第5章 主成分分析

主成分分析（principal components analysis）也称主分量分析，是由霍特林于1933年首先提出的。主成分分析是利用降维的思想，在损失很少信息的前提下，把多个指标转化为几个综合指标的多元统计方法。通常把转化生成的综合指标称为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，使得主成分比原始变量具有某些更优越的性能。这样在研究复杂问题时就可以只考虑少数几个主成分而不至于损失太多信息，从而更容易抓住主要矛盾，揭示事物内部变量之间的规律性，同时使问题得到简化，提高分析效率。

人们往往要考虑与其有关系的多个指标，这些指标在多元统计中也称为变量。这样就产生了如下问题：一方面人们为了避免遗漏重要的信息而考虑尽可能多的指标，另一方面考虑指标的增多增加了问题的复杂性，同时由于各指标均是对同一事物的反映，不可避免地造成信息的大量重叠，这种信息的重叠有时甚至会掩盖事物的真正特征与内在规律。基于上述问题，人们就希望在定量研究中涉及的变量较少，而得到的信息量又较多。主成分分析正是研究如何通过原来变量的少数几个线性组合来解释原来变量绝大多数信息的一种多元统计方法。

既然研究某一问题涉及的众多变量之间有一定的相关性，就必然存在着起支配作用的共同因素。根据这一点，通过对原始变量相关矩阵或协方差矩阵内部结构关系的研究，利用原始变量的线性组合形成几个综合指标（主成分），可以在保留原始变量主要信息的前提下起到降维与简化问题的作用，使得在研究复杂问题时更容易抓住主要矛盾。一般来说，利用主成分分析得到的主成分与原始变量之间有如下基本关系：

（1）每一个主成分都是各原始变量的线性组合。

（2）主成分的数目大大少于原始变量的数目。

（3）主成分保留了原始变量的绝大多数信息。

（4）各主成分之间互不相关。

实际上主成分分析的主要目的是希望用较少的变量去解释原来资料中的大部分变异，亦即期望能将手中许多相关性高的变量转化成彼此互相独立的变量，能由其中选取较原始变量个数少的，能解释大部分资料变异的几个新变量，也就是所谓的主成分，而这几个主成分也就成为我们用来解释资料的综合性指标。

5.1.2 主成分分析的基本理论

设对某一事物的研究涉及 p 个指标，分别用 X_1，X_2，…，X_p 表示，这 p 个指标构成的 p 维随机向量为 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$。设随机向量 $\boldsymbol{X}$ 的均值为 $\boldsymbol{\mu}$，协方差矩阵为 $\boldsymbol{\Sigma}$。

对 $\boldsymbol{X}$ 进行线性变换，可以形成新的综合变量，用 $\boldsymbol{Y}$ 表示，也就是说，新的综合变量可以由原来的变量线性表示，即满足下式：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\ Y_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\ \quad \vdots \\ Y_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p \end{cases} \tag{5.1}$$

当不对 u_i 进行限制时，$Var(Y_i)$ 会任意增大，因此线性变换约束在下面条件

1. $u_i'u_i = 1$ for $i = 1, 2, \cdots, p$.
2. Y_i 与 Y_j 相互无关
3. Y_1 是 $X_1, X_2, \cdots, X_p$ 线性组合中方差的最大者

   Y_2 是与 Y_1 不相关的 $X_1, X_2, \cdots, X_p$ 所有线性组合中的方差最大者

由于可以任意地对原始变量进行上述线性变换，由不同的线性变换得到的综合变量 $\boldsymbol{Y}$ 的统计特性也不尽相同。因此为了取得较好的效果，我们总是希望 $Y_i = \boldsymbol{u}_i'\boldsymbol{X}$ 的方差尽可能大且各 Y_i 之间互相独立，由于

$$\mathrm{var}(Y_i) = \mathrm{var}(\boldsymbol{u}_i'\boldsymbol{X}) = \boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i$$

而对任意的常数 c，有

$$\mathrm{var}(c\boldsymbol{u}_i'\boldsymbol{X}) = c^2\boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i$$

几何意义

这样，我们就对主成分分析的几何意义有了一个充分的了解。主成分分析的过程无非就是坐标系旋转的过程，各主成分表达式就是新坐标系与原坐标系的转换关系，在新坐标系中，各坐标轴的方向就是原始数据变差最大的方向。

$$\begin{cases} Y_1 = X_1\cos\theta + X_2\sin\theta \\ Y_2 = -X_1\sin\theta + X_2\cos\theta \end{cases}$$

其矩阵形式为：

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \boldsymbol{U}\boldsymbol{X}$$

