04

主成分分析

➢学习目标：

1.理解主成分分析的基本理论与方法;
2.了解主成分的性质;
3.理解主成分的求解方法;
4.掌握用R软件求解主成分的方法;
5.正确理解软件输出结果并对结果进行分析

4.1 主成分分析的基本原理

主成分分析

- ✓ 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。

4.1 主成分分析的基本原理

4.1.1 主成分分析的基本思想

◆ **思想：**

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

◆ **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

4.1 主成分分析的基本原理

4.1.2 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$X_1, X_2, \dots, Xp$表示,这p个指标构成的p维随机向量为$X = (X_1, X_2, \dots, Xp)'$。设随机向量X的均值为μ,协方差矩阵为Σ

对X进行线性变换,可以形成新的综合变量,用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:

$$\begin{cases} Y_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{P1}X_P \\ Y_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{P2}X_P \\ \cdots \\ Y_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{PP}X_P \end{cases}$$

4.1 主成分分析的基本原理

4.1.2 主成分分析的基本理论

◆ 我们将线性变换约束在下面的原则之下:

(1)$u_i' u_i = 1 (i = 1, 2, \dots, p)$

(2)Y_i与Y_j相互无关$(i \neq j; i, j = 1, 2, \dots, p)$

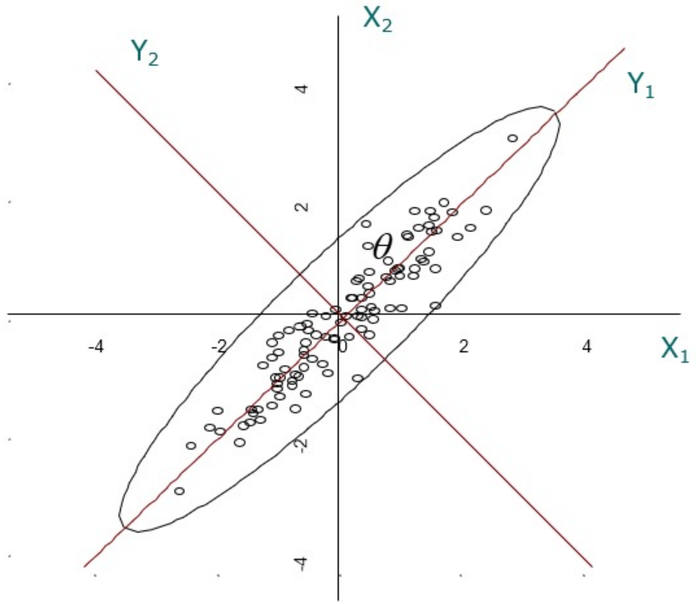
(3)Y_1是$X_1, X_2, \dots, Xp$的一切满足原则(1)的线性组合中方差最大者;Y_2是与Y_1不相关的$X_1, X_2, \dots, Xp$的所有线性组合中方差最大者……Y_p是与$Y_1, Y_2, \dots, Y_{P-1}$都不相关的$X_1, X_2, \dots, X_p$的所有线性组合中方差最大者。

基于以上三条原则确定的综合变量$Y_1, Y_2, \dots, Yp$分别称为原始变量的第一、第二……第p个主成分

4.1 主成分分析的基本原理念

4.1.3 主成分分析的几何意义

◆ 设有N个样品,每个样品有两个观测变量X_1, X_2,这样,在由变量X_1, X_2组成的坐标空间中,N个样品散布的情况如带状。



我们的目的是考虑X_1和X_2的线性组合,使原始样品数据可以由新的变量Y_1和Y_2来刻画。在几何上表示就是将坐标轴按逆时针方向旋转θ角度,得到新坐标轴Y_1和Y_2：

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

4.2 总体主成分及其性质

本节导论

- ✓ 对于随机变量$X_1, X_2, \ldots, Xp$而言,其协方差矩阵或相关矩阵正是对各变量离散程度与变量之间的相关程度的信息的反映,**相关矩阵不过是将原始变量标准化后的协方差矩阵**。
- ✓ 我们所说的保留原始变量尽可能多的信息,也就是指生成的较少的综合变量(主成分)的方差和尽可能接近原始变量方差的总和。因此在实际求解主成分的时候,总是从原始变量的**协方差矩阵**或**相关矩阵**的结构分析入手。

