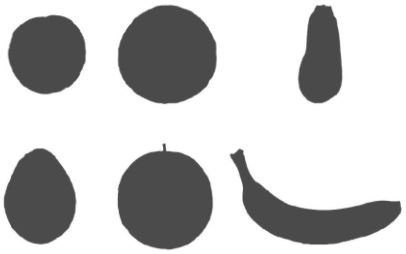
PCA, Principal Component Analysis

第十一讲 主成分分析

2024.4.10



方差=长度^2=信息

Recap

外积微分形式：

$$(dX) = \Lambda_{j=1}^{m} \Lambda_{i=1}^{n} dx_{ij} = J(X \to Y) \ \Lambda_{j=1}^{m} \Lambda_{i=1}^{n} dy_{ij} = J(dY)$$

计算积分（变量代换$X = g(Y)$）

$$I = \int_{X \in R^{n \times m}} f(X) dX = \int_{X \in R^{n \times m}} f(X)(dX)$$

$$= \int_{Y \in R^{n \times m}} f\big(g(Y)\big) J(X \to Y)(dY)$$

特别地，如果变量代换具有上三角形式（如球坐标变换，上三角分解），形式微分计算可大为简化（参见定理A5的证明）。

假设$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ iid $\sim N_m(0, \Sigma) \Leftrightarrow Z \sim N_{nm}(0, I_n \otimes \Sigma)$，　$A = W = Z^{\mathsf{T}} Z$，Z的概率密度$f(Z) = h(A)$仅依赖于A,求$A = W = Z^{\mathsf{T}} Z$ 的概率密度。

定理A8：Z唯一分解为(Schmidt正交化): $Z = H_1 T$，
其中 $H_1 = (\mathbf{h}_1, \ldots, \mathbf{h}_m)$是$n \times m$列正交矩阵，$H_1^{\mathsf{T}} H_1 = I_m$,$T$是$m \times m$上三角矩阵（对角元$> 0$），则

$$(dZ) = \prod_{i=1}^{m} t_{ii}^{n-i} \, (dT)(H_1^{\mathsf{T}} dH_1).$$

$$Z \longrightarrow (H_1, T) \longrightarrow (H_1, T^\top T = A)$$

$$Z = H_1 T \Rightarrow A = Z^\top Z = T^\top T$$

$$\Longrightarrow \ (dA) = 2^m \prod_{i=1}^{m} t_{ii}^{m-i+1} (dT)\ ,\ \text{定理A5}$$

$$\Longrightarrow (dZ) = \ 2^{-m} |A|^{\frac{n-m-1}{2}} (dA)(H_1^\top dH_1),\ \text{定理A8}$$

$$\Longrightarrow f(Z)(dZ) = h(A)\ 2^{-m} |A|^{\frac{n-m-1}{2}} (dA)(H_1^\top dH_1), \qquad (*)$$

此即定理A9, (*)右端对$(H_1^\top dH_1)$在$V_{m,n} = \{H_1 \in R^{n\times m}:\ \ H_1^\top H_1 =$

注：(*)式与上三角矩阵T无关，T只是起了方便计算微分外积的过渡作用。有理由相信，为了得到(*), 应该更简单的计算法则（外积微分形式还是过于复杂）。

主成分分析方法（PCA， principal component analysis）把多个相关变量线性组合成新的“变量”，如果少数几个组合能包含原来所有变量的大部分方差信息，那么我们可以用它们替代原来较多的变量。这些变量的组合称为主成分(PC, principal component)。

PCA是由K. Pearson、Hotelling独立发展出来的降维方法。

PCA实际上是SVD的一个特殊应用。

方差：随机变量的长度^2

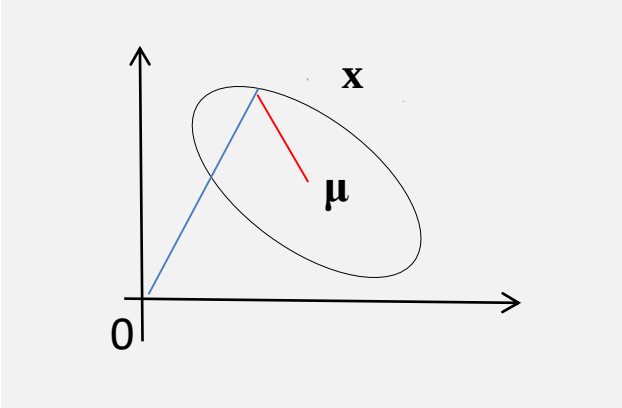
物体的长度

物体的长度以远端边界点之间的距离度量，或以边界点与中心的距离度量。

随机变量的“长度^2”：方差

随机变量作为函数，是有“形状”的数学对象，表现为其分布形状。

我们以方差$E(x-\mu)^2$而不是$E(x)^2$度量随机变量的“长度”（后者不代表几何形状的长度，参见下图）。



随机向量的各向“长度”