式中，$\boldsymbol{U}$ 为旋转变换矩阵，由上式可知它是正交阵，即满足

$$\boldsymbol{U}' = \boldsymbol{U}^{-1}, \quad \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$$

经过这样的旋转之后，N 个样品点在 Y_1 轴上的离散程度最大，变量 Y_1 代表了原始数据的绝大部分信息，这样，有时在研究实际问题时，即使不考虑变量 Y_2 也无损大局。因此，经过上述旋转变换就可以把原始数据的信息集中到 Y_1 轴上，对数据中包含的信息起到了浓缩的作用。主成分分析的目的就是找出变换矩阵 $\boldsymbol{U}$，而主成分分析的作用与几何意

由此进一步可知，主成分分析是把 p 个随机变量的总方差 $\sum_{i=1}^{p}\sigma_{ii}$ 分解为 p 个不相关的随机变量的方差之和，使第一主成分的方差达到最大。第一主成分是以变化最大的方向向量的各分量为系数的原始变量的线性函数，最大方差为 λ_1。$\alpha_1 = \dfrac{\lambda_1}{\sum\lambda_i}$ 表明了最大方差占总方差的比值，称 α_1 为第一主成分的贡献率。这个值越大，表明 Y_1 这个新变量综合 X_1，X_2，…，X_p 信息的能力越强，即由 Y_1 的差异来解释随机向量 $\boldsymbol{X}$ 的差异的能力越强。正因如此，才把 Y_1 称为 $\boldsymbol{X}$ 的主成分，进而我们就更清楚为什么主成分是按特征根 λ_1，λ_2，…，λ_p 取值的大小排序的。

进行主成分分析的目的之一是减少变量的个数，所以一般不会取 p 个主成分，而是取 m（$m<p$）个主成分。m 取多少比较合适，是一个很实际的问题，通常以所取 m 使得累积贡献率达到 85%以上为宜，即

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geqslant 85\% \tag{5.4}$$

这样，既能使信息损失不太多，又能达到减少变量、简化问题的目的。另外，选取主成分还可根据特征根的变化来确定。图 5－2 为 SPSS 统计软件生成的碎石图。

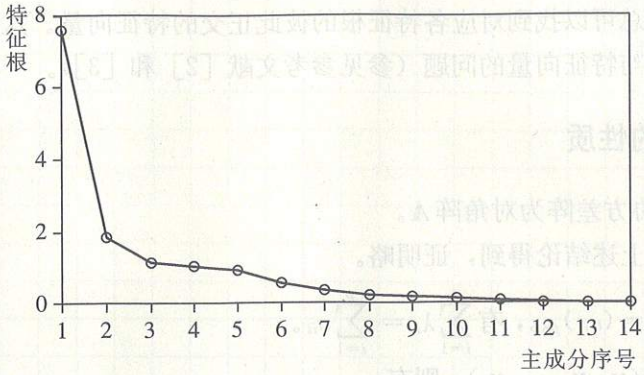

图 5－2　碎石图

由图 5－2 可知，第二个及第三个特征根变化的趋势已经开始趋于平稳，所以，取前两个或前三个主成分是比较合适的。采用这种方法确定的主成分个数与按累积贡献率确定的主成分个数往往是一致的。在实际应用中，有些研究者习惯于保留那些特征根大于 1 的主成分，但这种方法缺乏完善的理论支持。在大多数情况下，$m=3$ 即可使所选主成分保持信息总量的比重达到 85%以上。

定义 5.2　第 k 个主成分 Y_k 与原始变量 X_i 的相关系数 $\rho(Y_k, X_i)$ 称为因子负荷量。

因子负荷量是主成分解释中非常重要的解释依据，因子负荷量的绝对值大小刻画了该主成分的主要意义及其成因。在下一章中还将对因子负荷量的统计意义给出更详细的解释。由下面的性质我们可以看到，因子负荷量与系数向量成正比。

性质 3　$\rho(Y_k, X_i)=\gamma_{ik}\sqrt{\lambda_k}/\sqrt{\sigma_{ii}}$，　$k,i=1,2,\cdots,p$　　(5.5)

证明：$\sqrt{\mathrm{var}(Y_k)}=\sqrt{\lambda_k}$，$\sqrt{\mathrm{var}(X_i)}=\sqrt{\sigma_{ii}}$