4.2 总体主成分及其性质

4.2.1 从协方差矩阵出发求解主成分

引论

设矩阵$A' = A$,将A的特征根$\lambda_1, \lambda_2, \dots, \lambda_n$依大小顺序排列,不妨设$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \gamma_1, \gamma_2, \dots, \gamma_n$为矩阵$\Sigma$各特征根对应的标准正交特征向量,则对任意向量$x$,有

$$\max_{x \neq 0} \frac{x'Ax}{x'x} = \lambda_1, \quad \min_{x \neq 0} \frac{x'Ax}{x'x} = \lambda_n$$

4.2 总体主成分及其性质

4.2.1 从协方差矩阵出发求解主成分

结论

设随机向量$X=(X_1,X_2,\dots,X_p)'$的协方差矩阵为$\Sigma,\lambda_1,\lambda_2,\dots,\lambda_p(\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p)$为$\Sigma$的特征根,$\gamma_1,\gamma_2,\dots,\gamma_p$为矩阵$\Sigma$各特征根对应的标准正交特征向量,则第$i$个主成分为:$Y_i=\gamma_{1i}X_1+\gamma_{2i}X_2+\cdots+\gamma_{p_i}X_P,\quad i=1,2,\dots,\mathrm{p}$

令$P=(\gamma_1,\gamma_2,\dots,\gamma_p),\Lambda=diag(\lambda_1,\lambda_2,\dots,\lambda_p),Y$的分量$Y_1,Y_2,\dots,Y_p$依次是$X$的第一主成分、第二主成分……第$p$主成分的充分必要条件是:

(1)$Y=P'X$,即P为p阶正交阵;

(2)Y的分量之间互不相关,即$D(Y)=diag(\lambda_1,\lambda_2,\dots,\lambda p)$;

(3)Y的p个分量按方差由大到小排列,即$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$

4.2 总体主成分及其性质

4.2.2 主成分的性质

定义4.1

称$\alpha_k = \frac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}(k = 1, 2, \ldots, p)$为第k个主成分$Y_k$的**方差贡献率**，称$\frac{\sum_{j=1}^{m}\lambda_i}{\sum_{j=1}^{p}\lambda_i}$为主成分$Y_1, Y_2, \ldots, Y_m$的**累积贡献率**

- 其中：进行主成分分析的目的之一是减少变量的个数,因此一般不会取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。

4.2 总体主成分及其性质

4.2.2 主成分的性质

定义4.2

第k个主成分Y_k与原始变量X_i的相关系数$\rho(Y_k, Xi)$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；因子负荷量与系数向量成正比。

4.2 总体主成分及其性质

4.2.2 主成分的性质

定义4.3

X_i与前m个主成分$Y_1, Y_2, \dots, Y_m$的全相关系数平方和称为$Y_1, Y_2, \dots, Y_m$对原始变量X_i的**方差贡献率**v_i，即

$$v_i = \frac{1}{\sigma_{ii}} \sum_{k=1}^{m} \lambda_k \gamma_{ik}^2 \quad (i = 1, 2, \dots, p)$$

- 这一定义说明了前m个主成分提取了原始变量X_i中v_i的信息,由此可以判断我们提取的主成分说明原始变量的能力。

4.2 总体主成分及其性质

4.2.2 主成分的性质

性质1 $\boldsymbol{Y}$的协方差为对角阵$\boldsymbol{\Lambda}$

性质2 $\Sigma = (\sigma_{ij})_{P\times P}$，有 $\sum_{i=1}^{P}\lambda_i = \sum_{i=1}^{P}\sigma_{ii}$

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

性质4 $\sum_{i=1}^{P}\rho^2(Y_k, X_i)\sigma_{ii} = \lambda_k$

性质5 $\sum_{i=1}^{P}\rho^2(Y_k, X_i) = \frac{1}{\sigma_{ii}}\sum_{k=1}^{P}\lambda_k\gamma_{ik}^2 = 1$

