主成分分析

**一、主成分的几何意义与一般数学模型**

**1、主成分分析的几何意义；**  
　　例中的的数据点是六维的；也就是说，每个观测值是 6 维空间中的一个点。我们希望把 6 维空间用低维空间表示。

先假定只有二维，即只有两个变量，它们由横坐标和纵坐标所代表；因此每个观测值都有相应于这两个坐标轴的两个坐标值；如果这些数据形成一个椭圆形状的 点阵（这在变量的二维正态的假定下是可能的），那么这个椭圆有一个长轴和一个短轴。在短轴方向上，数据变化很少；在极端的情况，短轴如果退化成一点，那只 有在长轴的方向才能够解释这些点的变化了；这样，由二维到一维的降维就自然完成了。

当坐标轴和椭圆的长短轴平行，那么代表长轴的变量就描述了数据的主要变化，而代表短轴的变量就描述了数据的次要变化。但是，坐标轴通常并不和椭圆的长 短轴平行。因此，需要寻找椭圆的长短轴，并进行变换，使得新变量和椭圆的长短轴平行。如果长轴变量代表了数据包含的大部分信息，就用该变量代替原先的两个 变量（舍去次要的一维），降维就完成了。椭圆（球）的长短轴相差得越大，降维也越有道理。对于多维变量的情况和二维类似，也有高维的椭球，只不过无法直观地看见罢了。

　　首先把高维椭球的主轴找出来，再用代表大多数数据信息的最长的几个轴作为新变量；这样，主成分分析就基本完成了。  
 注意，和二维情况类似，高维椭球的主轴也是互相垂直的。这些互相正交的新变量是原先变量的线性组合，叫做主成分 (principal component) 。  
 正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。  
 选择越少的主成分，降维就越好。什么是标准呢？那就是这些被选的主成分所代表的主轴的长度之和占了主轴长度总和的大部分。有些文献建议，所选的主轴总长度占所有主轴长度之和的大约 85% 即可，其实，这只是一个大体的说法；具体选几个，要看实际情况而定。

**2、主成分的基本性质**

①各主成分之间互不相关，若原变量服从正态，则各主成分之间互相独立;  
　　②全部ｍ个主成分所反映的ｎ例样品的总信息，等于ｍ个原变量的总信息。信息量的多少，用变量的方差来度量。若将ｍ个原变量标准化后，每个变量的方差都为１，故方差之和为ｍ，此时，求得的ｍ个主成分的方差之和也为ｍ;  
　　③各主成分的作用大小是∶Z1≥Z2≥…≥Zm;  
　　④第ｉ个主成分的贡献率是(λi／ｍ)×100％;  
　　⑤前P个主成分的累计贡献率是((∑Ｐi＝1λi)／ｍ)×100％。在应用时,一般取累计贡献率为70～85％或以上所对应的前P个主成分即可。 在资料所含的变量个数、样品数及累计贡献率固定的前提下，P／ｍ的比值越小，则说明此资料用主成分分析越合适。

　　⑥r(Zi,xj)=cij，说明第i个主成分Zi与第j个标准化变量xj之间的相关系数就是表达式(3)中的系数cij;  
　　⑦∑ｍj＝1r2(Zi，xj)＝λi，说明第ｉ个主成分Zi与ｍ个标准化变量中的每一个变量之间的相关系数的平和为由大到小排列后的第ｉ个特征值λi;  
　　⑧∑ｍi＝1r2(Zi，xj)=１,说明ｍ个主成分分别与第ｊ个标准化变量的相关系数的平和为１，即每１个标准化变量的信息由全部主成分完全包含。

**3、proc princomp过程**

使用proc princomp过程进行主成分分析，其主要语句格式如下：

Proc princomp <选项列表>；

Var 变量列表；

Run;

其中：

（1）Proc princomp语句用来规定输入、输出和一些运行选项，其选项及功能如下：

①data=数据集名1：指明所要分析的数据集，若省略则表示分析最新生成的数据集。

②out=数据集名2：命名一个输出SAS数据集，其中包含原始数据以及各主成分得分（即各主成分的观测值）。

③outstat=数据集名3：命名一个包含各变量的均值、标准差、相关矩阵或协方差矩阵、特征值和特征向量的SAS数据集。

④covariance(cov)：要求从协方差矩阵出发作主成分分析，若省略此选项，则从相关矩阵出发进行分析；除非各变量的度量单位是可比较的或已经过某种方式的标准化，否则不宜使用此选项，应从相关矩阵出发作主成分分析。

⑤N=n：指定要计算的主成分的个数，其默认值为参与分析的变量的个数。

⑥prefix=name：规定各主成分名称的前缀。省略此句则SAS系统自动赋予各主成分名称分别为prin1,prin2,……；若“name=A”，则各主成分名称分别为A1，A2，……，前缀的字符个数加上后面数字位数应不超过8个字符。

（2）VAR variables;

此句中的“variables”部分列出数据集中参与主成分分析的变量名称。若省略此句，则被分析数据集中所有数值变量均参与分析。

实例

(中学生身体四项指标的主成分分析) 在某中学随机抽取某年级30名学生，测量其身高（X1）、体重（X2）、胸围（X3）和坐高（X4），（数据见下列程序）。试对这30名学生体四项指标数据做主成分分析。

**data** d721;

input number x1-x4 @@ ;

\*\*label x1='身高' x2='体重' x3='胸围' x4='坐高';

cards;

