第七章、主成分分析

December 11, 2018

7.1 总体的主成分

主成分分析试图在力保数据信息丢失最少的原则下，对这种多变量的截面数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。

问题：

1. 什么时候需要用主成分?
2. 什么时候不能用主成分?

1. 维数灾难问题: 变量过多，增加分析问题的复杂性，影响计算速度；
2. 多重共线性： 变量间常常存在着一定的相关性，使得观测到的数据在一定程度上反映的信息有所重叠，影响统计模型的精度。

案例一：主成分分析法在不同品牌啤酒风味差异性评价中的应用

研究背景： 啤酒是含酒精的饮料酒, 啤酒的风味是人们选择啤酒的主要影响因素。

- 啤酒不同于同浓度的酒精水溶液, 主要是因为啤酒除了含有酒精外还含有数以百计的微量成分, 例如醛、醇及酯类等。
- 对于啤酒生产企业来说, 把自己的啤酒和竞争啤酒的风味进行比较非常重要, 这样可以了解自己的啤酒和竞品的差异, 分析竞争啤酒受市场欢迎的原因, 以改进自己的产品, 或者找出自己啤酒的风格特点, 走差异化竞争之路。
- 为了解决此问题, 啤酒企业可以对啤酒的风味成分进行分析, 理论上讲, 分析的成分越多, 获得的信息量越大.

由于变量增加, 很难从总体上进行对比分析, 这时, 可以通过主成分分析法, 提取主要的综合成分, 然后在平面坐标系中画图进行比较。下面是对3个啤酒品牌分析的例子.

- 啤酒品牌：**百威啤酒、喜力啤酒和青岛啤酒** 是我国啤酒市场上的3 种知名品牌;
- 风味成分：乙醛、乙酸乙酯、异丁酯、乙酸异戊酯、异戊醇和己酸乙酯六种;
- 观测时间：跨度为半年, 隔一段时间测定三种啤酒上述六种风味成分含量;

通过**主成分分析法**后, 提取前两个主成分, 从图1第一主成分主要由乙酸乙酯、乙酸异戊酯和己酸乙酯构成的线性组合，代表了**啤酒的酯香** , 酯香越浓, 第一主成分的取值就越大。第二主成分主要由乙醛、异丁醇和异戊醇决定, 这些成分能够代表**啤酒的“酒劲”**, 这些成分含量越高, 第二主成分的了值就越大, 即啤酒的酒味就越重。这两个主成分可以反映全部信息的83.1%, 提取较为完全, 这说明这两个主成分替代原始的6 个风味成分反映的样品信息。