4.2 总体主成分及其性质

4.2.3 从相关矩阵出发求解主成分

◆ 对原始变量X进行如下标准化: $Z = (\Sigma^{1/2})^{-1}(X - \mu)$

经过上述标准化后,显然有: $E(Z) = 0$

$$cov(Z) = (\Sigma^{1/2})^{-1}\Sigma(\Sigma^{1/2})^{-1} = \begin{bmatrix} 1 & \rho_{12} & \rho_{1p} \\ \rho_{12} & 1 & \rho_{2p} \\ \cdots & \cdots & \cdots \\ \rho_{1p} & \rho_{2p} & 1 \end{bmatrix}$$

◆ 由于上面的变换过程,原始变量$X_1, X_2, \ldots, X_p$的相关阵实际上就是对原始变量标准化后的协方差矩阵.此时,求得的主成分与原始变量的关系式为:

$$Y_i = \gamma_i' Z = \gamma_i' (\Sigma^{1/2})^{-1} (X - \mu), i = 1, 2, \ldots, p$$

4.2 总体主成分及其性质

4.2.4 由相关阵求主成分时主成分性质的简单形式

性质

我们将由相关阵得到的主成分的性质总结如下:

(1) Y的协方差矩阵为对角阵Λ;

(2) $\sum_{i=1}^{m} var(Yi) = tr(\Lambda) = tr(R) = p = \sum_{i=1}^{p} var(Z_i)$;

(3) 第k个主成分的方差占总方差的比例,即第k个主成分的方差贡献率为$\alpha_k = \lambda_k/p$,前m个主成分的累积方差贡献率为 $\sum_{i=1}^{m} \frac{\lambda_i}{p}$

(4) $\rho(Yk, Zi) = \gamma\gamma_i \sqrt{\lambda_k}$

4.3 样本主成分的导出

符号定义

设有n个样品,每个样品有p个指标,这样共得到np个数据,:

$$X=\begin{bmatrix} x_{11} & x_{12} & x_{1p} \\ x_{12} & x_{22} & x_{2p} \\ \cdots & \cdots & \cdots \\ x_{1p} & x_{2p} & x_{np} \end{bmatrix}$$

记：$s=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\bar{X})\ (X_k-\bar{X})'$

$$\bar{X}=\frac{1}{n}\sum_{k=1}^{n}X_k=(\bar{X}_1,\bar{X}_{2,\ldots,}\bar{X}_p)'$$

S为样本协方差矩阵,作为总体协方差阵Σ的无偏估计;R为样本相关矩阵,为总体相关矩阵的估计。

4.3 样本主成分的导出

由相关阵R求解主成分

- 根据总体主成分的定义,主成分Y的协方差是: $cov(Y) = \Lambda$

$$\Lambda=\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

- 假定资料矩阵X为标准化后的数据矩阵,则可由相关矩阵代替协方差矩阵,于是上式可表示为: $P'RP = \Lambda$

故：所求的新的综合变量(主成分)的方差$\lambda_i(i = 1, 2, \dots, p)$是$|R - \lambda I| = 0$的$p$个根,$\lambda$为相关矩阵的特征根,相应的各个$\gamma_{ij}$是其特征向量的分量。

4.3 样本主成分的导出

由相关阵R求解主成分

● 因为R为正定矩阵,所以其特征根都是非负实数,将它们依大小顺序排列$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$,其相应的特征向量记为$\gamma_1, \gamma_2, \ldots, \gamma_p$,则

$$var(Yi) = var(\gamma_i' X) = \lambda_i$$

即对于Y_1有最大方差,Y_2有次大方差,……,并且协方差为:

$$cov(Yi, Yj) = \sum_{\alpha=1}^{p} \lambda_\alpha (\gamma'_i \gamma_\alpha)(\gamma'_\alpha \gamma_j) = 0, i \neq j$$

由此可有新的综合变量(主成分)$Y_1, Y_2, \ldots, Yp$彼此不相关,并且Y_i的方差为λ_i,则

$Y_1 = \gamma'_1 X, Y_2 = \gamma'_2 X, \ldots, Yp = \gamma'_p X$分别称为第一、第二…… 第$p$个主成分。

4.4 有关问题的讨论

4.4.1 关于由协方差矩阵或相关矩阵出发求解主成分

1.相同之处

- 求主成分的过程是一致的，实际就是对矩阵结构进行分析的过程，也就是求解特征根的过程。

2.如何选择

- 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
- 对同度量或取值范围在同量级的数据,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，要考虑实际情况