令 $\boldsymbol{e}_i=(0,\cdots,0,1,0,\cdots,0)'$ 为单位向量，则

$$X_i=\boldsymbol{e}_i'\boldsymbol{X}$$

又　　$$Y_k=\boldsymbol{\gamma}_k'\boldsymbol{X}$$

于是　　$$\mathrm{cov}(Y_k,X_i)=\mathrm{cov}(\boldsymbol{\gamma}_k'\boldsymbol{X},\boldsymbol{e}_i'\boldsymbol{X})=\boldsymbol{e}_i'D(\boldsymbol{X})\boldsymbol{\gamma}_k=\boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{\gamma}_k=\lambda_k\boldsymbol{e}_i'\boldsymbol{\gamma}_k=\lambda_k\gamma_{ik}$$

$$\rho(Y_k,X_i)=\frac{\mathrm{cov}(Y_k,X_i)}{\sqrt{\mathrm{var}(Y_k)}\sqrt{\mathrm{var}(X_i)}}=\frac{\gamma_{ik}\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}$$

由性质 3 知，因子负荷量 $\rho(Y_k, X_i)$ 与系数 γ_{ik} 成正比，与 X_i 的标准差成反比关系，因此，绝不能将因子负荷量与系数向量混为一谈。在解释主成分的成因或第 i 个变量对第 k 个主成分的重要性时，应当根据因子负荷量而不能仅仅根据 Y_k 与 X_i 的变换系数 γ_{ik}。

一般而言，对于度量单位不同的指标或取值范围彼此差异非常大的指标，不直接由其协方差矩阵出发进行主成分分析，而应该考虑将数据标准化。比如，在对上市公司的财务状况进行分析时，常常会涉及利润总额、市盈率、每股净利率等指标，其中利润总额取值常常从几十万元到上百万元，市盈率取值一般在5到六七十之间，而每股净利率在1以下，不同指标取值范围相差很大，这时若是直接从协方差矩阵入手进行主成分分析，利润总额将明显起到重要支配作用，而其他两个指标的作用很难在主成分中体现出来，此时应该考虑对数据进行标准化处理。

但是，对原始数据进行标准化处理后倾向于各个指标的作用在主成分的构成中相等。对于取值范围相差不大或度量相同的指标进行标准化处理后，其主成分分析的结果仍与由协方差阵出发求得的结果有较大区别。这是由于对数据进行标准化的过程实际上也就是抹杀原始变量离散程度差异的过程，标准化后的各变量方差相等，均为1，而实际上方差也是对数据信息的重要概括，也就是说，对原始数据进行标准化后抹杀了一部分重要信息，因此才使得标准化后各变量在对主成分构成中的作用趋于相等。由此看来，对同度量或取值范围在同量级的数据，还是直接从协方差矩阵求解主成分为宜。

对于从什么出发求解主成分，现在还没有一个定论，但是我们应该看到，不考虑实际情况就对数据进行标准化处理或者直接从原始变量的相关矩阵出发求解主成分是有其不足之处的，这一点需要注意。建议在实际工作中分别从不同角度出发求解主成分并研究其结果的差别，看看是否存在明显差异且这种差异产生的原因在何处，以确定哪种结果更为可信。

5.4.2 主成分分析不要求数据来自正态总体

由上面的讨论可知，无论是从原始变量协方差矩阵出发求解主成分，还是从相关矩阵出发求解主成分，均不涉及总体分布的问题。也就是说，与很多多元统计方法不同，主成分分析不要求数据来自正态总体。实际上，主成分分析就是对矩阵结构的分析，其中用到的主要是矩阵运算的技术及矩阵对角化和矩阵的谱分解技术。我们知道，对多元随机变量而言，其协方差矩阵或相关矩阵均是非负定的，这样，就可以按照求解主成分的步骤求出其特征根、标准正交特征向量，进而求出主成分，达到缩减数据维数的目的。同时，由主成分分析的几何意义可以看到，对来自多元正态总体的数据，我们得到了合理的几何解释，即主成分就是按数据离散程度最大的方向进行坐标轴旋转。