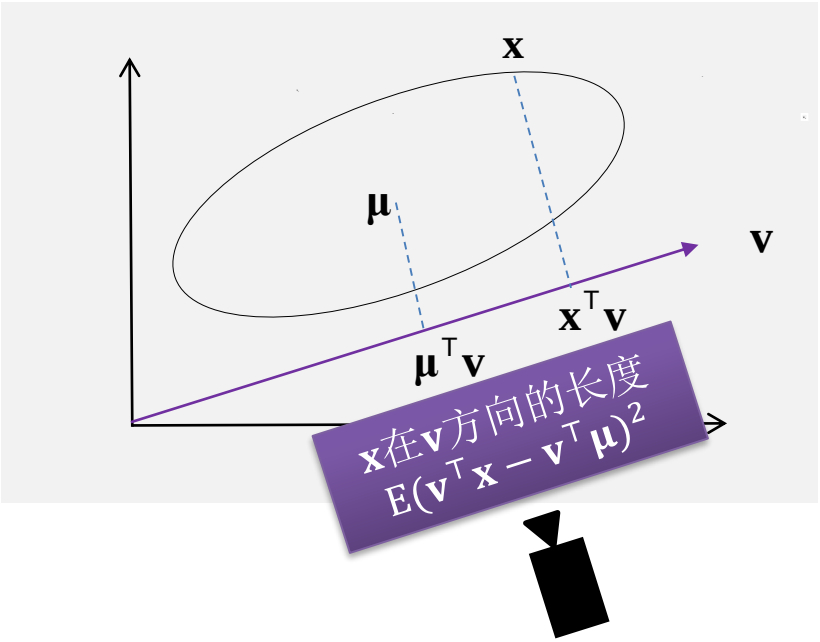
随机向量$\mathbf{x}$的空间轮廓在各个方向上都有长度，$E(\mathbf{x}) = \boldsymbol{\mu}, \mathrm{var}(\mathbf{x}) = \Sigma$。

对任何方向$\mathbf{v}$，$\|\mathbf{v}\| = 1$，

$\mathbf{x}$在$\mathbf{v}$上的投影坐标为$\mathbf{v}^{\mathsf{T}}\mathbf{x}$，中心$\boldsymbol{\mu}$的投影坐标为$\mathbf{v}^{\mathsf{T}}\boldsymbol{\mu}$，

随机向量在该方向上的方差或“长度^2”：

$$\mathrm{E}(\mathbf{v}^{\mathsf{T}}\mathbf{x} - \mathbf{v}^{\mathsf{T}}\boldsymbol{\mu})^2 = \mathbf{v}^{\mathsf{T}}\,\Sigma\,\mathbf{v}, \quad \|\mathbf{v}\| = 1,$$



PCA求解随机向量投影坐标方差最大的方向，此时投影坐标称为主成分。

$\mathbf{x}$与中心$\boldsymbol{\mu}$的平均距离$E\|\mathbf{x} - \boldsymbol{\mu}\|^2$不能体现出长度的方向性及分量之间的相依性。

总体PCA

我们首先考虑总体（单个随机向量）的主成分分析。

PCA寻找一个低维空间，特别地一组正交基，使得随机向量在其上的投影最大可能地保留原始数据的信息，信息以投影坐标的方差（长度^2）代表。

这些正交基称为主成分方向（或载荷），而随机向量在这些方向上的投影坐标称为主成分。

投影坐标

随机向量$\mathbf{x} \in R^p$在$\mathbf{v} \in R^p$上的投影为$P_{\mathbf{v}}\mathbf{x} = \mathbf{v}(\mathbf{v}^{\mathsf{T}}\mathbf{v})^{-1}\mathbf{v}^{\mathsf{T}}\mathbf{x} = \mathbf{v}c,$
投影坐标为$c = (\mathbf{v}^{\mathsf{T}}\mathbf{v})^{-1}\mathbf{v}^{\mathsf{T}}\mathbf{x}$。 特别地，若$\mathbf{v} \in S^{p-1}$,即$\| \mathbf{v} \| = 1,$
投影为$P_{\mathbf{v}}\mathbf{x} = \mathbf{v}[\mathbf{v}^{\mathsf{T}}\mathbf{x}]$， 投影坐标为$\mathbf{v}^{\mathsf{T}}\mathbf{x}$。