4.4 有关问题的讨论

4.4.2 主成分分析不要求数据来自正态总体

- 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
- 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。

4.4 有关问题的讨论

4.4.3 主成分分析与重叠信息

◆ 主成分分析方法适用于变量之间存在较强相关性的数据,如果原始数据相关性较弱,运用主成分分析不能起到很好的降维作用

◆ 虽然主成分分析不能有效地剔除重叠信息,但它至少可以发现原始变量是否存在重叠信息：如果所得到的样本协方差矩阵(或相关阵)最小的特征根接近零,那么就有:

$$\hat{\Sigma}\Upsilon_p=\frac{1}{n-1}(X-1\bar{X}')'(X-1\bar{X}')\,\Upsilon_p=\lambda_p\Upsilon_p\approx 0$$

其中,1是元素均为1的n元列向量。

上式中左乘Υ_p'可得$\Upsilon_p'(X-1\bar{X}')'(X-1\bar{X}')\,\Upsilon_p\approx 0$，进而推出$(X-1\bar{X}')\,\Upsilon_p\approx 0$

这就意味着,中心化以后的原始变量之间存在着多重共线性（即重叠信息）

4.5 主成分分析步骤及框图

4.5.1 主成分分析步骤

step1: 根据研究问题选取初始分析变量;

step2: 根据初始变量特性判断由协方差阵求主成分还是由相关阵求主成分;

step3: 求协方差阵或相关阵的特征根与相应标准特征向量;

step4: 判断是否存在明显的多重共线性,若存在,则回到第①步;

step5: 得到主成分的表达式并确定主成分个数,选取主成分;

step6: 结合主成分对研究问题进行分析并深入研究。

4.6 主成分分析的上机实现

【例5-1】为掌握我国各地区主要行业的城镇私营企业就业人员的平均工资水平,选取2016年我国30个省、直辖市、自治区(西藏地区数据缺失)9个行业就业人员的平均工资数据(数据来源于2017年《中国统计年鉴》)。我们用主成分分析方法处理该数据,以期用较少的变量描述这些行业各地区就业人员的工资水平。本例中各变量的量纲差别不大,为了保留各变量自身的变异,选择从协方差阵出发求解主成分。主要分为以下三步：

Step1：读入数据,计算特征值

Step2：计算特征向量和因子负荷量

Step3：第3步:进一步分析主成分的选择

表 5-1　2016 年分地区城镇私营企业就业人员平均工资单位:元

X_1　农、林、牧、渔业　　X_2　制造业　　X_3　电力、热力、燃气及水的生产和供应业

X_4　建筑业　　X_5　住宿和餐饮业　　X_6　金融业

X_7　房地产业　　X_8　教育业　　X_9　文化、体育和娱乐业

地区	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
北京	39138	58042	53062	49455	43187	143717	94956	65646	64250
天津	36007	61667	47103	50372	43400	68436	58365	44999	51602
河北	31330	37333	35800	36976	33168	37756	40386	35583	33065
山西	21145	30736	31722	35151	25722	34298	33398	27957	23396
内蒙古	31084	38296	39644	39719	32674	42301	35588	30959	31807
辽宁	28618	33884	32464	37930	28811	34981	35940	31771	29453
吉林	22137	29395	26099	32881	27225	36990	30983	35481	26079