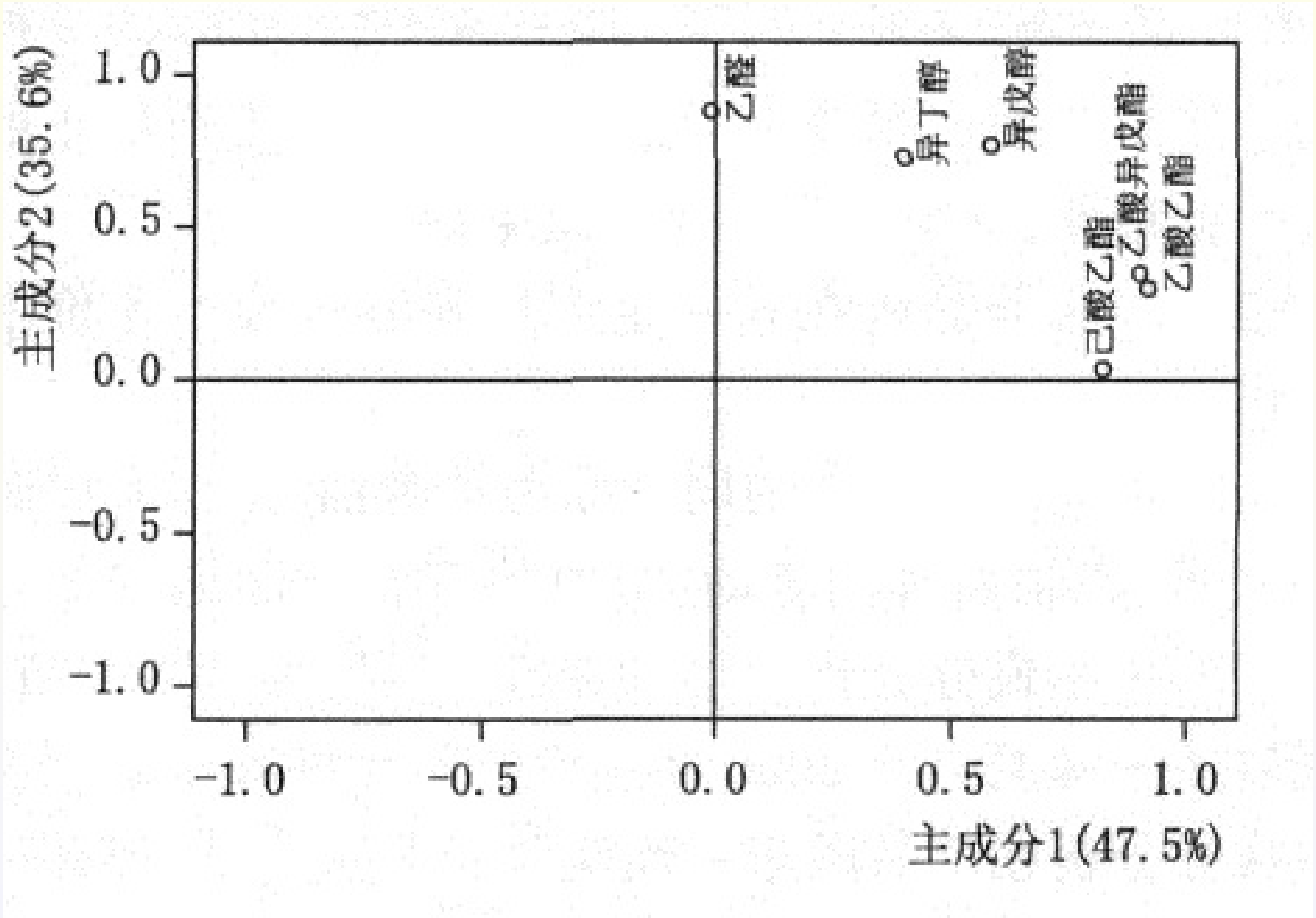


Figure: 不同品牌啤酒的主成分的载荷

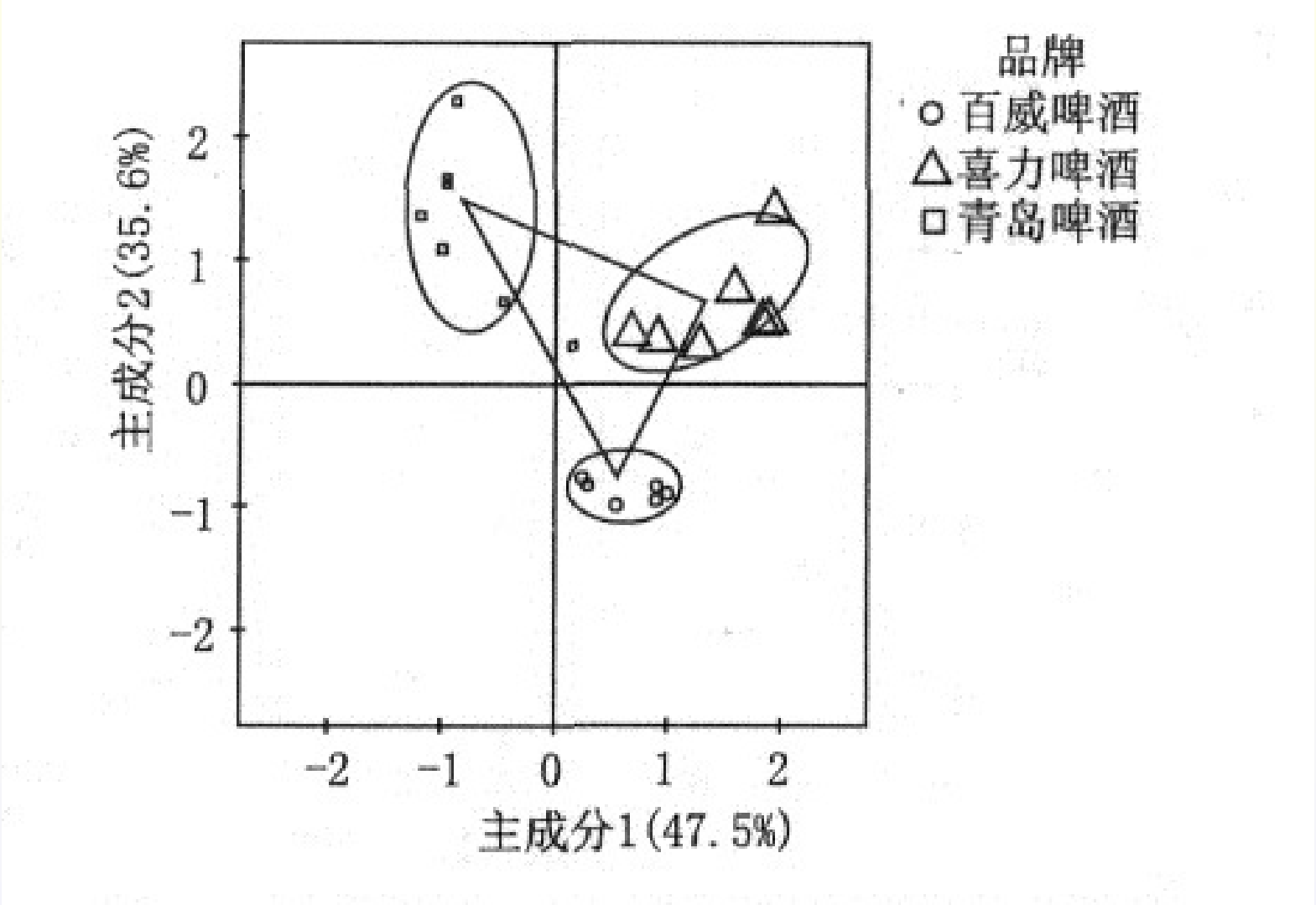


Figure: 不同品牌啤酒的主成分得分

对图2中的分类做出分析,

- 百威啤酒是酒味适中和酯香相对较浓的“浓香型”啤酒,
- 喜力啤酒是酒味和酯香均较浓的“浓醇型”啤酒,
- 青岛啤酒是酒味较重, 而酯香较弱的“ 醇型”啤酒.

案例二：主成分分析法回归

1. 研究背景：法国经济工作者希望通过国内总产值X_1、存储量X_2、总消费量X_3 去预测进口总额Y, 为此收集了1949 ~ 1958 年共十一年的数据。用回归分析得

$$\hat{Y} = -10.128 - 0.051X_1 + 0.587X_2 + 0.287X_3.$$

   但由于法国是一个原料进口国，因此当国内总产值X_1大时，进口总额Y也应大，所以X_1的系数大于0 才符合实际。
2. 原因分析：自变量X_1, X_2, X_3 存在多重共线性，是造成最小二乘法估计不好的原因。
3. 解决方法：采用主成分回归。

案例三：经济问题研究

1. 研究背景： 美国统计学家斯通(Stone)在1947年关于国民经济的研究。利用美国1929 – 1938年各年的数据，得到了17个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。
2. 主成分方法的应用： 通过主成分分析，以97.4%的精度，用3个新的综合变量取代了原始的17个变量。根据经济学知识，斯通给这三个新变量分别命名为总收入Z_1、总收入变化率Z_2和经济发展或衰退的趋势Z_3。
3. 统计分析： 斯通将他得到的主成分$Z_i,\ i=1,2,3$与实际测量的总收入I、总收入变化率ΔI以及时间t因素做相关分析，得：$r(Z_1,I)=0.995,\ r(Z_2,\Delta I)=0.948,\ r(Z_2,t)=-0.836.$

案例四：基于主成分分析的人脸特征提取

研究背景：人脸识别技术在信息安全领域应用较广，如公安系统、银行与海关的监控系统等。计算机人脸识别技术困难较多，关键在于人脸表情丰富；人脸受外部环境的影响等。另外，人脸识别还与图像处理、模式识别及神经网络等学科联系紧密。