投影坐标的方差

假设$\Sigma = \mathrm{var}(\mathbf{x})$，对任何$\mathbf{v} \in S^{p-1}$,随机向量$\mathbf{x}_{p\times 1}$在$\mathbf{v}$上的投影坐标$\mathbf{v}^{\mathsf{T}}\mathbf{x}$的方差/长度为 $\mathrm{var}(\mathbf{v}^{\mathsf{T}}\mathbf{x}) = \mathbf{v}^{\mathsf{T}}\Sigma\mathbf{v}.$

假设$\Sigma = \mathrm{var}(\mathbf{x}), \| \mathbf{v} \| = 1$, *PCA*极大化$\mathbf{x}$在$\mathbf{v}$上投影坐标的方差：

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \mathbf{v}^{\mathsf{T}} \Sigma \mathbf{v}$$

如果认为该方向上的投影(主成分)不足以代表原始$\mathbf{x}$，我们在$\mathbf{v}_1$的正交补空间上再次寻找方向$\mathbf{v}_2$，使投影的方差$\mathbf{v}^{\mathsf{T}}\Sigma\mathbf{v}$最大，表述为

$$\max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1} \mathbf{v}^{\mathsf{T}} \Sigma \mathbf{v}$$

依此类推，得到最优方向$\mathbf{v}_1, \mathbf{v}_2, ... \mathbf{v}_p$，定理1说明它们是$\Sigma$的特征向量。

二次型的极值

定理1：若$\Sigma_{p \times p} > 0$, 其特征根为$\lambda_1 \geq ... \geq \lambda_p > 0$, 对应的正交、单位长特征向量为$\mathbf{v}_1, ..., \mathbf{v}_p$，则对任何$1 \leq k \leq p$,

$$\max_{\|\mathbf{x}\|=1} \mathbf{x}^{\mathsf{T}} \Sigma \mathbf{x} = \lambda_1, ..., \quad \max_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x} \perp \mathbf{v}_1, ..., \mathbf{v}_k}} \mathbf{x}^{\mathsf{T}} \Sigma \mathbf{x} = \lambda_{k+1},$$

上述极值分别在特征向量为$\mathbf{v}_1, ..., \mathbf{v}_k$处达到。

定理1的证明：

记$\Lambda = \mathrm{diag}(\lambda_1,...,\lambda_p)$，$\lambda_1 \geq ... \geq \lambda_p > 0$，$V = (\mathbf{v}_1,...,\mathbf{v}_p)$，$V^{\mathsf{T}}V = V^{\mathsf{T}}V = I_p$。

即Σ的谱分解为$\Sigma = V\Lambda V^{\mathsf{T}}$。

令$\mathbf{y} = V^{\mathsf{T}}\mathbf{x}$，则$\mathbf{x}^{\mathsf{T}}\mathbf{x} = \mathbf{y}^{\mathsf{T}}\mathbf{y}$，$\mathbf{x}^{\mathsf{T}}\Sigma\mathbf{x} = \mathbf{y}^{\mathsf{T}}\Lambda\mathbf{y}$，故

$$\frac{\mathbf{x}^{\mathsf{T}}\Sigma\mathbf{x}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} = \frac{\mathbf{y}^{\mathsf{T}}\Lambda\mathbf{y}}{\mathbf{y}^{\mathsf{T}}\mathbf{y}} = \sum_{i=1}^{p}\lambda_i y_i^2 / \sum_{i=1}^{p} y_i^2 \leq \lambda_1,$$

当$y_2 = ... = y_p = 0$时等号成立，即$\mathbf{x} \perp \mathbf{v}_2,...,\mathbf{v}_p$，故当$\mathbf{x} \propto \mathbf{v}_1$时等号成立。