主成分分析的这一特性大大扩展了其应用范围，对多维数据，只要是涉及降维的处理，我们都可以尝试用主成分分析，而不用花太多精力考虑其分布情况。

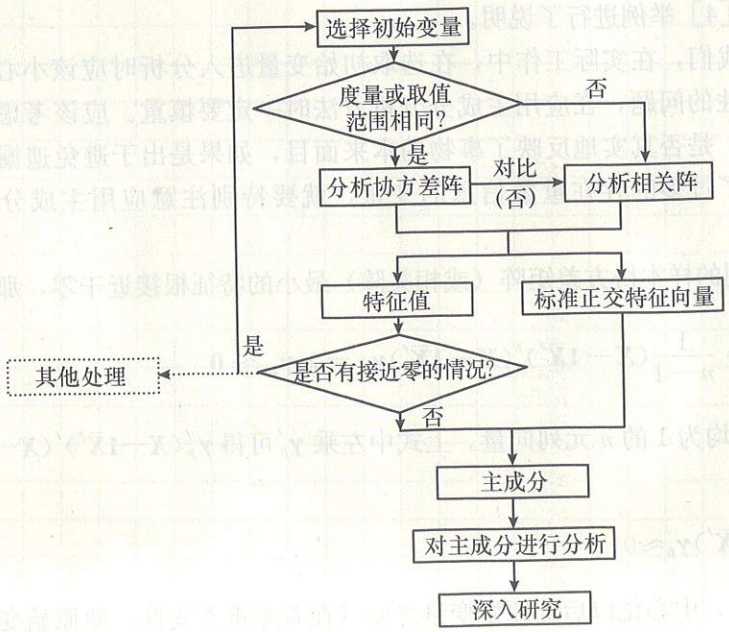

图5-3 主成分分析的逻辑框图

方差贡献率

（1）方差贡献

设 $\Sigma=(\sigma_{ij})_{p\times p}$，于是有 $\sum_{i=1}^{p}\sigma_{ii}=\sum_{i=1}^{p}\lambda_i$

即 $\sum_{i=1}^{p}Var(y_i)=\sum_{i=1}^{p}Var(x_i)$，也就是说，主成分把 p 个原始变量的总方差分解成 p 个不相关的新变量的方差之和。主成分分析就是为了减少变量的个数，忽略一些较小方差的主成分将不会给总方差带来大的影响。

（2）方差贡献率

定义 $\lambda_k/\sum_{i=1}^{p}\lambda_i$ 为第 k 个主成分 y_k 的方差贡献率（Proportion of Variance），第一个主成分的贡献率最大，表明 y_1 综合原始变量 $x_1,x_2,\cdots,x_p$ 的能力最强，而 $y_2,y_3,\cdots,y_p$ 的综合能力依次递减。

（3）方差累积贡献率

若只取 $m(<p)$ 个主成分，则称 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i$ 为主成分 $y_1,y_2,\cdots,y_m$ 的累积方差贡献率（Cumulative Proportion）。它表明 $y_1,y_2,\cdots,y_m$ 综合 $x_1,x_2,\cdots,x_p$ 的能力，通常取 m 使得累积贡献率不低于80%（一些文献中也认为只要特征根 λ_i 大于1即可）。

23.1.1　主成分分析原理

假设原始观测数据包括 p 个变量，每一个变量的数据由 n 次观测组成，其中 n 就是样本大小。

$$\begin{matrix} x_{11}, x_{21}, \cdots, x_{p1} \\ x_{12}, x_{22}, \cdots, x_{p2} \\ \vdots \\ x_{1n}, x_{2n}, \cdots, x_{pn} \end{matrix}$$

由 p 个变量组成的数据样本可看作是由 p 个向量 $X_1, X_2, \cdots, X_p$ 组成的数据，每个向量对应于数据表中的每一列。

$$X_i = (x_{i1}, x_{i2}, \cdots, x_{in}) \text{ 其中 } (i = 1, 2, \cdots, p)$$

要计算 p 个变量中任意两个变量 X_i，X_j 之间的协方差矩阵，可根据如下公式算出：

$$\mathrm{Cov}(X_i, X_j) = \frac{1}{n-1}\sum_{l=1}^{n}(x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_i) \quad \text{其中} \quad \bar{x}_i = \frac{1}{n}\sum_{l}^{n} x_{il} (l = 1, 2, \cdots, n)$$

而计算 p 个变量中任意两个变量 X_i, X_j 之间的相关系数矩阵，则可以根据如下公式算出：

$$r(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}} = \frac{\sum_{l=1}^{n}(x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{n}(x_{il} - \bar{x}_i)^2}\sqrt{\sum_{l=1}^{n}(x_{jl} - \bar{x}_j)^2}}$$

标准化数据计算出来的协方差矩阵跟未标准化数据计算出来的相关系数矩阵是等价的，也就是说协方差 $\mathrm{Cov}(X_i, X_j)$ 除以变量 X_i 的标准差 $\sqrt{\mathrm{Var}(X_i)}$ 和 X_j 的标准差 $\sqrt{\mathrm{Var}(X_j)}$，能消除了两个变量 X_i，X_j 的量纲影响。可以说相关系数矩阵是一种数据标准化后的特殊协方差矩阵，标准化数据计算出来的标准差为 1，对应上述公式中的分母为 1。

