大数据的矩阵计算基础——第10周

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

关注炼数成金企业微信

■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

多维数据

◆ 多维随机变量的矩阵表示

例 1 一个二维数据的例子是，N 个大学生关于体重和身高的一组数据，令 X_j 表示 $\mathbb{R}^2$ 中的观测向量，它列出第 j 个学生的体重和身高，如果用 w 表示体重，h 表示身高，那么观测矩阵的形式为：

$$\begin{array}{cccc} \begin{bmatrix} w_1 & w_2 & \cdots & w_N \\ h_1 & h_2 & \cdots & h_N \end{bmatrix} \\ \uparrow \quad\ \uparrow \quad\quad\quad\ \uparrow \\ X_1 \quad X_2 \quad\quad\quad X_N \end{array}$$

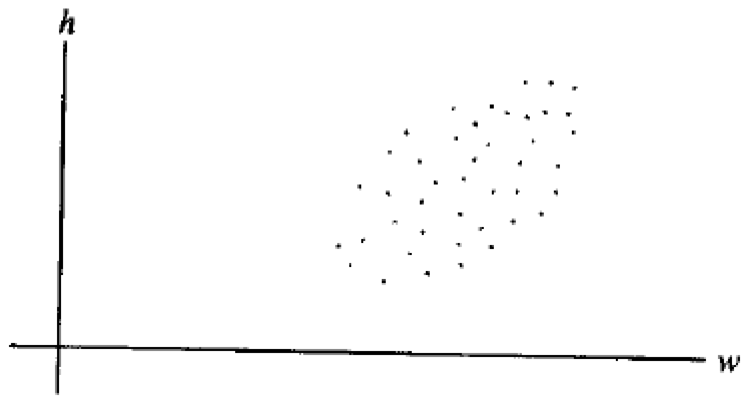


图 7-18 观测向量 $X_1,\cdots,X_N$ 的散列图

均值、协方差矩阵

- 样本均值

$$M = \frac{1}{N}(X_{.1} + X_{.2} + \cdots \ldots + X_{.N})$$

- 中心化

$$\hat{X}_{.k} = X_{.k} - M$$

$$B = [\hat{X}_{.1}, \hat{X}_{.2}, \ldots \ldots, \hat{X}_{.N}]$$

- 样本协方差

$$S = \frac{1}{N-1} BB^T$$

均值、协方差

例 3　从一个总体中随机取出 4 个样本作三次测量，每一个样本的观测向量为：

$$\boldsymbol{X}_1=\begin{bmatrix}1\\2\\1\end{bmatrix},\ \boldsymbol{X}_2=\begin{bmatrix}4\\2\\13\end{bmatrix},\ \boldsymbol{X}_3=\begin{bmatrix}7\\8\\1\end{bmatrix},\ \boldsymbol{X}_4=\begin{bmatrix}8\\4\\5\end{bmatrix}$$

计算样本均值和协方差矩阵.

协方差与协方差矩阵

- 协方差矩阵$S = (s_{ij})$
- 对角线元素s_{ii}
- 其他元素s_{ij}
- 总方差

主成分分析的实际背景

- 真实的训练数据总是存在各种各样的问题：
- 1、 比如拿到一个汽车的样本，里面既有以“千米/每小时”度量的最大速度特征，也有“英里/小时”的最大速度特征。
- 2、 拿到一个数学系的本科生期末考试成绩单，里面有三列，一列是对数学的兴趣程度，一列是复习时间，还有一列是考试成绩。
- 3、 拿到一个样本，特征非常多，而样例特别少。比如北京的房价：假设房子的特征是（大小、位置、朝向、是否学区房、建造年代、是否二手、层数、所在层数），有这么多特征，结果只有不到十个房子的样例，这时回归会出现问题
- 4、 在信号传输过程中，由于信道不是理想的，信道另一端收到的信号会有噪音扰动，那么怎么滤去这些噪音呢？