对任何$\mathbf{x} \perp \mathbf{v}_1$，$y_1 = \mathbf{v}_1^{\mathsf{T}}\mathbf{x} = 0$，所以$\frac{\mathbf{x}^{\mathsf{T}}\Sigma\mathbf{x}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} = \frac{\mathbf{y}^{\mathsf{T}}\Lambda\mathbf{y}}{\mathbf{y}^{\mathsf{T}}\mathbf{y}} = \sum_{i=2}^{p}\lambda_i y_i^2 / \sum_{i=2}^{p} y_i^2 \leq \lambda_2.$

其它类似。

主成分定义

假设$\Sigma_{p\times p}=\mathrm{var}(\mathbf{x})$的谱分解为$\Sigma=V\Lambda V^{\mathsf{T}}$，其中$\Lambda=\mathrm{diag}(\lambda_1,...,\lambda_p), V=(\mathbf{v}_1,...,\mathbf{v}_p)$, $V^{\mathsf{T}}V=VV^{\mathsf{T}}=I_p$, 假设特征根$\lambda_1\geq...\geq\lambda_p>0$, 对应的相互正交的单位特征向量为$\mathbf{v}_1,...,\mathbf{v}_p$。由定理1,

$$\mathbf{v}_1=\underset{\|\mathbf{v}\|=1}{\mathrm{argmax}}\,\mathbf{v}^{\mathsf{T}}\Sigma\mathbf{v}$$

$\mathbf{v}_1$称为第一主成分方向/载荷，$y_1=\mathbf{v}_1^{\mathsf{T}}\mathbf{x}$为第一主成分 (PC1);

$\mathbf{v}_k=\underset{\substack{\|\mathbf{v}\|=1,\\ \mathbf{v}\perp\mathbf{v}_1,...,\mathbf{v}_{k-1}}}{\mathrm{argmax}}\,\mathbf{v}^{\mathsf{T}}\Sigma\mathbf{v}$称为第$k$主成分方向/载荷，$y_k=\mathbf{v}_k^{\mathsf{T}}\mathbf{x}$为第$k$主成分; $k=2,...,p$

所有主成分：$\mathbf{y}=\begin{pmatrix}\mathbf{v}_1^{\mathsf{T}}\mathbf{x}\\ \vdots\\ \mathbf{v}_p^{\mathsf{T}}\mathbf{x}\end{pmatrix}=V^{\mathsf{T}}\mathbf{x}\in R^p.$

$\mathbf{y}=V^{\mathsf{T}}\mathbf{x}$ 或 $\mathbf{y}=V^{\mathsf{T}}(\mathbf{x}-E(\mathbf{x})$称为的**主成分变换**

主成分的方差

$\Sigma = \mathrm{var}(\mathbf{x})$的谱分解：$\Sigma = V\Lambda V^{\mathsf{T}}$，

所有主成分：$\mathbf{y} = V^{\mathsf{T}}\mathbf{x}$，

主成分的方差：$\mathrm{var}(\mathbf{y}) = V^{\mathsf{T}}\Sigma V = \Lambda$，各个主成分不相关。

因为$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$，第一主成分PC1方差$\lambda_1$最大，

方差贡献率最大；第二主成分PC)方差λ_2第二大,....

主成分分析

主成分分析（PCA）对随机向量进行主成分变换(V^{T}旋转变换)：

- $\Sigma = \mathrm{var}(\mathbf{x})$的谱分解：$\Sigma = V\Lambda V^{\mathsf{T}}$，
- 所有主成分：$\mathbf{y} = V^{\mathsf{T}}\mathbf{x}$，

PCA用前k（$k < p$）个方差最大的主成分替代原随机向量：

$$\mathbf{x} = (x_1, ..., x_p)^{\mathsf{T}} \rightarrow \mathbf{y}_k = (y_1, ..., y_k)^{\mathsf{T}}$$

其中k的选取使得前k个主成分的方差之和$\sum_{i=1}^{k} \mathrm{var}(y_i) = \lambda_1 + ... + \lambda_k$

在总方差$\sum_{i=1}^{p} \mathrm{var}(y_i) = \sum_{i=1}^{p} \mathrm{var}(x_i) = \lambda_1 + ... + \lambda_p$的占比大于某个阈值：

$$(\lambda_1 + ... + \lambda_k)/(\lambda_1 + ... + \lambda_p) > 80\%.$$

载荷作为PC的组合系数，其大小、符号代表了PC的含义。

载荷矩阵

$V=(\mathbf{v}_1,\dots,\mathbf{v}_p)=(v_{ij})$,
行下标 i：变量。列下标 j：主成分

命题1. $\mathrm{var}(\mathbf{x})=\Sigma=V\Lambda V^{\mathsf{T}}$，主成分$\mathbf{y}=V^{\mathsf{T}}\mathbf{x}$，即$y_j=\mathbf{v}_j^{\mathsf{T}}\mathbf{x}=\sum_{i=1}^{p}v_{ij}x_i$，则

$$\mathrm{cov}(\mathbf{x},y_j)=\mathbf{v}_j\lambda_j,\quad \mathrm{cov}(x_i,y_j)=\lambda_j v_{ij}.$$