主成分分析通过对协方差矩阵或相关系数矩阵可求解特征值 λ 和特征向量 U，主成分分析在数学上就是一个正交线性变换，其本质就是将数据从变量空间变换到一个新的正交坐标系中，并且使数据的任何投影的第一大方差落在第一主成分轴 F_1 上，第二大方差落在第二个主成分轴 F_2 上，依此类推（见图 23-2）。因此主成分具有 3 个特性：各个主成分彼此独立，即协方差 $\mathrm{Cov}(F_i, F_j)=0$；对任一主成分 F_j，它对应各变量系数的平方和 $\sum_{i=1}^{p} u_{ij}^{\ 2} = 1$；第 1 主成分到第 p 主成分的方差贡献依次递减。

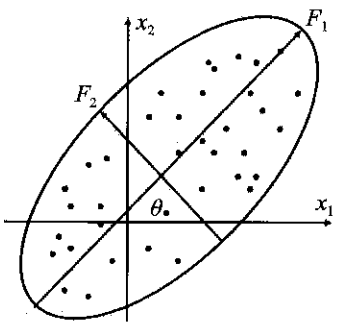

图 23-2　变量空间与主成分空间关系

主成分分析求解出的特征值 λ_i 之和恰好是原 p 个随机变量的方差之和，说明主成分分析将原来的 p 个随机变量的方差之和分解成了新的 p 个不相关的主成分方差贡献，每一个特征值 λ_i 就是主成分 F_i 对总体方差的贡献，因此 $\lambda_i / \sum_{i=1}^{p}\lambda_i$ 被称为主成分 F_i 的贡献率。前 k 个主成分的累积方差贡献等于 $\sum_{i=1}^{k}\lambda_i$，则它在全部方差中的累积贡献率为 $\sum_{i=1}^{k}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i$。它说明这 k 个主成分能够反映原来 p 个变量所表示的信息量的百分比。实践中通常要求选入的 k 个主成分，其方差累计贡献率必须在 80% 以上。

构造出来的主成分和原变量之间满足关系 $F=U'X$，U' 为矩阵 U 的转置矩阵，矩阵 $U=(u_1, u_2, \cdots, u_p)=[u_{ij}]$ 其中 $i, j=1, 2, \cdots, p$。则第 j 个主成分 F_j 与原变量之间满足如下线性组合关系：

$$F_j = \sum_{i=1}^{p} u_{ij} X_i \quad \text{其中 } j=1,2,\cdots,m, n \leqslant p$$

而反过来，原变量与主成分之间则有 $X=UF$，原变量与主成分的相关程度取决于对应的线性组合系数的大小：

$$\rho(X_i, F_j) = \mathrm{Cov}(X_i, F_j) \Big/ (\sigma_i \sqrt{\lambda_j}) = u_{ij}\lambda_j \Big/ (\sigma_i \sqrt{\lambda_j}) = u_{ij}\sqrt{\lambda_j} \Big/ \sigma_i$$

其中原变量 X_i 的方差为 $\sigma_i^2 = \sum_{j=1}^{p} u_{ij}{}^2 \lambda_j$，$u_{ij}{}^2\lambda_j$ 表示主成分 F_j 能解释的原始变量 X_i 的方差，如果提取了 m 个主成分，第 i 个原始变量 X_i 的信息被提取的比率为

$$\sum_{j=1}^{m} u_{ij}{}^2 \lambda_j \Big/ \sigma_i^2 = \sum_{j=1}^{m} \rho(X_i, F_j)^2$$

在 SAS 中有多个过程步能够做主成分分析，包括：

（1）PROC CORRESP 分类数据和频数数据的主成分分析，输入数据为列联表或原始数据，进行简单对应分析或多重对应分析。

（2）PROC PRINQUAL 使用于对定性数据执行主成分分析，包括类别数据和定序数据，也可用于多维偏好分析。

（3）PROC PRINCOMP 对连续型数据执行主成分分析，输出标准化或非标准化的主成分得分，是 SAS 中执行主成分分析的主要过程步。

（4）PROC FACTOR 执行各种探索性因子分析，支持旋转并输出主成分得分或因子得分的估计，需要注意该过程步默认选项执行的是主成分分析。

（5）PROC CANCORR 执行典型相关分析并输出典型变量得分。结果以表格展示，也可以将结果输出至数据集进行绘图。

（6）PROC PLS 使用包括偏最小二乘法在内的线性预测方法进行模型拟合，广泛应用于各种分析。PROC PLS 过程步也能执行主成分回归分析，回归的输出可用于预测。

23.1.2 主成分分析的具体步骤

下面以 SASHELP.CLASS 为例，用 PROC PRINCOMP 简单解释主成分分析的步骤

和逻辑。数据分析作为一种需要结合具体经验的解释型学科，分析方法本身对数据的语义并不敏感。因此才有“模型皆有误，或尤建奇功”这一至理名言。程序 23-1 和程序 23-2 分别基于相关系数和协方差矩阵对 SASHELP.CLASS 的量化变量进行主成分分析。