主成分方法的应用： 主成分分析(Principal Component Analysis, PCA) 是当前用得最多的特征提取方法。研究证明PCA 提取出来的图像特征与人类视觉上感知细胞感知的特征有很大的相似性。

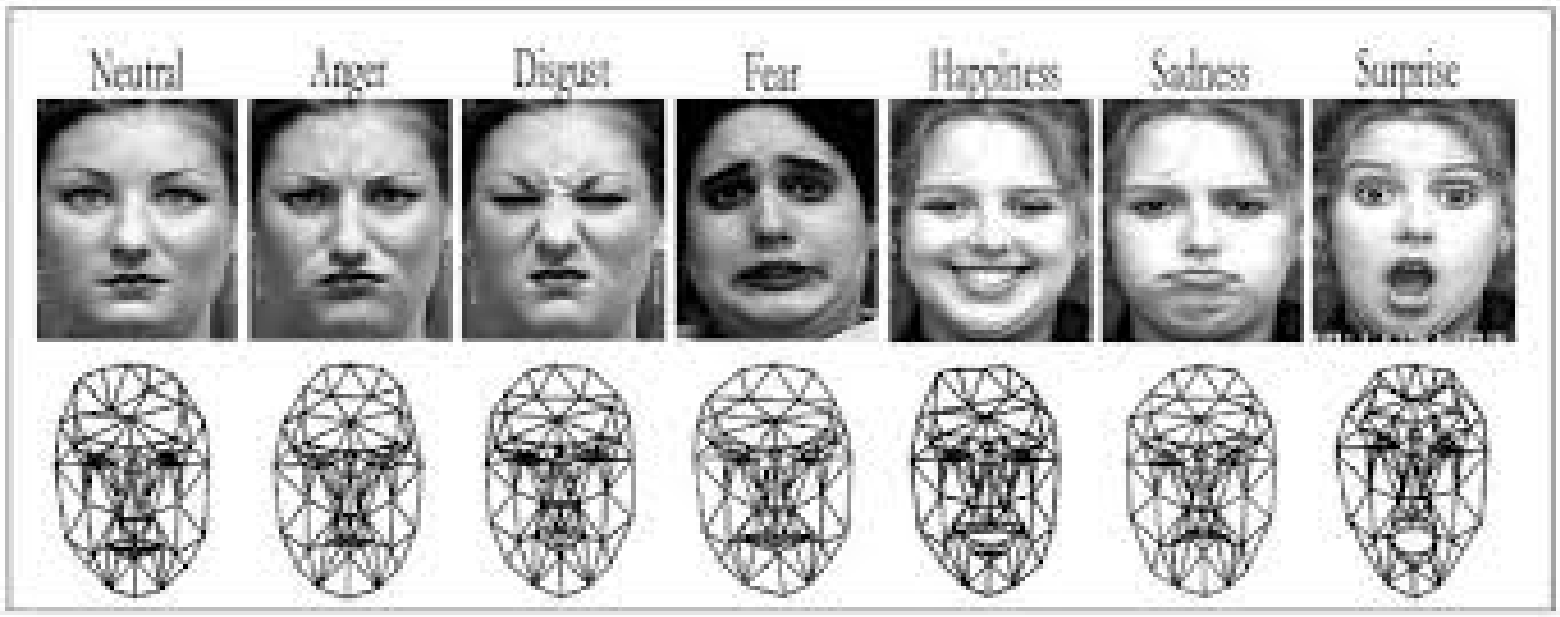


Figure: 人脸表情脸部特征提取

本课程中主成分分析方法与其它内容的联系

1. 多维正态分布检验：主成分检验法。
2. 判别分析：Fisher判别法。
3. 聚类分析：主成分聚类方法。
4. 回归分析：主成分回归法。
5. ……

在农业、生物、医学等领域中也有很多不适合用主成分分析的例子。如基因SNP的分析等。

一、主成分的定义

设$X=(X_1,X_2,\cdots,X_p)'$是p维随机向量,均值为$\mathrm{E}(X)=\mu$,协方差阵$D(X)=\Sigma$.考虑它的线性变换:

$$Z_1=a_1'X=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p,$$

$$Z_2=a_2'X=a_{12}X_1+a_{22}X_2+\cdots+a_{p2}X_p,$$

$$\cdots\cdots$$

$$Z_p=a_p'X=a_{1p}X_1+a_{2p}X_2+\cdots+a_{pp}X_p,$$

易见

$$\mathsf{Var}(Z_i)=a_i'\Sigma a_i\quad(i=1,2,\cdots,p)$$

$$\mathsf{Cov}(Z_i,Z_j)=a_i'\Sigma a_j\quad(i,j=1,2,\cdots,p)$$

用方差表示变量所包含的"信息".