证：(1) $\mathrm{cov}(\mathbf{x},\mathbf{y})=\mathrm{cov}(\mathbf{x},V^{\mathsf{T}}\mathbf{x})=\Sigma V=V\Lambda=(\mathbf{v}_1\lambda_1,\dots,\mathbf{v}_p\lambda_p)$,

$\mathrm{cov}(\mathbf{x},y_j)=\mathrm{cov}(\mathbf{x},\mathbf{v}_j^{\mathsf{T}}\mathbf{x})=\Sigma\mathbf{v}_j=\mathbf{v}_j\lambda_j$，其第$i$个分量为$\mathrm{cov}(x_i,y_j)=\lambda_j v_{ij}$。

注1：载荷$\mathbf{v}_j=\mathrm{cov}(\mathbf{x},y_j)/\lambda_j\propto\mathrm{cov}(\mathbf{x},y_j)$;

注2：方程$y_j=\mathbf{v}_j^{\mathsf{T}}\mathbf{x}=\sum_{i=1}^{p}v_{ij}x_i$中的变量的系数(载荷）

$$v_{ij}=\mathrm{cov}(x_i,y_j)/\mathrm{var}(y_j)\propto\rho_{x_i,y_j}$$

这与线性回归$x_i\sim y_j$的系数的表达完全相同。

例1. 若$\Sigma = \sigma^2 I_p$,所有特征根都是σ^2，特征向量可取为任意一组正交基，各个主成分重要性相同。若$\Sigma = \text{diag}(\sigma_{11},....,\sigma_{pp})$,所有特征根$\sigma_{11},....,\sigma_{pp}$，特征向量$\mathbf{v}_k = (0,...,1,...,0)^{\mathsf{T}}$,,主成分$y_k = x_k, k = 1,...., p.$

例2. $\mathbf{x} = (x_1, x_2)^{\mathsf{T}} \sim (\mathbf{0}, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, 则其特征根为

$1+\rho$和$1-\rho$，对应的特征向量分别是$\mathbf{v}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$

两个主成分别为

$$y_1 = \mathbf{x}^{\mathsf{T}}\mathbf{v}_1 = (x_1 + x_2)/\sqrt{2} \propto x_1 + x_2$$

$$y_2 = \mathbf{x}^{\mathsf{T}}\mathbf{v}_2 = (x_1 - x_2)/\sqrt{2} \propto x_1 - x_2$$

- 若$\rho > 0$, 如果只能用一个线性组合替代$\mathbf{x}$，那么我们应该使用第一主成分$y_1 = (x_1 + x_2)/\sqrt{2}$，或等价地，$x_1 + x_2$或$\bar{x} = (x_1 + x_2)/2.$
- 若$\rho < 0$,则第一主成分为 $y_2 = \mathbf{x}^{\mathsf{T}}\mathbf{v}_2 = (x_1 - x_2)/\sqrt{2}$。

例2(续). 假设$p\times 1$向量$\mathbf{x}$已经标准化，其协方差矩阵（相关系数矩阵）

$$\mathrm{cov}(\mathbf{x})=\Sigma=\begin{pmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & \cdots & \rho\\ \vdots & \vdots & \ddots & \vdots\\ \rho & \rho & \cdots & 1\end{pmatrix}=\rho\mathbf{1}\mathbf{1}^{\mathsf{T}}+(1-\rho)I_p,\quad \rho>-\frac{1}{p-1}$$

因为

$$\Sigma\mathbf{1}=\rho\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{1}+(1-\rho)\mathbf{1}=(1+(p-1)\rho)\mathbf{1},$$

所以$1+(p-1)\rho>0$是一个特征根，对应的特征向量为$\mathbf{1}/\sqrt{p}$.