程序23-1 基于相关系数矩阵对 SASHELP.CLASS 做主成分分析

```
proc princomp data=sashelp.class ;
run;
```

或

程序23-2 基于协方差矩阵对 SASHELP.CLASS 做主成分分析

```
proc princomp data=sashelp.class COV ;
run;
```

基于协方差矩阵的主成分分析结果如下（见图 23-3），结果各部分的解释如下：

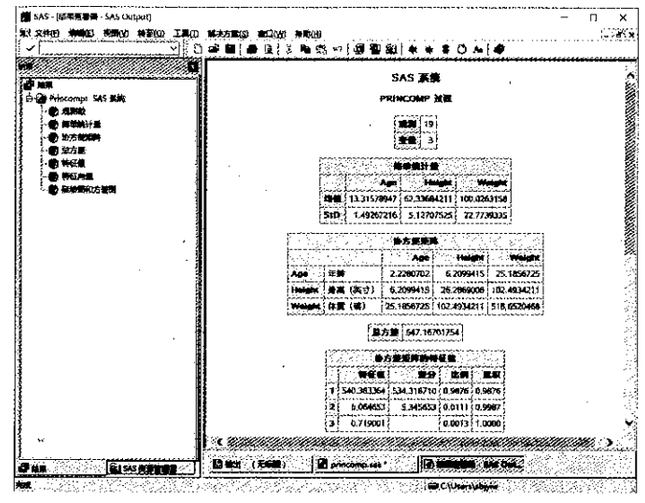

图 23-3 基于协方差矩阵的主成分分析结果

（1）观测数：其中可以解读出 PRINCOMP 使用 3 个定量变量 AGE、HEIGHT、WEIGHT 进行主成分分析，由于 PRINCOMP 是用于连续型变量的主成分分析，变量 NAME 和 SEX 为定类变量不参加分析，因此实际分析的观测数为 19，变量数为 3。

（2）简单统计量：包含各变量的均值和标准差，如 AGE 的均值为 13.31578947，标准差为 1.49267216，标准差的平方正是协方差矩阵（见图 23-4）对角线元素的值，如 AGE × AGE=2.2280702。

协方差矩阵

		Age	Height	Weight
Age	年龄	2.2280702	6.2099415	25.1856725
Height	身高（英寸）	6.2099415	26.2869006	102.4934211
Weight	体重（磅）	25.1856725	102.4934211	518.6520468

图 23-4 协方差矩阵

（3）协方差矩阵：列出了 3 个变量两两之间的协方差计算值，其中对角线元素为 3

个变量 AGE、HEIGHT 和 WEIGHT 各自的方差，如 AGE × AGE= 2.2280702。

协方差对角线元素之和为 3 个原变量的总方差 $\sigma_1^2+\sigma_2^2+\sigma_3^2$，即 547.16701754 = 2.2280702 + 26.2869006 + 518.5620468。

（4）总方差：列出了 3 个原始变量的总方差 547.16701754，在数据分析科学中，方差代表了数据所隐含的信息量，类似于化学中描述系统混乱程度的熵值。

（5）特征值和特征向量：基于协方差矩阵可计算出 3 个特征值 λ_1= 540.383364, λ_2=6.064653 和 λ_3= 0.719001（见图 23-5），分别代表主成分分析形成的 3 个主分量 PRIN1、PRIN2 和 PRIN3 对总体方差 547.16701754 的贡献，即 $\lambda_1+\lambda_2+\lambda_3$= 总方差。

每一个主成分 PRIN1、PRIN2 和 PRIN3 对总体方差的贡献比率为 λ_i / 总方差，它分别等于 0.9876、0.0111 和 0.0013，累积贡献比率为 0.9876、0.9987 和 1.0000。从中可以看出第一个分量 PRIN1 就已经贡献 98.76%，超过了 80% 阈值，因此根据选取准则我们只需要一个主分量 PRIN1 就可以了。

协方差矩阵的特征值				
	特征值	差分	比例	累积
1	540.383364	534.318710	0.9876	0.9876
2	6.064653	5.345653	0.0111	0.9987
3	0.719001		0.0013	1.0000

特征向量				
		Prin1	Prin2	Prin3
Age	年龄	0.048098	0.220785	0.974136
Height	身高（英寸）	0.195851	0.954248	-0.225948
Weight	体重（磅）	0.979453	-0.201653	-0.002657