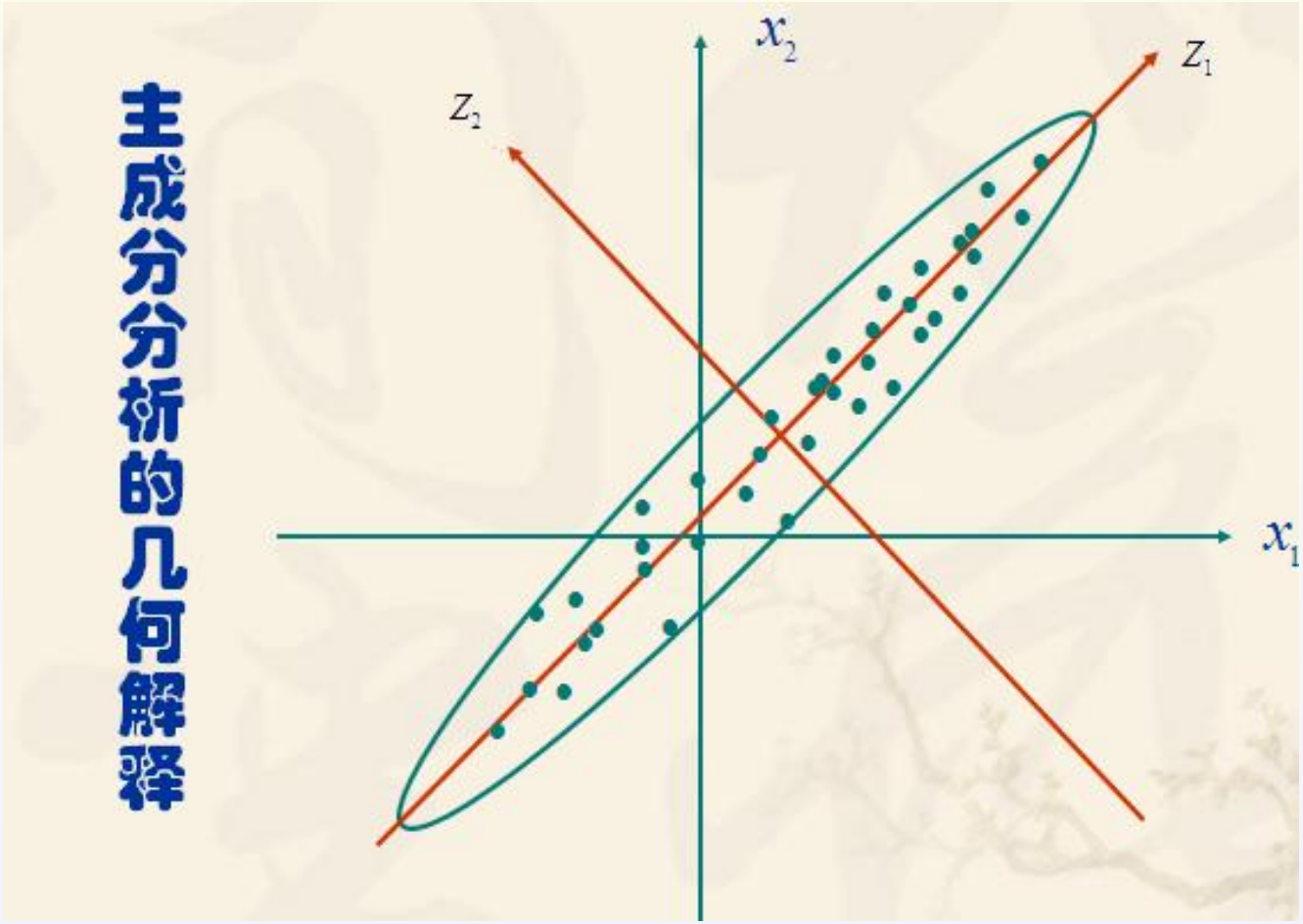
- 若希望用Z_1来尽可能反映原来的p个变量$X = (X_1, \cdots, X_p)'$, 必须对a_1 作出限制, 否则方差会趋于无穷大. 常用的限值为$a_1'a_1 = 1$. 在满足上述的约束条件的a_1,使$\mathrm{Var}(Z_1)$最大,称Z_1为第一主成分.
- 如果第一主成分不足以代表原始变量的信息,考虑Z_2, 但希望Z_2不重复包含Z_1的信息. 所以$\mathrm{Cov}(Z_1, Z_2) = a_2'\Sigma a_1 = 0$. 即求$Z_2$, 满足$a_2'a_2 = 1$ 和$\mathrm{Cov}(Z_1, Z_2) = a_2'\Sigma a_1 = 0$ 下, 求a_2使$\mathrm{Var}(Z_2)$ 最大.

定义7.1.1 设$X=(X_1,X_2,\cdots,X_p)'$是p维随机向量,称$Z_i=a_i'X$为X的第一主成分, 如果:

1. $a_i'a_i=1 \quad (i=1,\cdots,p)$
2. 当$i>1$时,$a_i'\Sigma a_j=0 \quad (j=1,\cdots,i-1)$
3. $\mathsf{Var}(Z_i)=\max_{\alpha'\alpha=1,\alpha'\Sigma a_j=0(j=1,\cdots,i-1)}\mathsf{Var}(\alpha'X)$

为了方便,我们在二维空间中讨论主成分的几何意义.

设有n 个样品，每个样品有两个观测变量x_1 和x_2, 在由变量x_1和x_2 所确定的二维平面中,n个样本点所散布的情况如椭圆状.



- 由图可以看出这n个样本点, 如果分别投影到z_1轴方向或z_2轴方向, n个样本点离散的程度可以z_1轴上投影点的方差和z_2轴上投影点的方差定量地表示.
- 如果只考虑z_1轴上投影和z_2轴上投影中的任何一个,那么包含在原始数据中的信息将会有较大的损失.
- 但如果投影和z_2轴上投影点的方差很小, 则说明将大部分的信息集中z_1轴上投影点上. 有时为了方便数据处理, 只考虑z_1轴上投影点.(起到了降维的作用)

如何设定旋转轴？

作旋转变换:

$$\begin{cases} Z_1 = \cos(\theta)X_1 + \sin(\theta)X_2 \\ Z_2 = -\sin(\theta)X_1 + \cos(\theta)X_2 \end{cases}$$

选择适当的旋转角度θ, 使原始数据经上述的旋转变换, 将的大部分信息集中到Z_1轴上(即Z_1的数据投影的方差远大于Z_2上的), 即Z_1上的投影点对数据中包含的信息起到了浓缩作用.

旋转变换的目的:

- 浓缩作用: 为了使得n个样品点在Z_1 轴方向上的离散程度最大，即Z_1 的方差最大。变量Z_1 代表了原始数据的绝大部分信息，在研究实际问题时，即使不考虑变量Z_2 也无损大局。经过上述旋转变换原始数据的大部分信息集中到Z_1 轴上，对数据中包含的信息起到了浓缩作用.
- 不相关的性质: Z_1, Z_2 除了可以对包含在X_1, X_2 中的信息起着浓缩作用之外，还具有不相关的性质，这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。二维平面上的个点的方差大部分都归结在Z_1 轴上，而Z_2轴上的方差很小。Z_1 和Z_2 称为原始变量X_1 和X_2 的综合变量。Z简化了数据结构，抓住了主要矛盾。

二、主成分的求法

设$X=(X_1,X_2,\cdots,X_p)'$是p维随机向量,均值为$\mathrm{E}(X)=\mu$,协方差阵$D(X)=\Sigma>0$.

用拉格朗日乘子法,求第一主成分$Z_1=a'X$.

$$\varphi(a_1)=\mathsf{Var}(a_1X)-\lambda(a_1'a_1-1)=a_1'\Sigma a_1-\lambda(a_1'a_1-1),$$

考虑

$$\begin{cases}\dfrac{\partial\varphi}{\partial a_1}=2(\Sigma-\lambda I)a_1=0\\ \dfrac{\partial\varphi}{\partial\lambda}=a_1'a_1-1=0\end{cases}$$

可以看出,λ是Σ的特征值,a_1是相应的单位特征向量.

