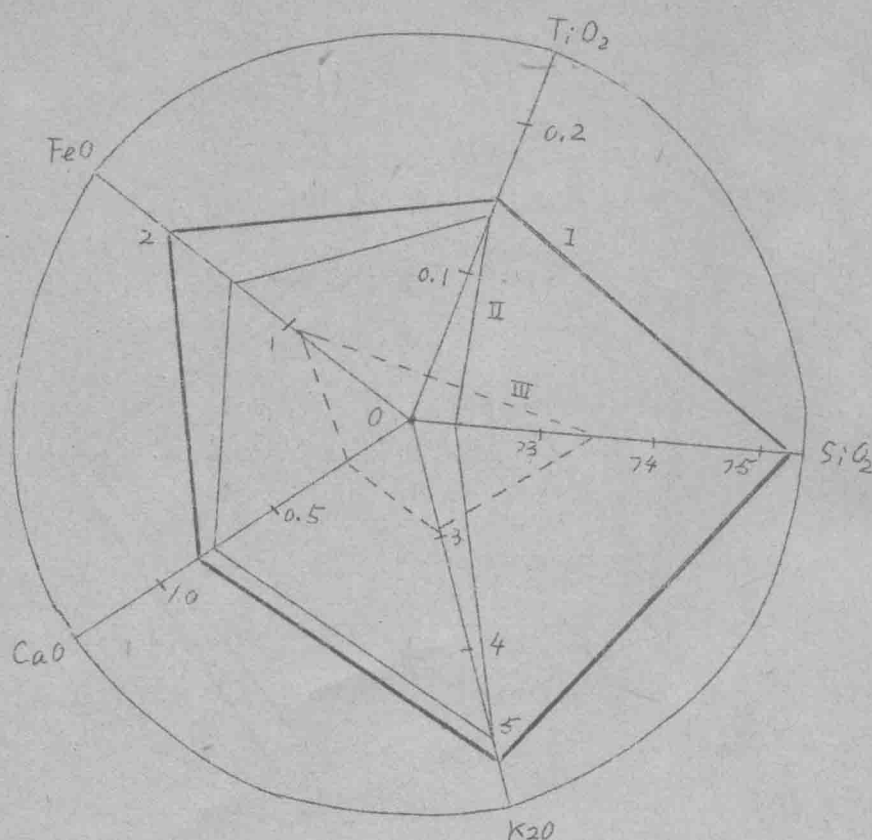


图 1.2 类均值雷达图.

二. 星座图 (Constellation graph).

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, 很象天文学上表示星座的图象。

星座图在作图之前先将数据 $\{x_{ij}\}$ 作一变换, 使其取值范围落到 $[0, \pi]$, 即取适当的函数 $f_j(\cdot)$ 使

$$\begin{cases} \bar{x}_{ij} = f_j(x_{ij}), & i = 1, 2, \dots, n \\ 0 \leq \bar{x}_{ij} \leq \pi & j = 1, 2, \dots, p \end{cases}$$

常取

$$f_j(x_{ij}) = \frac{x_{ij} - x_{\min j}}{x_{\max j} - x_{\min j}} \pi \quad (1)$$

其中

$$X_{\max j} = \max_{1 \leq i \leq n} X_{ij}, \quad X_{\min j} = \min_{1 \leq i \leq n} X_{ij} \quad (2)$$

相应于每个变易取一个权 $w_j \geq 0$, 使得 $\sum_{j=1}^p w_j = 1$ 。

有了上述准备工作之后, 星座图的作法如下: 以 O 为圆心, 以 1 为半径, 作一上半圆, 并作相应于这半圆的直径。所有的星(样本)就落在这个圆内。每个样本对应于一个星, 比如为了求 x_1 的星, 先以 O 为圆心, w_1 为半径作一上半圆, 在圆上对应弧度为 β_{11} 的点记作 O_1 ; 再以 O_1 为心, w_2 为半径作一上半圆, 在弧上对应弧度为 β_{12} 的点记作 O_2 ; 再以 O_2 为心, w_3 为半径作上半圆, 取弧度为 β_{13} 的点记作 $O_3 \cdots$ 最后一直到 O_{p-1} , 这一点就代表了 x_1 的星座, 记作 β_1 。对其余的样本作法是类似的。由 O 点通过上述作图步骤, 到达星座的路线(即 $OO_1O_2 \cdots O_{p-1}$) 称做该星的路径, 通过星的位置和路径就会刻画到了每个样本的特征。