图 23-5　特征值与特征向量

实际上，给定协方差矩阵或相关系数矩阵，在 SAS 中可以直接使用 PROC IML 过程步求解特征值和特征向量。比如上面的协方差矩阵，可用程序 23-3 的 PROC IML 过程步求解特征值和特征向量，其结果等价（见图 23-6）。

程序23-3　PROC IML 计算特征值和特征向量

```
proc iml;
  A={
    2.2280702    6.2099415   25.1856725,
    6.2099415   26.2869006  102.4934211,
   25.1856725  102.4934211  518.6520468
  };
  call eigen(eigenvalues, eigenvectors, A);
  print A eigenvalues eigenvectors;
quit ;
```


A			eigenvalues	eigenvectors		
2.2280702	6.2099415	25.185673	540.38336	0.0480984	0.2207849	0.9741358
6.2099415	26.286901	102.49342	6.0646532	0.1958508	0.9542484	-0.225948
25.185673	102.49342	518.65205	0.7190007	0.9794534	-0.201653	-0.002657

图 23-6　PROC IML 计算特征值和特征向量

基于协方差矩阵的主成分分析结果，各主成分 PRIN1, PRIN2 和 PRIN3 是原变量 AGE, HEIGHT, WEIGHT 的线性组合。主分量和原变量之间的系数由输出的特征向量给出，此时主成分和原变量有如下数学关系：

```
Prin1=0.048098 * (Age - Age_Mean) + 0.195851 * (Height - Height_Mean)
    + 0.979453 * (Weight - Weight_Mean)
```

比如对于第 1 个观测，其主成分得分 Prin1 可由如下公式计算出：

```
Prin1=0.048098 * (14 - 13.31578947) + 0.195851 * (69 - 62.33684211)
    + 0.979453 * (112.5 - 100.0263158)=13.5553
```

（6）陡坡图和方差图：各特征值 λ_i 和主成分能够解释的方差可由陡坡图和方差图（见图 23-7）给出，其中方差图中上面的虚线为累计方差贡献，下面的实线为各分量的方差贡献，形状与陡坡图是一样的。

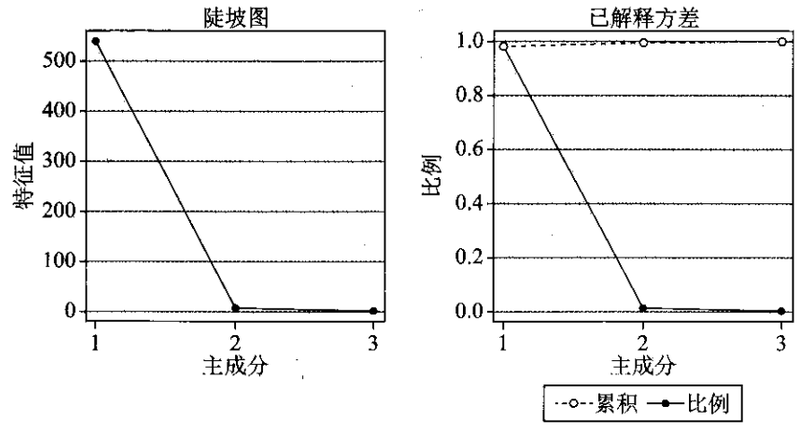

图 23-7　陡坡图和方差图

上面为采用协方差矩阵进行主成分分析，可以看出变量 WEIGHT 在主成分 PRIN1 中的系数最大。原因是变量 WEIGHT 的数值量级比较大（其平均值为 100 左右），而变量 AGE 数值较小（其平均值为 13）。为了消除量纲和单位的影响，我们通常需要将数据标准化（变换后数据的均值为 0，方差为 1）后再执行主成分分析，才不会出现这种偏差。程序 23-4 首先标准化 3 个量化原变量 AGE、HEIGHT、WEIGHT，然后再调用协方差方法进行主成分分析：

程序23-4 数据标准化后用协方差矩阵做主成分分析

```
proc standard data = sashelp.class out=class_std mean = 0 std=1;
  var Age Height Weight;
run;

proc princomp data=class_std cov ;
run;
```

由数学上推导可知，基于标准化数据计算的协方差矩阵等价于相关系数矩阵，因此上面的程序输出完全等价于如下使用相关系数矩阵进行的主成分分析，这也是SAS默认的主成分分析方法选项（见程序23-5）。

程序23-5 默认使用相关系数矩阵做主成分分析

```
proc princomp data=sashelp.class ;
run;
```

从相关系数矩阵（见图23-8）可以看出，HEIGHT和WEIGHT关系最紧密为0.8778，其次为HEIGHT和AGE 0.8114，再其次为WEIGHT和AGE 0.7409。