1 148 41 72 78 2 139 34 71 76

3 160 49 77 86 4 149 36 67 79

5 159 45 80 86 6 142 31 66 76

7 153 43 76 83 8 150 43 77 79

9 151 42 77 80 10 139 31 68 74

11 140 29 64 74 12 161 47 78 84

13 158 49 78 83 14 140 33 67 77

15 137 31 66 73 16 152 35 73 79

17 149 47 82 79 18 145 35 70 77

19 160 47 74 87 20 156 44 78 85

21 151 42 73 82 22 147 38 73 78

23 157 39 68 80 24 147 30 65 75

25 157 48 80 88 26 151 36 74 80

27 144 36 68 76 28 141 30 67 76

29 139 32 68 73 30 148 38 70 78

;

**\*\*从相关系数矩阵出发做主成分分析 ;**

**proc** **princomp** data=d721 prefix=z out=o721 ;

var x1-x4;

**run**;

**\*\*从协方差矩阵出发做主成分分析 ;**

**proc** **princomp** data=d721 cov prefix=z out=o721 ; /\*cov：要求从协方差矩阵出发作主成分分析，若省略此选项，则从相关矩阵出发进行分析；\*/

var x1-x4;

**run**;

options ps=**32** ls=**85**; /\*表示输出屏幕定义为一页32行,每行85字符\*/

**proc** **plot** data=o721;

plot z2\*z1 $ number='\*'/href=-**1** href=**2** vref=**0**;

**run**;

/\* plot语句用来作图， z2\*z1 表示以z2作为纵轴变量，z1作为横轴变量.

$后面接标签变量 ，此处以number 作为标签，用”\*”标出其对应点；

href=和vref=表示画一条垂直于坐标轴横轴值（-1,2,）、纵轴数值（0）的基准线

该图形主要是用来初步的查看样本点的分布情况。

\*/

**proc** **sort** data=o721;

by z1;

**run**;

**proc** **print** data=o721;

var number z1 z2 x1-x4;

**run**;

**quit**;

**从R进行主成分分析：**

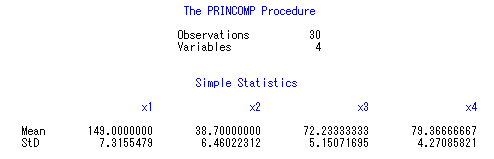


图1 描述性统计量

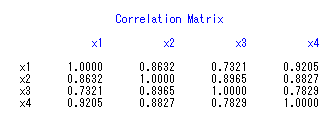
****

图2 样本相关矩阵

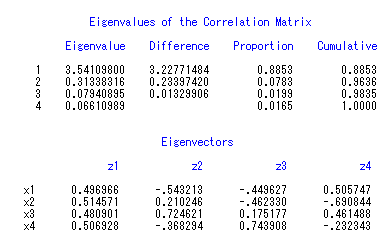


图3 样本相关矩阵的特征值与特征向量

由输出中相关阵的特征值可以看出,第一主成分的贡献率已高达88.53%;且前两个主成分的累计贡献率已达96.36%.因此只需要用前两个主成分就能很好地概括这一组数据.

另由第四个特征值（0.06610989）近似为0，可以看出这4个标准化后的身体指标变量有近似的共线性。由最大的两个特征值对应的特征向量可以写出第一和第二主成分：



利用特征向量的值对各个主成分进行分析。第一主成分特征值对应的第一个特征向量的各个分量均在0.5附近，而且都是正值，它反映学生的魁梧程度。身体高大的学生，他的4个部位的尺度都比较大；而身体矮小的学生，他的4个部位的尺寸都比较小。因此我们可以称第一主成分为大小因子。第二大特征值对应的特征向量中第一和第四个分量均为负值，其它的都为正值，它反映学生的胖瘦情况，可称为形态因子。

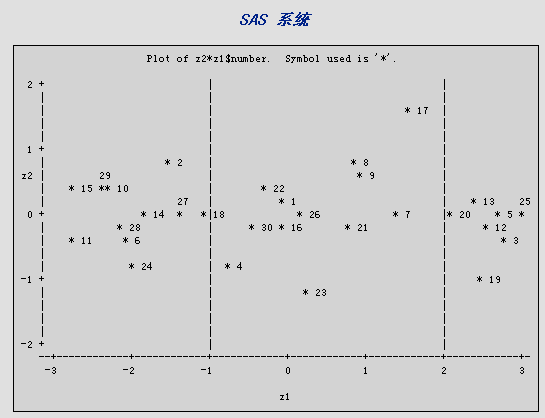


图4 第二主成分得分对第一主成分得分的散布点图

从图中可以直观地看出,按学生的身体指标尺寸,这30名学生大致可以分成三组（以第一主成分得分值为分界点).

每一组包含哪几名学生由每个散点旁边的序号可以得知,更详细的信息可以从下面的数据输出列表中得到:第11，15，29，10，28，6，24，14，2，27，18号样本为一组(第一主成分得分值<-1)；第4，30，22，1，16，26，23，21，8，9，7，17为第二组(第一主成分得分值介于-1与2之间)；第20，13，19，12，5，3，25为第三组(第一主成分得分值>2）。

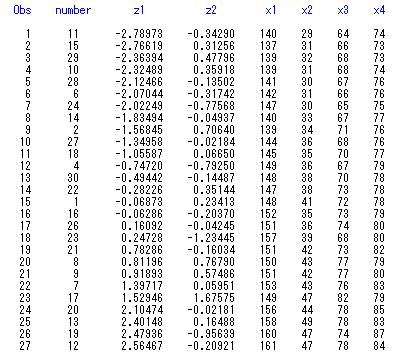


图5 按第一主成分得分排序后的主成分得分和原始数据

**从S进行主成分分析得到的结果分析方法与上面类似**