通过作图来求路径和星座是不方便的, 应当通过计算来实现, 如将 O 点看成原点, 半径看成复半径, 则 x_α 的星座 β_α 为

$$\beta_\alpha = \sum_{j=1}^p w_j e^{i\beta_{\alpha j}} \quad (3)$$

其中 $i = \sqrt{-1}$, 因 $e^{i\beta_{\alpha j}} = \cos \beta_{\alpha j} + i \sin \beta_{\alpha j}$, β_α 的实部和虚部分别为

$$\sum_{j=1}^p w_j \cos \beta_{\alpha j} \quad \text{和} \quad \sum_{j=1}^p w_j \sin \beta_{\alpha j}$$

β_2 的路径为

~ 6 ~

$$\sum_{j=1}^l \omega_j \cos \beta_{xj} \quad \text{和} \quad \sum_{j=1}^l \omega_j \sin \beta_{xj} \quad (l=1, 2, \dots, p)$$

现在我们仍用例 1 来说明星座图的应用, 由表 1.2 算得 $\{\beta_{ij}\}$ 的值列于表 2.1, 并取 $\omega_1 = \omega_2 = \omega_3 = \omega_4 = \omega_5 = 1/5$, 由它作成图 2.1, 我们看到三类分得很清, 似乎比雷达图表达得更为直观和醒目。

星座图最初是由 Wakimoto, K and Taguri, M (1978) 提出来的, 并在医学, 行动科学, 保险公司, 中学教育等方面作了不少应用。

表 2.1 β_{ij} 值

	SiO ₂	TiO ₂	FeO	CaO	K ₂ O
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

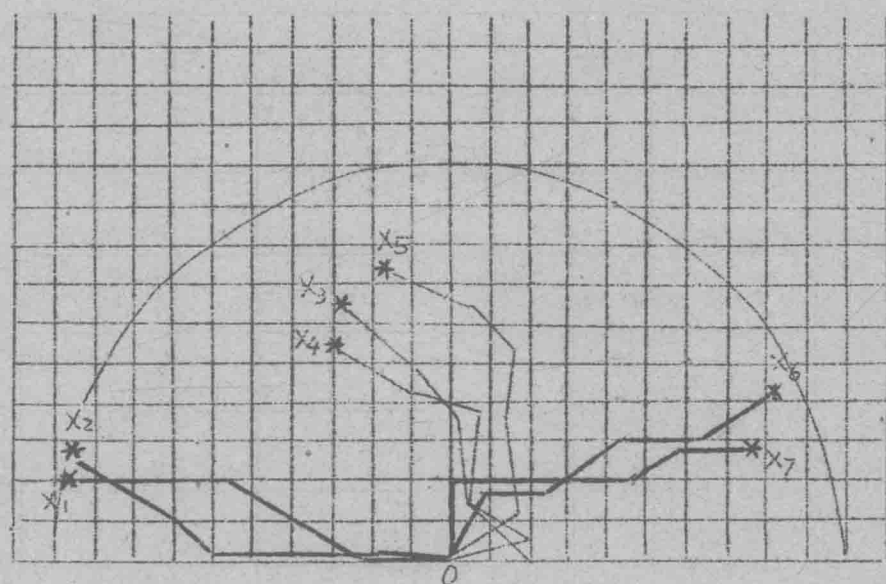


图 2.1 星座图

例2. 为了研究人类的进化, 测了现代男人和古代猿猴的下颌和牙齿的十七个印位, 有关数据可参看王令红、方开泰、郑玉颖(1979)运用星座图得图2.2, 其中编号1~20为人, 21~38为猿。我们看到一个在图的右半, 另一个在左半, 可谓泾渭分明。