定理7.1.1 设$X=(X_1,X_2,\cdots,X_p)'$是p维随机向量，$D(X)=\Sigma>0$. $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$是$\Sigma$的特征值，$a_1,a_2,\cdots,a_p$为相应的单位正交特征向量. 则$X$的第$i$主成分为

$$Z_i=a_i'X.$$

推论 设$Z=(Z_1,Z_2,\cdots,Z_p)'$是p维随机向量,则其分量Z_i $(i=1,\cdots,p)$ 依次是X的第i主成分的充要条件是:

1. $Z=A'X$, A为正交矩阵;
2. $D(Z)=\mathrm{diag}(\lambda_1,\cdots,\lambda_p)\hat{=}\Lambda$,即随机变量$Z$的协方差矩阵为对角矩阵;
3. $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$.

三、主成分的性质

记$\Sigma = (\sigma_{ij})$, $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$,其中$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$为$\Sigma$的特征值,$a_1, a_2, \cdots, a_p$为相应的单位正交特征向量.

性质1 $D(Z) = \Lambda$,即p 个主成分的方差为: $\mathsf{Var}(Z_i) = \lambda_i$ $(i = 1, \cdots, p)$, 且它们是互不相关的.

性质2 $\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i$,通常称$\sum_{i=1}^{p} \sigma_{ii}$为**原总体$X$的总方差(或称总惯量)**

注:

- 说明主成分分析把p个随机变量的总方差分解成为p个不相关的随机变量的方差之和。
- 如果A 是特征向量矩阵，Λ 是特征根所构成的对角阵，则有

$$\mathrm{tr}[\Sigma] = \mathrm{tr}[A\Lambda A'] = \mathrm{tr}[\Lambda].$$

性质3 主成分Z_k与原始变量X_i的相关系数$\rho(Z_k, X_i)$为

$$\rho(Z_k, X_i) = \sqrt{\lambda_k} a_{ik} / \sqrt{\sigma_{ii}}$$

并把主成分Z_k与原始变量X_i相关系数称为因子负荷量

证明： 显然

$$\rho(Z_k, X_i) = \frac{\mathsf{Cov}(Z_k, X_i)}{\sqrt{\mathsf{Var}(Z_k)\mathsf{Var}(X_i)}} = \frac{\mathsf{Cov}(a_k' X, e_i' X)}{\sqrt{\lambda_k \sigma_{ii}}},$$

其中：$e_i = (0, \cdots, 0, 1, 0, \cdots, 0)$.

$$\mathsf{Cov}(a_k' X, e_i X) = a_k' D(X) e_i = a_k' \Sigma e_i = \lambda_k e_i' a_k = \lambda_k a_{ik}.$$

性质4 $\sum_{k=1}^{p}\rho^2(Z_k,X_i)=\sum_{k=1}^{p}\frac{\lambda_k a_{ik}^2}{\sigma_{ii}}=1(i=1,2,\cdots,p)$

证明 设A是由特征向量组成的矩阵，Λ 是由特征根组成的对角阵，显然有$A'\Sigma A=\Lambda$. 即有$\Sigma=A\Lambda A'$，故有

$$\sigma_{ii}=(a_{i1},\cdots,a_{ip})\Lambda(a_{i1},\cdots,a_{ip})'=\sum_{k=1}^{p}\lambda_k a_{ik}^2.$$

性质5 $\sum_{i=1}^{p}\sigma_{ii}\rho^2(Z_k,X_i)=\lambda_k \quad (k=1,\cdots,p).$

证明 由**性质3** 可知，

$$\rho(Z_k,X_i)=\sqrt{\lambda_k}a_{ik}/\sqrt{\sigma_{ii}}$$

并且$\sum_{i=1}^{p}a_{ik}^2=1,$

$$\sum_{i=1}^{p}\sigma_{ii}\rho^2(Z_k,X_i)=\sum_{i=1}^{k}\lambda_k a_{ik}^2=\lambda_k.$$

定义7.1.2 我们称$\lambda_k/\sum_{i=1}^{p}\lambda_i$称为主成分$Z_k$的**贡献率**，称$\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i$为主成分$Z_1,\cdots,Z_m$的**累计贡献率**.

定义7.1.3 将前m个主成分$Z_1,\cdots,Z_m$对原始变量X_i的贡献率$\nu_i^{(m)}$定义为X_i与$Z_1,\cdots,Z_m$的相关系数平方和,它等于

$$\nu_i^{(m)}=\sum_{k=1}^{m}\lambda_k a_{ik}^2/\sigma_{ii}.$$

例7.1.1

设随机向量$X=(X_1,X_2,X_3)'$的协方差阵为

$$\Sigma=\begin{bmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{bmatrix}$$

试求X的主成分及主成分对变量X_i的贡献率$\nu_i(i=1,2,3)$.

解

求得Σ的特征值为$\lambda_1=3+\sqrt{8},\ \lambda_2=2,\ \lambda_3=3-\sqrt{8},$ 相应的单位正交向量为

$$a_1=(0.383,-0.924,0.000)' \quad a_2=(0,0,1)' \quad a_3=(0.924,0.383,0.000)'$$

故主成分为

$$\begin{aligned} Z_1 &= 0.383X_1 - 0.924X_2 \\ Z_2 &= X_3 \\ Z_3 &= 0.924X_1 + 0.383X_2 \end{aligned}$$

取$m = 1$时,Z_1对X的贡献率达$(3+\sqrt{8})/8 = 72.8\%$

取$m = 2$时,Z_1和Z_2对X的贡献率达97.85%.

列出m个主成分对对X_i的贡献率$\nu_i^{(m)}$.

i	$\rho(Z_1, X_i)$	$\rho(Z_2, X_i)$	$\nu_i^{(1)}$	$\nu_i^{(2)}$
1	0.925	0	0.856	0.856
2	-0.998	0	0.996	0.996
3	0.000	1	0.000	1.000

由上可见,

- 当$m = 1$时, Z_1 的贡献率已达72.8%, 比较理想, 但Z_1 对X_3 的贡献率为0. 这是因为Z_1 中没有包含X_3 的任何信息.
- 当取两个主成分, 且Z_1, Z_2 对X_i 的贡献率也较高.