4.6 主成分分析的上机实现

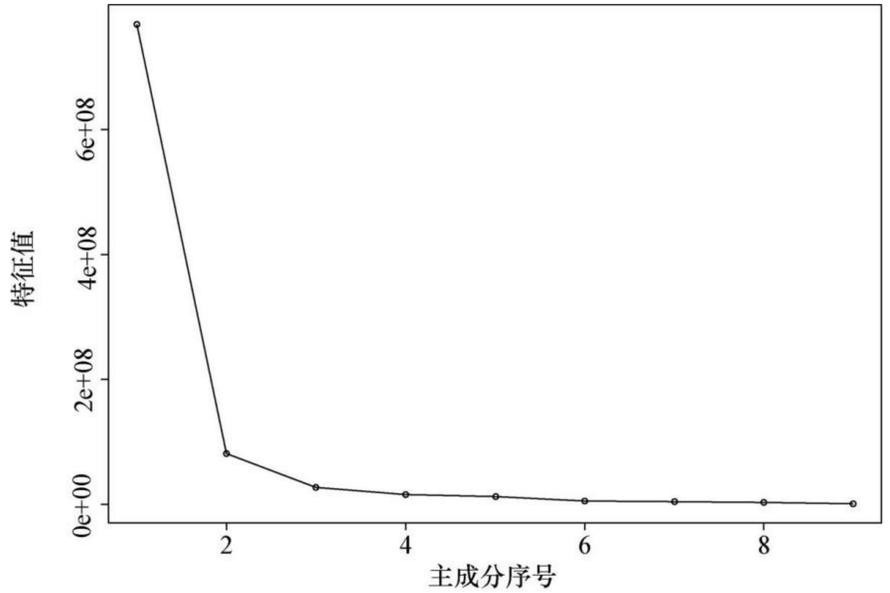
◆ Step1：读入数据,计算特征值，输入如下：

```
1.> rm(list=ls())
2.> ex5.1 <- read.table("例5-1.txt", head=TRUE, fileEncoding="utf8")
3.> dat51 <- ex5.1[, -1]
4.> rownames(dat51) <- ex5.1[, 1]
5.>#协方差矩阵
6.> sigm <- cov(dat51)
7.> my51 <- eigen(sigm)
8.>#特征值
9.> lam <- my51$values
10.> p  <- length(lam)
11.>#方差贡献率
12.> cumlam <- cumsum(lam)/sum(lam)
13.> VE <- data.frame(lam, lam/sum(lam), cumlam)
14.> colnames(VE) <- c("特征根","贡献率","累计贡献率")
15.> print(VE)
```

4.6 主成分分析的上机实现

- Step1：读入数据,计算特征值，我们使用eigen函数对协方差矩阵进行特征值分解。第一主成分的方差贡献率为82.302%,是保留的特征根占所有特征根的和的比值,由此可见第一主成分解释原始变量总差异的效果比较好。第二个主成分的方差贡献率为8.857%,这个相对第一主成分贡献率低很多。碎石图显示选择2个主成分比较好。

16.1	特征根	贡献率	累计贡献率
17.2	768365315	0.823019789	0.8230198
18.2	82685252	0.088566724	0.9115865
19.3	29249216	0.031329738	0.9429163
20.4	17215863	0.018440442	0.9613567
21.5	14296274	0.015313180	0.9766699
22.6	7511072	0.008045341	0.9847152
23.7	6169207	0.006608028	0.9913232
24.8	5030735	0.005388576	0.9967118
25.9	3069824	0.003288183	1.0000000
26.#碎石图			
27.plot(lam, type="o", xlab="主成分序号", ylab="特征值")			



4.6 主成分分析的上机实现

◆ Step2：计算特征向量和因子负荷量。

```
1.>#特征向量
2.> gam <- my51$vectors
3.> colnames(gam) <- paste("vec",sep = "", 1:p)
4.> print(gam[, 1:2])
5.              vec1          vec2
6.   [1,] -0.1191864 -0.31997109
7.   [2,] -0.2320962 -0.35693452
8.   [3,] -0.1694318 -0.35315311
9.   [4,] -0.1452976 -0.38420063
10. [5,] -0.1302961 -0.27546205
11. [6,] -0.7438316  0.51064145
12. [7,] -0.4091319  0.07373654
13. [8,] -0.2533333 -0.12321860
14. [9,] -0.2833341 -0.37501342
```

4.6 主成分分析的上机实现

◆ Step2：计算特征向量和因子负荷量。

```
15.>#因子负荷量
16.> lam_ma <- matrix(lam, p, p, byrow = TRUE)
17.> sigmai <- (diag(sigm))^0.5
18.>##特征向量*特征根的算数平方根
19.> gamsla <- gam * sqrt(lam_ma)
20.> load <- gamsla / sigmai
21.> colnames(load) <- paste("load",sep = "", 1:p)
22.> print(load[,1:2])
23.              load1      load2
24. [1,] -0.6478731 -0.57056382
25. [2,] -0.8469924 -0.42729737
26. [3,] -0.6590387 -0.45061871
27. [4,] -0.6636656 -0.57567732
28. [5,] -0.7368664 -0.51103336
29. [6,] -0.9717312  0.21883548
30. [7,] -0.9320512  0.05510478
31. [8,] -0.9099041 -0.14518100
32. [9,] -0.8757826 -0.38025452
```