另外，对任何$\mathbf{v}\perp\mathbf{1}$,

$$\Sigma\mathbf{v}=\rho\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{v}+(1-\rho)\mathbf{v}=(1-\rho)\mathbf{v},$$

所以其它$p-1$个特征根都是$1-\rho$，任何$\mathbf{v}\perp\mathbf{1}$都是特征向量。

(1) $\rho>0$的情形: 最大特征根为$\lambda_1=1+(p-1)\rho$,

第一主成分： $y_1=\mathbf{1}^{\mathsf{T}}\mathbf{x}/\sqrt{p}=(x_1+\ldots+x_p)/\sqrt{p}$, 方差占比: $\dfrac{1+(p-1)\rho}{p}$.

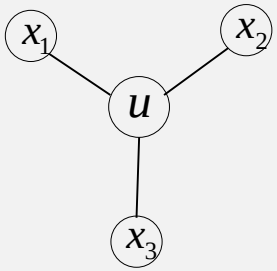
正定性要求

$$\rho>-\frac{1}{p-1}$$

$\rho > 0$时，由$\Sigma = \rho \mathbf{1}\mathbf{1}^{\mathrm{T}} + (1-\rho)I_p$的形式知$\mathbf{x}$可以表示为生成模型

$$\mathbf{x} = \mathbf{1}u + \boldsymbol{\varepsilon}, \quad \mathrm{var}(u) = \rho, \quad \mathrm{var}(\varepsilon_i) = 1-\rho,$$

其中$u, \varepsilon_1, ..., \varepsilon_p$独立，即$x's$由一个共同的变量$u$和若干独立的误差$\varepsilon's$决定。诸$x's$的对称性说明第一主成分取为$\bar{x}$是合理的。



(2) $-\dfrac{1}{p-1} < \rho < 0$的情形：

最大特征根为$1-\rho$(重数$1-p$)，最小特征根$1+(p-1)\rho$，类似于$\Sigma = \sigma^2 I_p$情形，少数几个主成分难以替代原始变量$\mathbf{x}$。

- 问题：两两负相关且相关性相同的生成模型？

例3. $\Sigma = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix}$，$\rho < 0$。$\lambda_1 = 2 > \lambda_2 = 1-\rho > \lambda_3 = 1+\rho$,

$$\mathbf{v}_1 = (0,\ 0,\ 1)^{\mathsf{T}},\ \ \mathbf{v}_2 = (1,-1,\ 0)^{\mathsf{T}}/\sqrt{2},\ \ \mathbf{v}_3 = (1,\ \ 1,\ 0)^{\mathsf{T}}/\sqrt{2}$$

主成分：

$$y_1 = \mathbf{v}_1^{\mathsf{T}}\mathbf{x} = x_3,\quad y_2 = \mathbf{v}_2^{\mathsf{T}}\mathbf{x} = (x_1 - x_2)/\sqrt{2},\quad y_3 = \mathbf{v}_3^{\mathsf{T}}\mathbf{x} = (x_1 + x_2)/\sqrt{2},$$

三个主成分解释总方差的比例分别是$\frac{2}{4}, \frac{1-\rho}{4}, \frac{1+\rho}{4}$.

软件不输出公式，只输出数值结果，基于输出载荷解释PC含义：

载荷矩阵$V = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$:

$$V = \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \end{matrix} \begin{matrix} \begin{matrix} \text{PC1} & \text{PC2} & \text{PC3} \end{matrix} \\ \begin{pmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \end{pmatrix} \end{matrix}$$

⇨

PC1的载荷

$$\mathbf{v}_1 = (0,0,1)^{\mathsf{T}} = \mathrm{cov}(\mathbf{x}, y_1)/\lambda_1$$

PC1: $y_1 = \mathbf{v}_1^{\mathsf{T}}\mathbf{x} = x_3$，与$x_1, x_2$无关，只与$x_3$有关

样本 PCA

样本PCA与总体PCA基本相同，只需将总体协方差矩阵Σ替换为样本协方差矩阵S，不同的是每个样本点都需做主成分变换。对样本方差矩阵

样本：$\mathbf{x}_1,...,\mathbf{x}_n \in R^p, \text{var}(\mathbf{x}_i) = \Sigma,$

数据矩阵：$X_{n\times p} = (\mathbf{x}_1,...,\mathbf{x}_n)^{\mathsf{T}},$

X的中心化：$X_c = (\mathbf{x}_1 - \bar{\mathbf{x}},...,\mathbf{x}_n - \bar{\mathbf{x}})^{\mathsf{T}} = X - \mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}},$