主成分分析PCA

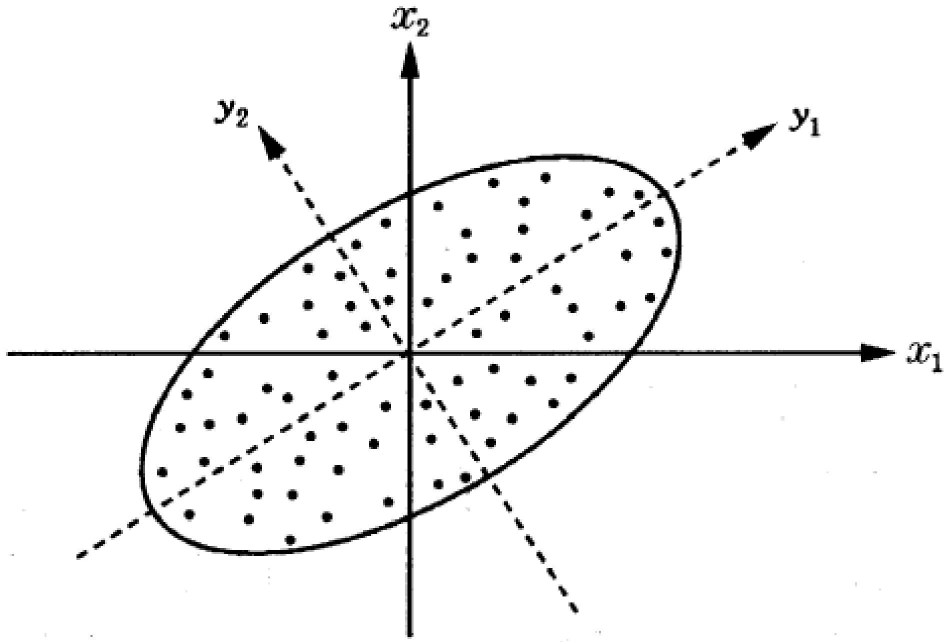
- 主成分分析（或称主分量分析，principal component analysis）由皮尔逊（Pearson,1901）首先引入，后来被霍特林（Hotelling,1933）发展了。
- 主成分分析是一种通过降维技术把多个变量化为少数几个主成分（即综合变量）的统计分析方法。这些主成分能够反映原始变量的绝大部分信息，它们通常表示为原始变量的某种线性组合。
- 主成分分析的一般目的是：

  (1)变量的降维；

  (2)主成分的解释。

主成分分析PCA

- 旋转变换



主成分分析PCA

- 对于p维随机变量$X=(x_1,x_2,\dots\dots,x_p)^T$,数学期望$E(X)=\boldsymbol{\mu}$,

协方差矩阵$V(X)=\Sigma=\frac{1}{N}(X-\mu)(X-\mu)^T$

- 考虑这样的线性变换

$$\begin{cases} Z_1=a_1^T X \\ Z_2=a_2^T X \\ \quad\vdots \\ Z_p=a_p^T X \end{cases},$$

- 显然

$$\begin{aligned} \mathrm{Var}(Z_i) &= a_i^T\Sigma a_i, \quad i=1,2,\cdots,p, \\ \mathrm{Cov}(Z_i,Z_j) &= a_i^T\Sigma a_j, \quad i,j=1,2,\cdots,p, \quad i\neq j. \end{aligned}$$

主成分分析PCA

DATAGURU炼数成金

- 我们希望寻找合适的a_1使得Z_1方差最大，即a_1是约束优化问题

$$\begin{array}{ll} \max & a^T\Sigma a \\ \text{s.t.} & a^T a = 1 \end{array}$$

的解 ⟶ 二次型的条件优化问题

$$m = \min\{\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x} : \|\boldsymbol{x}\| = 1\} \text{ , } M = \max\{\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x} : \|\boldsymbol{x}\| = 1\} \qquad (2)$$

定理 6 设 A 是对称矩阵，且 m 和 M 的定义如(2)式所示，那么 M 是 A 的最大特征值 λ_1，m 是 A 的最小特征值，如果 $\boldsymbol{x}$ 是对应 M 的单位特征向量 $\boldsymbol{u}_1$，那么 $\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x}$ 的值等于 M，如果 $\boldsymbol{x}$ 是对应 m 的单位特征向量，$\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x}$ 的值等于 m.

故a_1是协方差矩阵最大的特征值对应的单位特征向量。称$Z_1 = a_1^T X$为第一主成分。

主成分分析PCA

- 类似的，可以求出与第一主成分正交的第二主成分

定理 7 设 A, λ_1 和 u_1 如定理 6 所示. 在如下条件限制下

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}=1,\ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{u}_1=0$$

$\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x}$ 的最大值是第二大特征值 λ_2，且这个最大值，可以在 $\boldsymbol{x}$ 是对应 λ_2 的特征向量 $\boldsymbol{u}_2$ 处达到.

- 还有第三主成分，第四主成分......