4.6 主成分分析的上机实现

- Step3：进一步分析主成分的选择。第一主成分和第二主成分对原始各变量方差贡献率的和,即0.412+…+ 0.145+=7.483,以及该主成分占所有主成分对原始变量方差贡献率总和(等于9)的比值为0.832。这说明前面选择两个主成分比较合适。

```
1.>#定义5.3 第一和二主成分对X1-X9的方差贡献率
2.> VV <- load ^ 2
3.> print(VV[, 1:2])
4.      load1     load2
5. [1,] 0.4197396 0.325543073
6. [2,] 0.7173962 0.182583046
7. [3,] 0.4343320 0.203057225
8. [4,] 0.4404520 0.331404379
9. [5,] 0.5429721 0.261155094
10. [6,] 0.9442615 0.047888965
11. [7,] 0.8687194 0.003036536
12. [8,] 0.8279255 0.021077522
13. [9,] 0.7669951 0.144593498
14.> sum(VV[, 1:2])/9
15.[1] 0.8314592
```

4.6 主成分分析的上机实现

【例5-2】在工业企业经济效益的评价中,设计的指标往往较多。为了简化系统结构,抓住经济效益评价中的主要方面,我们可由原始数据出发求主成分。在对我国各地区规模以上工业企业的经济效益评价中,包含8项指标,原始数据如表5-2所示(数据来源于2017年《中国工业统计年鉴》),其中,前7项指标的单位是亿元,最后一项指标的单位是万人。由于原始数据量纲差别较大,需要对数据进行标准化。步骤和程序与前面差不多，分为以下三步：

Step1：读入数据,并输出变量之间的相关性

Step2：计算特征值并输出特征根及对应主成分的方差贡献率和累积贡献率

Step3：计算特征向量和因子负荷量

表 5-2　2016 年各地区规模以上工业企业主要经济指标

地区	工业销售产值 X_1	资产总计 X_2	负债合计 X_3	所有者权益合计 X_4	主营业务收入 X_5	利润总额 X_6	投资收益 X_7	平均用工人数 X_8
北京	17837.50	43093.68	19798.13	23272.45	19746.96	1608.26	635.55	104.45
天津	26654.45	25075.09	15385.02	10095.20	25888.20	2046.69	-6.52	146.98

4.6 主成分分析的上机实现

- Step1：读入数据,并输出变量之间的相关性，输出结果显示8个变量之间存在较强的相关关系,适合进行主成分分析。

```
1.> rm(list=ls())
2.> ex5.2 <- read.table("例5-2.txt", head=TRUE, fileEncoding="utf8")
3.> dat52 <- ex5.2[, -1]
4.> rownames(dat52) <- ex5.2[, 1]
5.> dat52 <- scale(dat52, scale = TRUE, center = TRUE)
6.> #协方差
7.> sigm <- cov(dat52)
8.> print(sigm, digits=3)
9.          X1    X2    X3    X4    X5    X6    X7    X8
10.X1 1.000 0.958 0.936 0.962 1.000 0.990 0.345 0.952
11.X2 0.958 1.000 0.991 0.988 0.959 0.954 0.492 0.937
12.X3 0.936 0.991 1.000 0.959 0.937 0.923 0.455 0.928
13.X4 0.962 0.988 0.959 1.000 0.964 0.969 0.520 0.927
14.X5 1.000 0.959 0.937 0.964 1.000 0.990 0.354 0.949
15.X6 0.990 0.954 0.923 0.969 0.990 1.000 0.405 0.945
16.X7 0.345 0.492 0.455 0.520 0.354 0.405 1.000 0.385
17.X8 0.952 0.937 0.928 0.927 0.949 0.945 0.385 1.000
```