样本协方差矩阵：$S = X_c{}^{\mathsf{T}} X_c/(n-1),$

样本PCA

假设样本协方差矩阵S的特征根为$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, 对应的正交单位特征向量$V = (\mathbf{v}_1, \mathbf{v}_2,..., \mathbf{v}_p)$, 则$V$的各列是主成分方向/载荷, $\mathbf{x}_i - \bar{\mathbf{x}}$的主成分变换$\mathbf{y}_i = V^{\mathsf{T}}(\mathbf{x}_i - \bar{\mathbf{x}})$, 其第$j$分量称为第$j$主成分，$j = 1,.., p.$ 所有样本的所有主成分构成主成分矩阵：

$$Y = (\mathbf{y}_1,...,\mathbf{y}_n)^{\mathsf{T}} = X_c V$$

(有时，XV也称为是主成分矩阵). Y的第 j列为第j主成分, 其样本方差为λ_j, $j = 1,..., p$。

降维 - 选取部分主成分

主成分变换是1-1变换，通常选取累计方差贡献比率接近100%的几个主成分代表原来的所有变量，从而达到降维的目的。

主成分的方差

$Y = XV$的第j列为所有样本点的第j主成分 $\mathbf{y}_{(j)} = (y_{1j},...,y_{nj})^{\mathsf{T}}$ $= X\mathbf{v}_j = (\mathbf{x}_1^{\mathsf{T}}\mathbf{v}_j,...,\mathbf{x}_1^{\mathsf{T}}\mathbf{v}_j)^{\mathsf{T}}$的样本方差为$\lambda_j$。所有主成分的总方差为：

$$\lambda_1 + ... + \lambda_p = tr(V^{\mathsf{T}}SV) = tr(S)$$

第j个主成分所能解释的方差占比： $\lambda_j / (\lambda_1 + ... + \lambda_p)$。

前k个主成分的累计方差贡献比率：$\dfrac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p}$

主成分的选取

主成分个数的选取：取k使得累积方差贡献率 $\dfrac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p} > 80\%,$

画图时，选$k = 2$或3.

```
# PCA函数： princomp （n > p）
mypca = princomp(covmat=R) # R: 协方差矩阵，总体PCA
mypca = princomp(x) #x: 数据矩阵，样本PCA
summary(mypca) #特征根/方差、贡献率
mypca$loadings #载荷
mypca$scores #PC
```

例4. 112个孩子进行了6项测试

general:综合，picture:绘画，blocks:积木，
maze:迷宫， reading:阅 读, vocab:词汇量

其中general是智商测试。各项测试的成绩总分不同，故我们考虑标准化数据的协方差矩阵，即原始数据的相关系数矩阵：

$R =$

	general	picture	blocks	maze	reading	vocab
general	1	0.47	0.55	0.34	0.58	0.51
picture	0.47	1	0.57	0.19	0.26	0.24
blocks	0.55	0.57	1	0.45	0.35	0.36
maze	0.34	0.19	0.45	1	0.18	0.22
reading	0.58	0.26	0.35	0.18	1	0.79
vocab	0.51	0.24	0.36	0.22	0.79	1

相关系数阵的特征根：

主成分	1	2	3	4	5	6
特征根λ_j/方差	3.07	1.14	0.82	0.41	0.35	0.20
累积方差贡献率	0.51	0.70	0.84	0.91	0.97	1.00

总方差$=\lambda_1+\ldots+\lambda_6=tr(R)=6$

第一主成分能解释总方差的$3.07/6=51\%$,

第二主成分解释$1.14/6=19\%$, 累积70%

第三主成分解释$0.82/6=14\%$, 累积84%.

各个主成分的含义可参看载荷矩阵（下页）

载荷矩阵V

	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{v}_5$	$\mathbf{v}_6$
	pc1	pc2	pc3	pc4	pc5	pc6
general	0.47	0	0	0.86	0	-0.19
picture	0.36	0.4	0.6	-0.24	0.54	0
blocks	0.43	0.4	0	-0.25	-0.76	0
maze	0.29	0.41	-0.79	0	0.35	0
reading	0.44	-0.51	0	0	0	0.73
vocab	0.43	-0.5	0	-0.37	0	-0.65