(图2.2 见下页)。

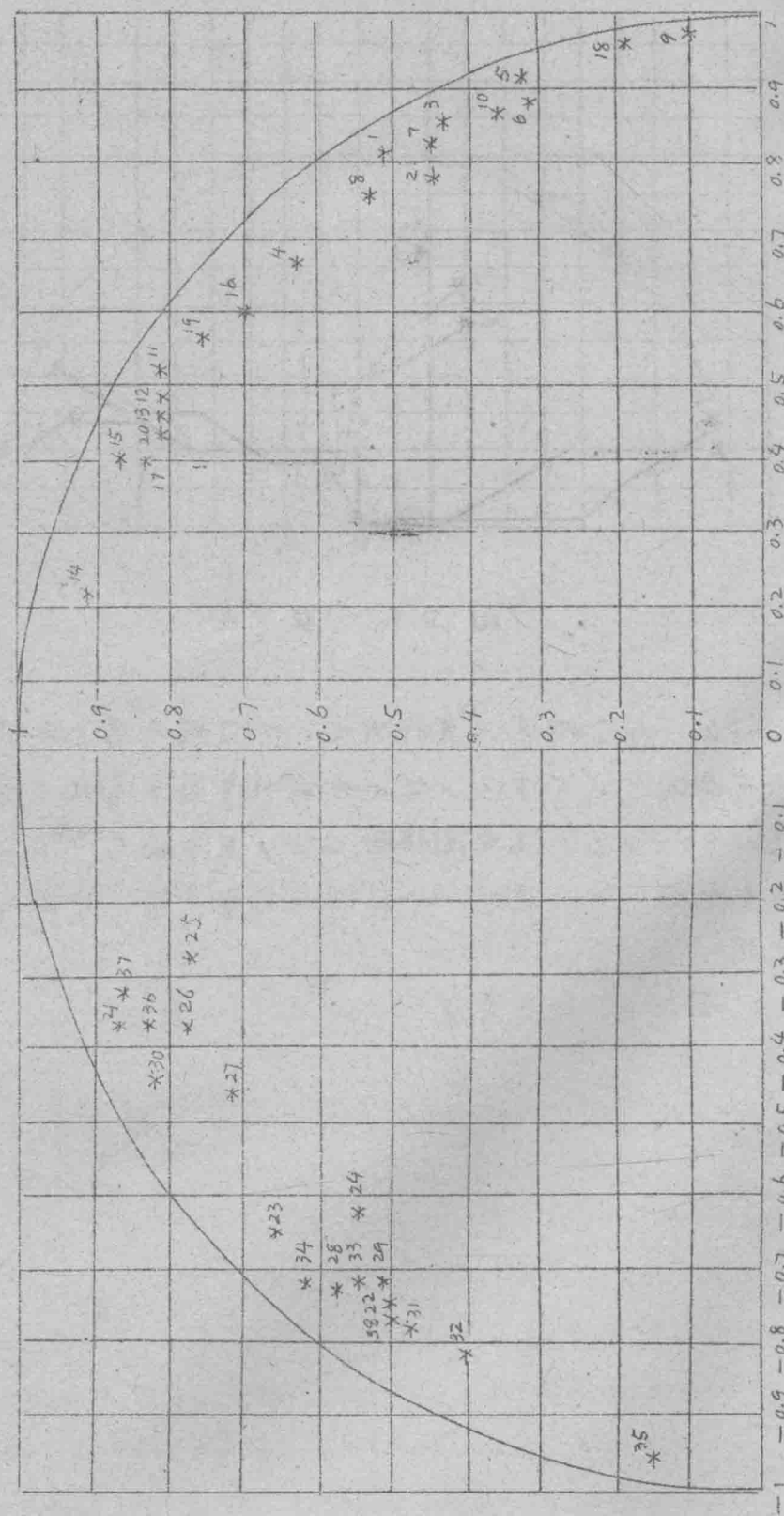


图 2.2

三. 脸谱图 (face graph)

将每个指标用人脸的某一下位的形状或大小来表达, 通过 n 个指标的大小就构成一个脸谱, 如果样本比较相似, 它们构成的脸谱也很相象, 反之, 如样本的性质相差甚远, 则各自的脸谱就很不一样, 由于脸谱是一种很形象的东西, 给人会留下较深的印象。

脸谱图各下位分用 18 个变量 ($x_1, x_2, \dots, x_{18}$) 来构成, 当变量数 $p < 18$ 时, 可将脸中某几个下位固定, 当 $p > 18$ 时, 自己可以在脸中再添加一些下位。脸谱由六下分来构成: 脸的轮廓、鼻、嘴、眼、眼珠和眉。现在分别来介绍每一下位的画法。

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, 两椭圆交于 P 和 P' , P 和 P' 关于 y 轴对称, P 到原点 O 的位置由 h^* 和角 θ^* (参图 3.1) 来决定, 而 O 是脸的中心也是鼻的中心。其中 $h^* = \frac{1}{2}(1+x_1)H$, H 的大小是由作画者决定的, H 大脸也大。 OP 和水平线的角度 $\theta^* = (2x_2-1)\pi/4$, 从脸的中心到头的上下的距离 $OU (= OL)$ 为 h , $h = \frac{1}{2}(1+x_3)H$ 。