4.6 主成分分析的上机实现

- Step2：计算特征值并输出特征根及对应主成分的方差贡献率和累积贡献率。我们可提取1个主成分,其方差贡献率为86.981%,说明该第一主成分基本上提取了原始变量的大部分信息。这样由分析原来的8个变量转化为仅需分析1个综合变量,极大地起到了降维的作用。

```
1.> my52 <- eigen(sigm)
2.> #特征值
3.> lam <- my52$values
4.> p <- length(lam)
5.> #方差解释
6.> cumlam <- cumsum(lam)/sum(lam)
7.> VE <- data.frame(lam, lam/sum(lam), cumlam)
8.> colnames(VE) <- c("特征值","比例","累计比例")
9.> print(VE,digits = 5)
```

```
10.    特征值     比例      累计比例
11.1 6.9584e+00 8.6981e-01 0.86981
12.2 8.2794e-01 1.0349e-01 0.97330
13.3 1.0895e-01 1.3619e-02 0.98692
14.4 7.8891e-02 9.8613e-03 0.99678
15.5 1.7949e-02 2.2436e-03 0.99902
16.6 7.5446e-03 9.4308e-04 0.99996
17.7 2.7163e-04 3.3954e-05 1.00000
18.8 9.3902e-06 1.1738e-06 1.00000
```

4.6 主成分分析的上机实现

◆ Step3：计算特征向量和因子负荷量。

```
1.>#特征向量
2.> gam <- my52$vectors
3.> colnames(gam) <- paste("vec",sep = "", 1:p)
4.> print(gam[, 1:2], digits = 4)
5.          vec1    vec2
6. [1,] -0.3721  0.173239
7. [2,] -0.3755 -0.022033
8. [3,] -0.3684  0.008133
9. [4,] -0.3752 -0.053518
10.[5,] -0.3724  0.161586
11.[6,] -0.3722  0.099774
12.[7,] -0.1844 -0.958285
13.[8,] -0.3645  0.110445
```

4.6 主成分分析的上机实现

◆ Step3：计算特征向量和因子负荷量。

```
14.>#因子负荷量
15.> lam_ma <- matrix(lam, p, p, byrow = TRUE)
16.> sigmai <- (diag(sigm))^0.5
17.>##特征向量*特征根的算数平方根
18.> gamsla <- gam * sqrt(lam_ma)
19.> load <- gamsla / sigmai
20.> colnames(load) <- paste("load",sep = "", 1:p)
21.> print(load[, 1:2], digits = 4)
22.          load1    load2
23.[1,] -0.9814  0.15763
24.[2,] -0.9904 -0.02005
25.[3,] -0.9717  0.00740
26.[4,] -0.9898 -0.04870
27.[5,] -0.9824  0.14703
28.[6,] -0.9819  0.09079
29.[7,] -0.4863 -0.87195
30.[8,] -0.9616  0.10049
```

4.6 主成分分析的上机实现

【例5-3】试利用主成分综合评价全国各地区水泥制造业规模以上企业的经济效益,原始数据来源于2014年《中国水泥年鉴》，与上面例子一样,我们将数据进行标准化处理。

Step1：读入数据,并输出变量之间的相关性

Step2：计算特征值

Step3：计算特征向量和因子负荷量

Step4：进一步分析主成分。

4.6 主成分分析的上机实现

- Step1：读入数据,并输出变量之间的相关性，输出结果显示除X_7与各变量的相关性不强外,其他变量之间均存在较强的相关关系,因此原始数据适合做主成分分析。

```
1.> rm(list=ls())
2.> ex5.3 <- read.table("例5-3.txt", head=TRUE, fileEncoding="utf8")
3.> dat53 <- ex5.3[, -1]
4.> rownames(dat53) <- ex5.3[, 1]
5.> dat53 <- scale(dat53, center = TRUE, scale = TRUE)
6.>#协方差
7.> sigm <- cov(dat53)
8.> print(sigm, digits=3)
9.          X1      X2      X3      X4      X5      X6     X7
10.X1 1.000 0.7629 0.8518 0.7950 0.902 0.821  0.1570
11.X2 0.763 1.0000 0.9234 0.8967 0.881 0.715  0.0248
12.X3 0.852 0.9234 1.0000 0.9809 0.875 0.694  0.0252
13.X4 0.795 0.8967 0.9809 1.0000 0.810 0.582 -0.0506
14.X5 0.902 0.8809 0.8750 0.8102 1.000 0.903  0.1884
15.X6 0.821 0.7155 0.6945 0.5818 0.903 1.000  0.4282
16.X7 0.157 0.0248 0.0252-0.0506 0.188 0.428  1.0000
```