四、标准化变量的主成分及性质

$$X_i^* = \frac{X_i - E(X_i)}{\sqrt{\mathrm{Var}X_i}} = \frac{X_i - \mu_i}{\sigma_i}(i = 1, 2, \cdots, p)$$

为标准化的随机变量.

性质1 $D(Z^*) = \Lambda^* = \mathrm{diag}(\lambda_1^*, \cdots, \lambda_p^*)$,其中$\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_p^*$为相关阵$R$ 的特征值.

性质2 $\sum_{i=1}^{p} \lambda_i^* = p.$

性质3 主成分Z_k^*与标准化变量X_i^*的相关系数为

$$\rho(Z_k^*, X_i^*) = \sqrt{\lambda_k^*} a_{ik^*},$$

其中$a_k^* = (a_{1k}^*, \cdots, a_{pk}^*)'$是R对应与$\lambda_k^*$的单位正交特征向量.

性质4 $\sum_{k=1}^p \rho^2(Z_k^*, X_i^*) = \sum_{k=1}^p \lambda_k^* (a_{ik}^*)^2 = 1,$
$(i = 1, 2, \cdots, p).$

性质5 $\sum_{i=1}^p \rho^2(Z_k^*, X_i^*) = \lambda_k^*, \quad (k = 1, \cdots, p).$

7.2 样本的主成分

在实际问题中,一般协方差阵Σ未知,需要通过样本来估计.

设样本数据阵为:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} X'_{(1)} \\ X'_{(2)} \\ \vdots \\ X'_{(n)} \end{bmatrix}$$

样本协方差阵和相关系数阵:

$$\begin{aligned} S &= \frac{1}{n-1}\sum_{t=1}^{n}(X_{(t)}-\bar{X})(X_{(t)}-\bar{X})' \overset{def}{=} (s_{ij})_{p\times p}; \\ R &= (r_{ij})_{p\times p}, \text{其中} r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \end{aligned}$$

我们可以用样本协方差S作为Σ的估计或用R作为总体相关阵的估计,然后按上节的方法即可获得样本的主成分.

一、样本主成分

假定观测数据都已经标准化,这时，样本协方差阵S 就是样本相关阵R.

$$R = \frac{1}{n-1} X'X$$

记:

- 相关阵R的p个主成分为$Z_1, Z_2, \cdots, Z_p$,
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 为R 的特征根,
- $a_1, a_2, \cdots, a_p$ 为对应特征根的单位正交特征向量,
- $A = (a_1, a_2, \cdots, a_p)$ 为正交矩阵.

二、主成分得分

- 第i个主成分$Z_i = a_i' X$.将第t个样品$X_{(t)} = (x_{t1}, \cdots, x_{tp})$的值代入$Z_i$的表达式,得到的值称为第$t$个样品在$i$个主成分的得分,记为$z_{ti}$.
- 主成分得分阵:

$$\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} Z'_{(1)} \\ Z'_{(2)} \\ \vdots \\ Z'_{(n)} \end{bmatrix}$$

$$\text{或} \overset{def}{=} (\mathbf{Z_1}, \mathbf{Z_2}, \cdots, \mathbf{Z_p})$$

三、样本主成分的性质

性质1

$$\bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_{(t)} = 0,$$

$$\mathbf{Z_i'Z_j} = \begin{cases} 0, \\ (n-1)\lambda_i \end{cases}$$

说明当$i \neq j$时,第i个主成分和第j个主成分正交.

性质2 $\sum_{i=1}^{p} \lambda_i = p$.称$\frac{\lambda_k}{p}$为样本主成分$Z_k$的**贡献率**;又称$\sum_{i=1}^{m} \lambda_i/p$为样本主成分$Z_1, \cdots, Z_m$的**累计贡献率**.

性质3 样本主成分具有使残差平方和最小的优良性.

主成分分析的目的之一是简化数据结构, 即用尽可能少的主成分$Z_1, \cdots, Z_m (m < p)$代替原来的p个变量,这样就把p个变量的n次观测数据简化为m个主成分的得分矩阵数据.

要求:

1. m个主成分反映的信息与原来p个变量所含信息差不多.
2. m个主成分又能够对数据所具有的意义进行解释.

主成分个数m的选取:

1. 按累计贡献率达到一定程度,如0.7或0.8.来确定.
2. 先计算p个特征值的平均数$\bar{\lambda}$,取大于$\bar{\lambda}$的特征值个数m.

R语言中有关主成份分析的函数

(1) princomp函数其格式为

$$princomp(formula, data = NULL, subset, na.action)$$

- formula 是没有响应变量的公式(类似回归分析、方差分析、但无响应变量)
- data是数据框(类似于回归分析、方差分析)

(2)princomp()函数格式为

$$princomp(x, cor = FALSE, scores = TRUE, covmat = NULL, subset = rep(TRUE, nrow(as.matrix(x))), \cdots)$$

- x是用于主成份分析的数据，以数值矩阵或表单的形式给出;
- cor 是逻辑变量，当为TRUE时表示用样本的相关矩阵R作主成份分析，当为FALSE时表示用样本的协方差阵S作主成份分析;
- covmat 是协方差阵，如果不用数据集x提供，可由协方差阵提供;

prcomp()函数与princomp()函数用法一样.

(3) summary()函数目的是提取主成份的信息，其使用格式为

$summary(object, loadings = FALSE, cutoff = 0.1, \cdots$

- object 是由princomp()得出的对象;
- loadings是逻辑变量，TRUE显示loadings函数的内容，FALSE则不显示.

(4) loadings()函数显示主成份或因子分析中载荷的内容，主成份分析中该内容实际上是主成份对应的各列，即前面分析的正交阵A(很多书上是记为Q)，在因子分析中，其内容就是载荷因子矩阵。格式：

$loadings(x)$ x是princomp()或factanal()得到的对象.

(5)predict()函数预测主成份的值，其使用格式为:

$predict(object, newdata, \cdots)$

- object是由princomp()得到的对象;
- newdata 是由预测值构成的数据框，当newdata为缺省时，预测已有数据的主成份值.