定理 8 设 A 是一个 $n\times n$ 对称矩阵，且其正交对角化为 $A=PDP^{-1}$，将对角矩阵 D 上的元素重新排列，使得 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$，且 P 的列是其对应的单位特征向量 $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n$. 那么对 $k=2,\cdots,n$ 时，在以下限制条件下

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}=1,\ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{u}_1=0,\ \cdots,\boldsymbol{x}^{\mathrm{T}}\boldsymbol{u}_{k-1}=0$$

$\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x}$ 的最大值是特征值 λ_k，且这个最大值在 $\boldsymbol{x}=\boldsymbol{u}_k$ 处可以达到.

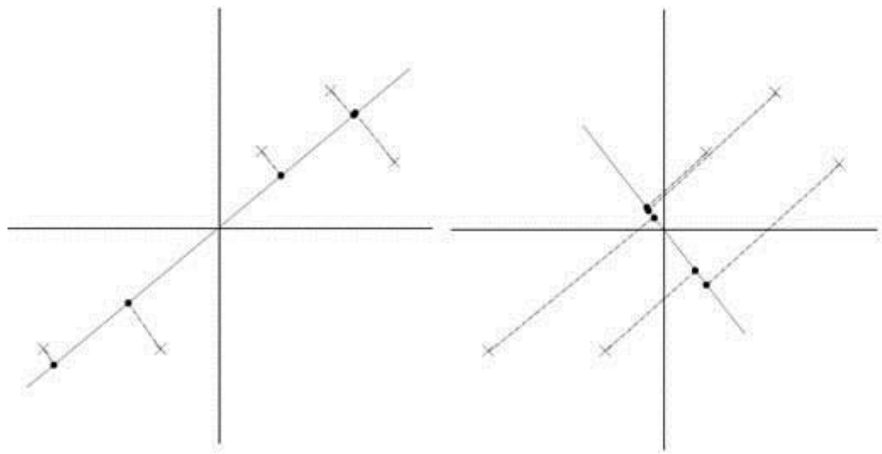
主成分分析PCA

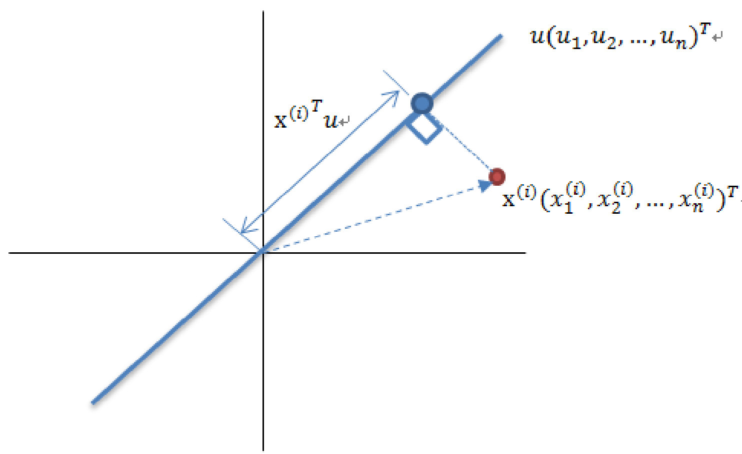
- 对于变量代换$Z = Q^T X$
- 协方差矩阵可正交对角化

$$Q^T \Sigma Q = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix}, \qquad (9.4)$$

且 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. 则矩阵 Q 的第 i 列就对应于 a_i, 相应的 Z_i 为第 i 主成分.

为什么要最大方差





主成分的性质

- 均值
- 协方差阵
- 总方差
- 变量与主成分的相关系数
- 方差贡献率

协方差阵与相关阵

- 标准化变量 $X_j^* = \dfrac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}, \quad j = 1, 2, \cdots, p.$

- 相关矩阵 $X^* = (X_1^*, X_2^*, \cdots, X_p^*)^T$ 的方差矩阵就是 X 的相关矩阵 R.

相关阵

◆ 相关矩阵的性质

(1) $E(Z^*)=0$, $\mathrm{Var}(Z^*)=\Lambda^*$, 其中 $\Lambda^*=\mathrm{diag}(\lambda_1^*,\lambda_2^*,\cdots,\lambda_p^*)$.

(2) $\sum_{i=1}^{p}\lambda_i^*=p$.