脸的上半 $P'UP$ 的椭圆的离心率为 x_4 , 脸下半 PUP 的离心率为 x_5 。

2. 鼻: 以 O 为中心的垂直线 (Y 轴) 上下各取 $h x_6$ 长。

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, 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。如果口太大, 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 $y_e = h[x_6 + (1-x_{10})x_6]$ 和 $\pm x_e = \pm w_e(1+2x_{11})/4$, 长轴与水平线的夹

~10 ~

角为 θ 和 $\pi - \theta$ ，其中 $\theta = (2\chi_{12} - 1)\pi/5$ ，眼的长轴为 L_e ， $L_e = \chi_{14} \min(\chi_e, w_e - \chi_e)$ ，其中 w_e 为 y 轴上由刻度为 y_e 的点至脸轮廓的水平距离，眼的椭圆的离心率为 χ_{13} 。

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

$$\gamma_e = (\cos^2 \theta + \sin^2 \theta / \chi_{13}^2)^{-1/2} L_e$$

在作图时，注意眼珠是左右对称的。

b. 眉：从眼的椭圆中心向上至 y_b 的高度，这点是眉的中心，其长度为 $2L_b$ ，它与水平的夹角为 θ^{**} ，其中 $\theta^{**} = \theta + 2(\chi_{17} - 1)\pi/5$ ， $y_b = 2(\chi_{16} + 0.3)L_e \chi_{13}$ ， $L_b = \gamma_e(2\chi_{18} + 1)/2$ 。

由于直接用实际的数据来画图，经常使脸谱失去通常的志义，为此，需要将每个变易通过一个线性变换变到予定的范围。

设数据如表 1.1 所规定的记号，如第 j 个变易要变化到 (a_j, b_j) 内，则令

$$x'_{ij} = a_j + (b_j - a_j) \frac{x_{ij} - x_{\min j}}{x_{\max j} - x_{\min j}} \quad i = 1, 2, \dots, n$$