(6)screeplot()函数是画出主成份的碎石图，其使用格式为:

$$screeplot(x, npcs = min(10, length(x\$sdev)), type = c(\text{"}barplot\text{"}, \text{"}lines\text{"}), main = deparse(substitute(x)), \cdots)$$

- x是由princomp()得到的对象;
- nepcs 是画出的主成份的个数;
- type是描述画出的碎石图的类型，其中"barplot"是直方图类型，"lines"是直线图类型.

(7) biplot()函数画出数据关于主成份的散点图和原坐标在主成份下的方向，其使用格式为：

$biplot(x)$

- x 是由princomp()得到的对象;
- pc.biplot 是逻辑变量(缺省值为FALSE), 如果是TRUE, 用Gabriel(1971)提出的画图方法.

例7.2.1 (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纪30名学生,测量其身高(X_1),体重(X_2),胸围(X_3)和坐高(X_4),数据见下表.试对这30名中学生身体四项指标数据做主成分分析.

序号	X1	X2	X3	X4	序号	X1	X2	X3	X4
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

7.3 主成分分析的应用

假定本节的所有数据均已经标准化

- 样本协方差阵S就是样本相关阵R.

$$R = \frac{1}{n-1} X'X = (r_{ij})_{p\times p}$$

- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$为$R$的特征根,$a_1, a_2, \cdots, a_p$为相应的单位正交特征向量,
- 样本主成分为

$$Z_j = a_j' X \quad (j = 1, 2, \cdots, p)$$

- 根据满足累计贡献率$>P_0$,取前m个主成分,有样本观测数据可求得m个主成分的得分值:

$$z_{ij} = a_j' X_{(i)} = a_{1j}X_{i1} + a_{2j}X_{i2} + \cdots + a_{pj}X_{ip}$$

$$(i = 1, \cdots, n \quad j = 1, \cdots, m)$$

- $\mathbf{z_j} = (z_{1j}, z_{2j}, \cdots, z_{nj})'$为第$j$ 个主成分的得分值($j = 1, 2, \cdots, m$). 利用样本主成分性质3,可得出由主成分得分值估计变量X_k 的得分.

记

$$\mathbf{X_k}^* = (x_{1k}^*, x_{2k}^*, \cdots, x_{nk}^*)'$$

其中

$$x_{ik}^* = a_{k1}z_{i1} + \cdots + a_{km}z_{im}$$

$$X^* = \begin{bmatrix} x_{11}^* & \cdots & x_{1p}^* \\ \vdots & & \vdots \\ x_{n1}^* & \cdots & x_{np}^* \end{bmatrix} \stackrel{def}{=} (\mathbf{X_1^*}, \cdots, \mathbf{X_p^*})$$

可以证明:

$$\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x_{ij}^{*})^{2}=(n-1)\sum_{k=m+1}^{p}\lambda_{k}$$

故有

$$X\cong X^{*},\quad \text{且}\mathbf{X_k^*}=Z^{*}\begin{bmatrix}a_{k1}\\ \vdots\\ a_{km}\end{bmatrix}=\sum_{t=1}^{m}a_{kt}\mathbf{Z_t}$$

其中

$$Z^{*}=\begin{bmatrix}z_{11} & \cdots & z_{1m}\\ \vdots & & \vdots\\ z_{n1} & \cdots & z_{nm}\end{bmatrix}\overset{def}{=}(\mathbf{Z_1},\cdots,\mathbf{Z_m})$$

一、指标分类(变量分类)

如果第i个变量和第j个变量的相关系数$r_{ij} \approx 1$,显然这两个变量应归为一类.

(1)$\mathbf{X_i}$和$\mathbf{X_j}$表示两个变量的n次观测向量,考虑它们之间的距离:

$$||\mathbf{X_i} - \mathbf{X_j}||^2 = \mathbf{X_i}'\mathbf{X_i} - 2\mathbf{X_i}'\mathbf{X_j} + \mathbf{X_j}'\mathbf{X_j}$$

$$= (n-1)(r_{ii} - 2r_{ij} + r_{jj}) = 2(n-1)(1 - r_{ij})$$

$$\begin{aligned} 2(1-r_{ij}) &= ||\frac{1}{\sqrt{n-1}}\mathbf{X_i} - \frac{1}{\sqrt{n-1}}\mathbf{X_j}||^2 \\ &\approx ||\frac{1}{\sqrt{n-1}}\mathbf{X_i^*} - \frac{1}{\sqrt{n-1}}\mathbf{X_j^*}||^2 \\ &= ||\frac{1}{\sqrt{n-1}}[(a_{i1} - a_{j1})\mathbf{Z_1} + \cdots + (a_{im} - a_{jm})\mathbf{Z_m}]||^2 \\ &= \lambda_1(a_{i1} - a_{j1})^2 + \cdots + \lambda_m(a_{im} - a_{jm})^2 \end{aligned}$$

因为

$$\rho(Z_k, X_i) = \sqrt{\lambda_k} a_{ik} \overset{def}{=} \rho_{ik}$$

$\rho(Z_k, X_i)$是第k个主成分对第i个原始变量的因子负荷量.

$$2(1 - r_{ij}) = (\rho_{i1} - \rho_{j1})^2 + \cdots + (\rho_{im} - \rho_{jm})^2$$

若$r_{ij} \approx 1$,则有

$$(\rho_{i1} - \rho_{j1})^2 + \cdots + (\rho_{im} - \rho_{jm})^2 \approx 0$$

亦即$||\mathbf{X_i} - \mathbf{X_j}|| \approx 0$,即把第$i$个变量和第$j$个变量应归为一类.

例7.3.1服装定型分类问题

- 对128个成年男子的身材进行测量,每人各侧得16项指标: 身高(X_1),坐高(X_2),胸围(X_3),头高(X_4), 裤长(X_5),下档(X_6),手长(X_7),领围(X_8),前胸(X_9),后背(X_{10}),肩厚(X_{11}),肩宽(X_{12}),袖长(X_{13}), 肋围(X_{14}),腰围(X_{15}),和腿肚(X_{16}).
- 16项相关阵R见程序(因相关阵为对称矩阵,只给出相关阵的上三角部分).
- 试从相关阵R出发进行主成分分析,并对16项指标进行分类.