$\mathbf{v}_1$：各个分量符号相同，取值近似，故PC1代表平均成绩,综合能力。

$\mathbf{v}_2 = (0,0.4,0.4,0.4,-0.5,-0.5)^{\mathsf{T}}$：general载荷0，PC2与*general*无关，与*picture*,*maze*,*blocks*正相关，与*reading*,*vocab*负相关：

$$\mathrm{PC2} \approx 0.4*(\mathit{picture}+\mathit{maze}+\mathit{blocks})-0.5*(\mathit{reading}+\mathit{vocab})$$

故PC2代表了行为能力(数学、空间)与语言能力的差别。

PC2，PC3在general上载荷为0，度量了综合能力之外的特殊能力

如果对成绩的协方差矩阵（而不是相关系数矩阵）进行PCA分析，前两个主成分贡献89%, 但载荷的含义不容易解释。所以，当随机向量的各个分量尺度(scale)不同时，最好基于相关系数矩阵进行PCA分析。

协方差矩阵$S=$

	general	picture	blocks	maze	reading	vocab
general	24.641	5.991	33.52	6.023	20.755	29.701
picture	5.991	6.7	18.137	1.782	4.936	7.204
blocks	33.52	18.137	149.831	19.424	31.43	50.753
maze	6.023	1.782	19.424	12.711	4.757	9.075
reading	20.755	4.936	31.43	4.757	52.604	66.762
vocab	29.701	7.204	50.753	9.075	66.762	135.292

S的特征根

主成分	1	2	3	4	5	6
特征根λ_j/方差	237.09	102.04	17.57	11.66	9.39	4.02
累积方差贡献率	0.62	0.89	0.93	0.96	0.99	1.00

S的特征向量/载荷矩阵$V=$

	pc1	pc2	pc3	pc4	pc5	pc6
general	0.23	0	0.55	0.65	-0.42	-0.19
picture	0	0	0	0	-0.22	0.96
blocks	0.63	-0.74	-0.13	-0.18	0	0
maze	0	0	0	0.55	0.82	0.14
reading	0.37	0.3	0.67	-0.47	0.32	0.03
vocab	0.63	0.59	-0.47	0.13	-0.11	0

例5.课本Table8.4给出的是5个公司股票的周收益率数据（103周），5家公司为JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, Exxon Mobil, 各周之间的数据是独立的。5个股票的协方差矩阵 ($x10^{-4}$) 如下

	JP Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
JP Morgan	4.33	2.76	1.59	0.64	0.89
Citibank	2.76	4.39	1.8	1.81	1.23
Wells Fargo	1.59	1.8	2.24	0.73	0.61
Royal Dutch Shell	0.64	1.81	0.73	7.22	5.08
Exxon Mobil	0.89	1.23	0.61	5.08	7.66

```
## R 函数princomp
Weekly.return = read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/T8-4.DAT")
dimnames(Weekly.return)[[2]]=c("JP Morgan","Citibank", "Wells Fargo", "Royal Dutch Shell","Exxon Mobil")

cov(Weekly.return)->v  ##计算协方差矩阵
mypca = princomp(covmat =v)  ##总体PCA：基于协方差矩阵求解主成分
Mypca =princomp(Weekly.return) ## 样本PCA：基于原始数据，与上一行结果相同
summary(mypca,loading=T) ##结果汇总
```

> summary(mypca, loading=T) #R 函数princomp分析结果输出：

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	0.037	0.026	0.016	0.012	0.011
Proportion of Variance	0.53	0.27	0.10	0.06	0.05
Cumulative Proportion	0.53	0.80	0.90	0.95	1.00

$\leftarrow \sqrt{\lambda_j}$

$\leftarrow \lambda_j / (\lambda_1 + \ldots + \lambda_5)$

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
JP Morgan	0.223	+0.625	0.326	0.663	0.118
Citibank	0.307	+0.570	-0.250	-0.414	-0.589
Wells Fargo	0.155	+0.345	0	-0.497	0.780
Royal Dutch Shell	0.639	-0.248	-0.642	0.309	0.148
Exxon Mobil	0.651	-0.322	0.646	-0.216	

PC1可称为“**市场成分**”：

载荷$\mathbf{v}_1$的各分量符号相同，PC1是各股票的加权平均，该加权平均具有最大的方差，反映了市场波动性。其中最后2个股票对PC1(市场波动性)影响最大.

PC2可称为“**行业成分**”：

载荷$\mathbf{v}_2$的最后2个股票为负(石油)，前三个为正(为银行)，故这两类股票的差距具有第二大的方差或变化，