则 $\{x'_{ij}\}$ 已落入 (a_j, b_j) 内。上述 18 个变易的取值范围列入表 3.1，对某一个具体问题，一般比表中的范围还取得小一点。

例：表 3.2 列举了 20 名球员的九项技战术指标，希用脸谱图来表达它们的特点。为了与脸谱图的变易取得一致，表中变易的编号不是

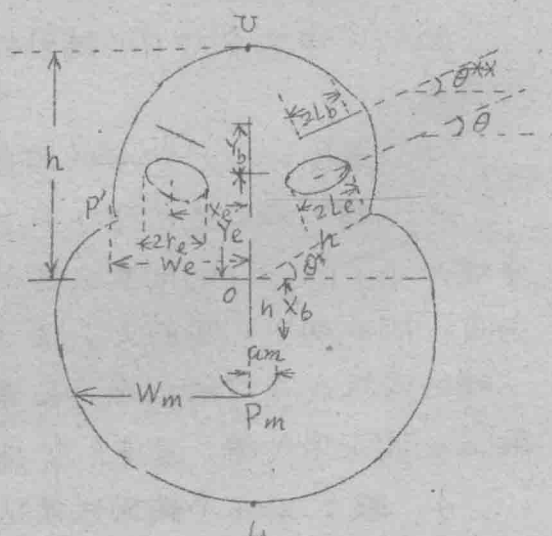


图 3-1

按自然顺序的。

由于脸谱图需要 18 个变量，现仅 9 个变量，因此，让其中的九个变量采取定值，取值情况为：

表 3.1 脸谱图各变量的定义

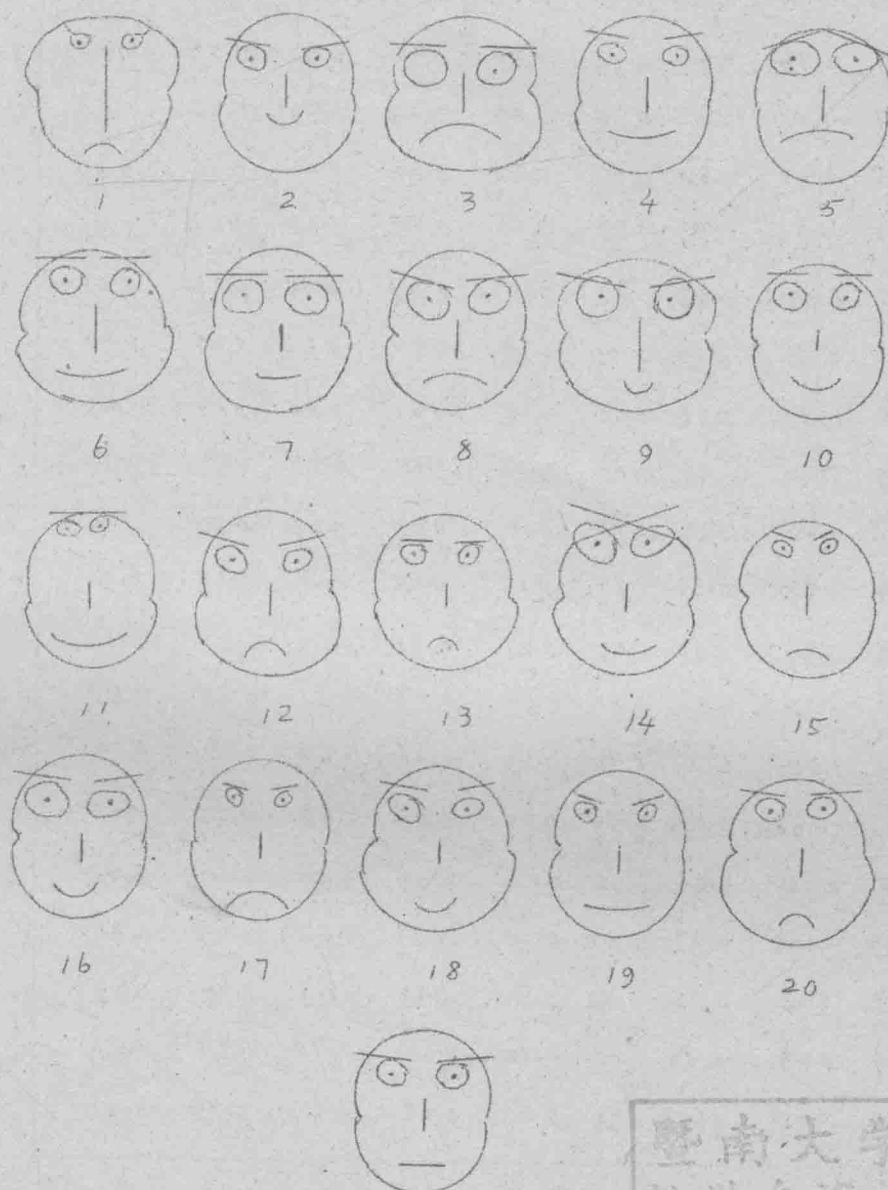
变量	描述脸的变量	变换公式	取值范围	备 注
X_1	h^*	$h^* = \frac{1}{2}(1+X_1)H$	0—1	OP 之长, H 为脸的大小比例数
X_2	θ^*	$\theta^* = (2X_2-1)\pi/4$	0—1	X 轴与 OP 的角度
X_3	h	$h = \frac{1}{2}(1+X_3)H$	0—1	OU = OL 之长
X_4	X_4		0.2—0.8	脸的上半椭圆离心率
X_5	X_5		0.2—0.8	脸的下半椭圆离心率
X_6	X_6	$2hX_6$	0.1—0.7	鼻的长
X_7	p_m	$p_m = h(X_7 + (1-X_7)X_6)$	0—1	嘴的位置
X_8	X_8		-5—5	嘴的曲率(半径 = $h/(1-X_8)$)
X_9	a_m	$a_m = X_9(h/(1-X_8))$ 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(纵座标)
X_{11}	x_e	$x_e = w_e(1+2X_{11})/4$	0—1	眼的位置(横座标)
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
X_{15}	X_{15}		0—1	眼珠的位置
X_{16}	y_b	$y_b = 2(X_{16} + 0.5)L_eX_{13}$	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	$L_b = y_e(2X_{18}+1)/2$	0—1	眉的长($2L_b$)