二、样品分类

对p个变量观测n次,得n个样品,记$X_{(i)} = (x_{i1}, \cdots, x_{ip})$为第$i$个样品,按距离相近的程度分类,即若

$$||X_{(i)} - X_{(j)}|| \approx 0$$

把第i个样品和第j个样品归为一类.

因原始数据阵$X \approx X^*$,故

$$||X_{(i)} - X_{(j)}|| \approx ||X^*_{(i)} - X^*_{(j)}||$$

因为

$$X^*_{(i)} = \begin{bmatrix} x^*_{i1} \\ \vdots \\ x^*_{ip} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pm} \end{bmatrix} = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{im} \end{bmatrix}$$

$$= (a_1, \cdots, a_m) \begin{bmatrix} z_{i1} \\ \vdots \\ z_{im} \end{bmatrix}$$

因

$$||X^*_{(i)} - X^*_{(j)}||^2 = ||a_1(z_{i1} - z_{j1} + \cdots + a_m(z_{im} - z_{jm})||^2$$

$$= (z_{i1} - z_{j1})^2 + \cdots + (z_{im} - z_{jm})^2$$

这样就把考察两个p维空间点的靠近程度转化为两个$m(m < p)$维空间点的靠近程度.若取$m = 2$,n个样品点可在平面上表示出,利用点的分布规律对其分类.

例7.3.2

服装定型分类问题（续例7.3.1）。仍然利用128人16项指标的观测数据，试对128人的服装尺寸进行分类（即样品分类问题：把128人分为几类，每类找出典型代表，以该代表的服装尺寸作为这一类的尺寸）。

解： 取m=2,求出两个主成分，并计算样本主成分得分值$Z_{(i)} = (z_{i1}, z_{i2})'(i = 1, 2, \cdots, 128)$. 这把128个点全部表示在平面上，利用平面散布图，把128个点分为七类：

第一类共有25个点，聚集中心是$Z_{(25)}$;

第二类共有14个点，聚集中心是$Z_{(114)}$;

第三类共有9个点，聚集中心是$Z_{(89)}$;

第四类共有7个点，聚集中心是$Z_{(112)}$;

第五类共有12个点，聚集中心是$Z_{(9)}$;

第六类共有20个点，聚集中心是$Z_{(47)}$;

第七类共有8个点，聚集中心是$Z_{(118)}$;

各种型号服装的生产数量也按25:14:9:7:12:20:8这样的比例来生产。

注意：这七类并没有把128个点全部包括在内，还有33个样品不能归入这七个类，可认为它们是一些特殊体形的样品。

三、样品排序或系统评估

对p元总体X的样本进行主成分分析往往不是最终目的，而常常是完成某个实际问题的一种手段。

在实际工作中常常会遇到多指标系统的排序问题，比如:

- 对某类企业的经济效益进行评估比较，影响企业经济效益的指标有很多，
- 更科学、更客观地将一个多指标问题综合为单个指数的形式。

主成分分析法为样品排序或多指标系统评估提供可行的方法。

- 设Z_1是标准化随机向量$X = (X_1, X_2, \cdots, X_p)'$的第一主成分。由主成分的性质可知，$Z_1$与原始标准化变量$X_1, X_2, \cdots, X_p$ 的综合相关程度最强.如果只选一个综合变量来代表原来所有的原始变量，最佳选择就是Z_1。
- 由于第一主成分Z_1对应于数据变异最大的方向，这说明Z_1是使数据信息损失最小、精度最高的一味综合变量，因此它可用于构造系统排序评估指数。

几点说明：

(1)第一主成分Z_1并不是总能够被用来作为排序评估指数. 如果

$$Z_1 = a_{11}X_1 + a_{21}X_2 + a_{31}X_3 + \cdots + a_{p1}X_p$$

中的系数$a_{i1}(i = 1, \cdots, p)$既有正又有负或者近似为零，说明Z_1与原始变量$X_1, X_2, \cdots, X_p$中有一部分为正相关，而另一部分为负相关或不相关，这是Z_1有可能是无序指数,不能用Z_1作为排序评估指数.

(2)一般情况下，$Z_2, Z_3, \cdots, Z_p$不适合用来构造排序评估指数。因$Z_k(k = 2, \cdots, p)$一般与原始变量$X_1, X_2, \cdots, X_p$中有一部分为正相关，而另一部分为负相关或不相关.

(3)传统的专家评估和第一主成分评估法的结合。把传统的专家调查研究的信息用来对主成分评估法进行修正.

四、主成分回归

在考虑因变量Y与p个自变量$X_1, \cdots, X_p$的回归模型中，当自变量间有较强的线性相关（多重共线性）时，利用经典的回归方法求回归系数的最小二乘估计，一般效果较差。利用主成分的性质，可由前m个主成分来建立主成分回归模型：

$$Y = b_0 + b_1 Z_1 + \cdots + b_m Z_m (m \leq p)$$

这样既简化了回归方程的结构，且消除了变量间相关性带来的影响；但另一方面，主成分回归也给回归模型的解释带来一定的复杂性，因为主成分是原始变量的线性组合，不是直接观测的变量，其含义有时不明确。在求得主成分回归方程后，经常又使用逆变换将其变为原始变量的回归方程。

当原始变量间有较强的多重共线性，其主成分又有特殊的含义时，往往采用主成分回归，其效果比较好。

例7.3.3

经济分析数据的主成分回归。考察进口总额Y与三个自变量：国内总产值x_1，存贮量x_2，总消费量x_3（单位均为10亿法郎）之间的关系。现收集了1949至1959年共11年的数据。对表7.6的数据试用主成分回归分析方法求进口总额与总产值、存贮量和总消费量的定量关系式。

五、主成分检验法

设$D(X)=\Sigma$,如果Σ是对角矩阵，即p维向量的分量间不相关，这时把p元正态性检验问题可转化为p个一元正态性检验问题。但一般Σ不是对角矩阵，即分量间是相关的。利用主成分分析方法，求得X的p个主成分$Z_1, Z_2, \cdots, Z_p$(不相关)，并由原样本值计算p个主成分得分值，作为p个不相关的综合变量的样本值。这时就把p元正态性检验问题化为p个一元综合变量（主成分）的正态性检验。这就是多元正态性检验的主成分检验法。实际检验时，利用主成分的性质，只需对前$m(m<p)$个主成分得分数据逐个做正态性检验。