从特征值累积方差贡献上也可以看出只需要一个主成分PRIN1即可达到87.38%(0.8738)，此时主成分与原变量的标准化数据有如下关系：

```
Prin1 = 0.560811 * Age_Std + 0.593307 * Height_Std + 0.577476 * Weight_Std
```

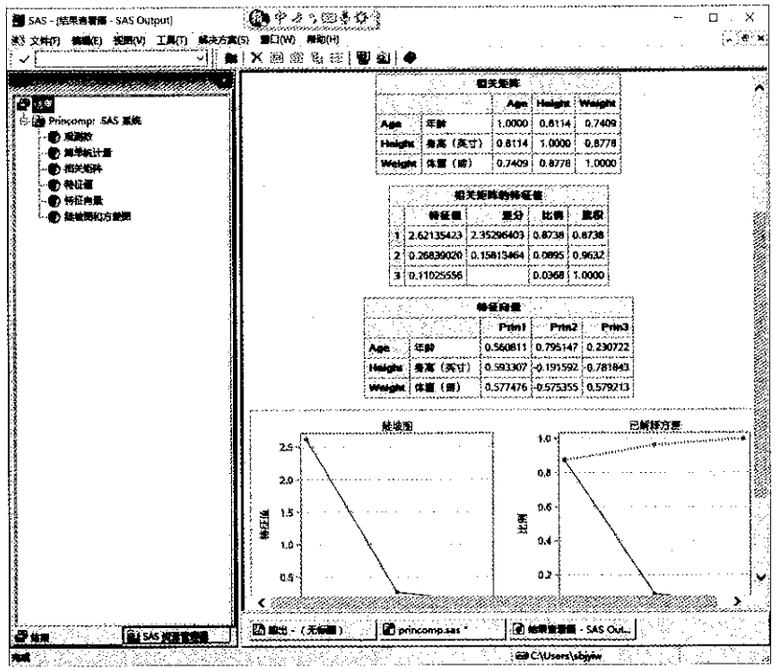

图23-8 基于相关系数矩阵的主成分分析

主成分分析的统计量和结果可以输出到数据集中（见程序23-6），OUTSTAT=将各种统计量包括相关系数，特征值和特征向量（即各主成分和对应变量的系数）输出到CLASS_OUTSTAT中，OUT=将原变量和对应的主成分输出到指定数据集CLASS_OUT中。

程序23-6 输出统计量主成分得分

```
proc princomp data=sashelp.class outstat=class_outstat out=class_out;
run;
proc print data=class_outstat;
run;
```

上面的过程步输出如下各种统计量（见图23-9）。

Obs	_TYPE_	_NAME_	Age	Height	Weight
1	MEAN		13.3158	62.3368	100.026
2	STD		1.4927	5.1271	22.774
3	N		19.0000	19.0000	19.000
4	CORR	Age	1.0000	0.8114	0.741
5	CORR	Height	0.8114	1.0000	0.878
6	CORR	Weight	0.7409	0.8778	1.000
7	EIGENVAL		2.6214	0.2684	0.110
8	SCORE	Prin1	0.5608	0.5933	0.577
9	SCORE	Prin2	0.7951	-0.1916	-0.575
10	SCORE	Prin3	0.2307	-0.7818	0.579

图23-9 输出统计量

输出的数据集CLASS_OUT中包括3个原始变量和3个新的主成分得分，用如下方法验证各主成分之间是互相独立的，即各主成分是线性无关的，其单元格中第一排的相关系数为0（见程序23-7和图23-10）。

程序23-7 验证主成分之间的相关性

```
proc corr data=class_out out=class_out_corr_prin_prin;
  var prin1-prin3;
  with prin1-prin3;
run ;
```

Pearson 相关系数，N = 19
Prob > |r| under H0: Rho=0

	Prin1	Prin2	Prin3
Prin1	1.00000	0.00000 1.0000	0.00000 1.0000
Prin2	0.00000 1.0000	1.00000	0.00000 1.0000
Prin3	0.00000 1.0000	0.00000 1.0000	1.00000

图23-10 主成分之间的相关系数

同理，可以检查各主成分和原来的各变量之间的相关关系（见程序23-8），输出结果（见图23-11）的第1行(0.90799)表示因子载荷，反映了主成分受某个原始变量影响的程度，绝对值越大表示影响越强烈，而符号则表示受影响的方向，正相关还是负相关。第2行(<0.0001)表示零假设下的检验概率，表示在0.01%水平上显著。

程序23-8 验证主成分和变量之间的相关关系

```
proc corr data=class_out out=class_out_corr_prin_var;
  var prin1-prin3;
  with Age Height Weight;
run ;
```