表 3.2 球员考核指标

球 员	x_7	x_8	x_{16}	x_{14}	x_6	x_{10}	x_{12}	x_5	x_4
1	2.98	0.467	0.017	0.000	0.441	0.211	0.0437	0.0468	0.215
2	2.99	0.692	0.163	0.143	0.056	0.241	0.0376	0.0556	0.162
3	3.11	0.500	0.200	0.300	0.125	0.214	0.0274	0.1001	0.223
4	3.26	0.571	0.706	0.167	0.091	0.248	0.0319	0.0617	0.189
5	3.32	0.533	0.238	0.400	0.000	0.256	0.0124	0.0838	0.106
6	3.32	0.567	0.381	0.188	0.154	0.251	0.0217	0.0929	0.205
7	3.35	0.556	0.146	0.333	0.029	0.244	0.0298	0.0445	0.161
8	3.61	0.526	0.146	0.333	0.029	0.246	0.0382	0.0603	0.145
9	3.70	0.789	0.087	0.250	0.190	0.228	0.0339	0.0925	0.261
10	3.75	0.625	0.211	0.250	0.000	0.261	0.0294	0.0671	0.125
11	3.80	0.600	0.125	0.200	0.000	0.303	0.0308	0.0454	0.080
12	3.81	0.433	0.390	0.125	0.080	0.236	0.0415	0.0898	0.167
13	3.82	0.313	0.118	0.000	0.000	0.261	0.0289	0.0856	0.165
14	3.92	0.629	0.300	0.500	0.000	0.278	0.0393	0.0486	0.118
15	3.93	0.519	0.316	0.000	0.077	0.270	0.0424	0.0344	0.088
16	3.94	0.667	0.240	0.333	0.000	0.251	0.0351	0.0876	0.159
17	3.96	0.500	0.095	0.000	0.000	0.270	0.0424	0.0639	0.073
18	3.97	0.700	0.324	0.182	0.000	0.255	0.0339	0.0947	0.121
19	3.98	0.571	0.161	0.000	0.000	0.263	0.0436	0.0719	0.170
20	4.09	0.393	0.192	0.200	0.071	0.271	0.0313	0.0948	0.146

$x_1=0.8$, $x_2=0.5$, $x_3=1$, $x_9=0.5$, $x_{11}=0.5$, $x_{13}=0.6$,
 $x_{15}=0.5$, $x_{17}=1$, $x_{18}=0.8$; 其余 9 个变量的取值范围为 x_4 ,
 x_5 , $x_7(0.2, 0.8)$, $x_6(0.1, 0.7)$, $x_8(-5, 5)$, x_{10} , $x_{16}(0.1, 0.9)$

$x_{12}(0.3, 0.7)$, $x_{14}(0.3, 0.9)$ 。将表 3.2 的数据变换到所规定的范围，变换后的数据列于表 3.3，由这些数据作图得图 3.2。



总平均

图 3.2 脸谱图

表 3.3 变换后作脸谱图的数据

球员	x_7	x_8	x_{16}	x_{14}	x_6	x_{10}	x_{12}	x_5	x_4
1	0.20	-1.77	0.10	0.30	0.70	0.10	0.70	0.31	0.65
2	0.21	2.96	0.41	0.47	0.18	0.36	0.62	0.39	0.48
3	0.27	-1.07	0.49	0.66	0.27	0.13	0.49	0.80	0.68
4	0.35	0.43	0.82	0.50	0.22	0.42	0.55	0.45	0.57
5	0.39	-0.37	0.57	0.78	0.10	0.49	0.30	0.65	0.31
6	0.39	0.33	0.88	0.52	0.31	0.45	0.42	0.73	0.62
7	0.40	0.10	0.78	0.70	0.14	0.38	0.52	0.29	0.48
8	0.54	-0.52	0.38	0.70	0.14	0.41	0.63	0.44	0.43
9	0.59	5.00	0.25	0.60	0.36	0.25	0.57	0.73	0.80
10	0.62	1.55	0.52	0.60	0.10	0.53	0.52	0.50	0.37
11	0.64	1.03	0.33	0.54	0.10	0.90	0.54	0.30	0.22
12	0.65	-2.47	0.90	0.45	0.21	0.32	0.67	0.71	0.53
13	0.65	-5.00	0.32	0.30	0.10	0.53	0.51	0.67	0.50
14	0.71	1.55	0.71	0.90	0.10	0.68	0.64	0.33	0.34
15	0.71	-0.68	0.74	0.30	0.20	0.62	0.68	0.20	0.25
16	0.72	2.43	0.58	0.70	0.10	0.45	0.59	0.69	0.47
17	0.73	-1.07	0.27	0.30	0.10	0.61	0.68	0.47	0.20
18	0.74	3.12	0.76	0.52	0.10	0.48	0.57	0.75	0.35
19	0.74	0.43	0.41	0.30	0.10	0.56	0.70	0.54	0.51
20	0.80	-3.32	0.48	0.54	0.20	0.62	0.54	0.75	0.43

四. 连接向量图 (linked vector graph)

上述三种作图都是对 p 个平等的变量而言的, 有时要表达 p 个自变量 $x_1, x_2, \dots, x_p$ 对因变量 y 的相关关系, 当然