4.6 主成分分析的上机实现

- Step2：计算特征值。结果可以看到,本例保留了前两个主成分,它们解释了全部变量总方差的91.036%,说明这2个主成分代表原来的7个指标评价企业的经济效益已经足够。

```
1.> my53 <- eigen(sigm)
2.>#特征值
3.> lam <- my53$values
4.> p <- length(lam)
5.>#方差解释
6.> cumlam <- cumsum(lam)/sum(lam)
7.> VE <- data.frame(lam, lam/sum(lam), cumlam)
8.> colnames(VE) <- c("特征值","比例","累计比例")
9.> print(VE,digits = 5)
10.   特征值      比例 累计比例
11.1 5.1633892 0.7376270 0.73763
12.2 1.2091445 0.1727349 0.91036
13.3 0.3418942 0.0488420 0.95920
14.4 0.1947949 0.0278278 0.98703
15.5 0.0490616 0.0070088 0.99404
16.6 0.0341498 0.0048785 0.99892
17.7 0.0075659 0.0010808 1.00000
```

4.6 主成分分析的上机实现

◆ Step3：计算特征向量和因子负荷量。

```
1.>#特征向量
2.> gam <- my53$vectors
3.> colnames(gam) <- paste("vec",sep = "", 1:p)
4.> print(gam[, 1:2], digits = 4)
5.       vec1     vec2
6. [1,] -0.40708  0.04364
7. [2,] -0.40956 -0.15500
8. [3,] -0.42124 -0.17831
9. [4,] -0.39992 -0.26946
10.[5,] -0.42630  0.07045
11.[6,] -0.37688  0.35951
12.[7,] -0.07354  0.85759
```

4.6 主成分分析的上机实现

◆ Step3：计算特征向量和因子负荷量。

```
13.>#因子负荷量
14.> lam_ma <- matrix(lam, p, p, byrow = TRUE)
15.> sigmai <- (diag(sigm))^0.5
16.>##特征向量*特征根的算数平方根
17.> gamsla <- gam * sqrt(lam_ma)
18.> load <- gamsla / sigmai
19.> colnames(load) <- paste("load",sep = "", 1:p)
20.> print(load[, 1:2], digits = 4)
21.    load1    load2
22.[1,] -0.9250  0.04799
23.[2,] -0.9306 -0.17044
24.[3,] -0.9572 -0.19607
25.[4,] -0.9087 -0.29630
26.[5,] -0.9687  0.07747
27.[6,] -0.8564  0.39532
28.[7,] -0.1671   0.94301
```

4.6 主成分分析的上机实现

- Step4：进一步分析主成分。当主成分有两个时,将各样品的主成分得分在平面直角坐标系上描出来,就可得到各样品的分布情况,然后可以对样品进行分类。将标准化后的原始数据代入两个主成分的线性表达式,计算各样品的两个主成分得分。现将各样品的主成分得分在平面直角坐标系上描出来(使用R软件画散点图并添加辅助线)

```
1.y1 <- dat53 %*% as.matrix(gam[, 1], p, 1)
2.y2 <- dat53 %*% as.matrix(gam[, 2], p, 1)
3.#data.frame(y1, y2)#表5-2
4.plot(y1, y2, pch="+", xlab = "第一主成分", ylab="第二主成分")
5.abline(h=0, lty=2)
6.abline(v=0, lty=2)
7.text(y1,y2, ex5.3[, 1], adj= -0.05)
```

4.6 主成分分析的上机实现

- Step4：
- 由图可知,分布在第一象限的地区是广西、江西、安徽、湖南、浙江、广东、湖北、江苏和山东,说明这些省区的规模以上的水泥企业的经济效益较好,企业整体规模大且收入高,盈利能力强;
- 分布在第三象限的地区是黑龙江、陕西、福建、云南、重庆、山西、新疆、上海、天津、北京,说明这些地区的规模以上的水泥企业的经济效益较差,企业整体规模小且盈利能力弱,尤其是北京地区的水泥企业的经济效益最差,主要是由于北京地区较大规模的水泥企业比较少。