(3) 变量 X_j^* 与主成分 Z_i^* 之间的相关系数

$$\rho(X_j^*,Z_i^*)=\sqrt{\lambda_i^*}q_{ji}^*,\quad j,i=1,2,\cdots,p.$$

(4) 主成分 $Z_1^*,Z_2^*,\cdots,Z_m^*$ 对 X_j^* 的贡献率

$$\rho_{j\cdot 1\cdots m}^2=\sum_{i=1}^{m}\rho^2(X_j^*,Z_i^*)=\sum_{i=1}^{m}\lambda_i^*{q_{ji}^*}^2.$$

(5)

$$\rho_{j\cdot 1\cdots p}^2=\sum_{i=1}^{p}\rho^2(X_j^*,Z_i^*)=\sum_{i=1}^{p}\lambda_i{q_{ji}^*}^2=1.$$

样本主成分

- 如何求样本数据的主成分
- 1. 将样本数据中心化
- 2. 计算样本数据的协方差矩阵
- 3. 求出协方差矩阵的特征值与正交单位特征向量

例子

例 4 铁路峡谷（例 2）的多谱图像的初始数据包含 $\mathbb{R}^3$ 中 4 百万个向量，其协方差矩阵是

$$S=\begin{bmatrix} 2\,382.78 & 2\,611.84 & 2\,136.20 \\ 2\,611.84 & 3\,106.47 & 2\,553.90 \\ 2\,136.20 & 2\,553.90 & 2\,650.71 \end{bmatrix}$$

求数据的主成分，且列出由第一主成分确定的新变量.

数据降维

- 正交变换X=AY不改变数据的总方差

- 通过选取前k个主成分，包含了数据大部分的信息，达到了降维的效果

例 4　铁路峡谷（例 2）的多谱图像的初始数据包含 $\mathbb{R}^3$ 中 4 百万个向量，其协方差矩阵是

$$S=\begin{bmatrix} 2\,382.78 & 2\,611.84 & 2\,136.20 \\ 2\,611.84 & 3\,106.47 & 2\,553.90 \\ 2\,136.20 & 2\,553.90 & 2\,650.71 \end{bmatrix}$$

求数据的主成分，且列出由第一主成分确定的新变量.

奇异值分解（SVD分解）

- $m \times n$对角矩阵$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$
- $m \times m$正交矩阵U——左奇异向量
- $n \times n$正交矩阵V——右奇异向量
- A的一个奇异值分解：$A = U\Sigma V^T$

奇异值分解（SVD分解）

- 奇异值分解三部曲：
- 1. 将矩阵$A^T A$正交对角化
- 2. 算出V和Σ
- 3. 构造U

奇异值分解的应用

- 计算存储图形——将图形分解成象素（pixels）的一个矩形的数阵，其中的信息就可以用一个矩阵A=（a_{ij}）$_{m\times n}$来存储。矩阵A的元素a_{ij}是一个正的数，它相应于象素的灰度水平（gray level）的度量值。
- 由于一般来讲，相邻的象素会产生相近的灰度水平值，因此有可能在满足图像清晰度要求的条件下，将存储一个m×n阶矩阵需要存储的m×n个数减少到n+m+1的一个倍数。
- 原矩阵$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$
- 压缩矩阵$A_k = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T, k \le r$
- 应用实例：https://yihui.shinyapps.io/imgsvd/

SVD与PCA

- 实际应用中，SVD分解是PCA的主要工具
- SVD分解的迭代计算比特征值分解更快更准确

- 若B是中心化后的p*n阶观测矩阵，$A=\frac{1}{\sqrt{N-1}}B^T$，A的SVD分解等价于B的协方差阵特征值分解

例子

- 在制定服装标准的过程中，对128名成年男子的身材进行了测量，每人测得的指标中含有这样六项：身高（x_1）、坐高（x_2）、胸围（x_3）、手臂长（x_4）、肋围（x_5）和腰围（x_6）。所得样本相关矩阵列于表7.3.1。

表7. 3. 1　　男子身材六项指标的样本相关矩阵

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.000					
x_2	0.79	1.000				
x_3	0.36	0.31	1.000			
x_4	0.76	0.55	0.35	1.000		
x_5	0.25	0.17	0.64	0.16	1.000	
x_6	0.51	0.35	0.58	0.38	0.63	1.000

例子

◆ 表7.3.2　的前三个特征值、特征向量以及贡献率

特征向量			
：身高	0.469	−0.365	0.092
：坐高	0.404	−0.397	0.613
：胸围	0.394	0.397	−0.279
：手臂长	0.408	−0.365	−0.705
：肋围	0.337	0.569	0.164
：腰围	0.427	0.308	0.119
特征值	3.287	1.406	0.459
贡献率	0.548	0.234	0.077
累计贡献率	0.548	0.782	0.859

例子

- 前3个主成分

$$\hat{y}_1 = 0.469x_1^* + 0.404x_2^* + 0.394x_3^* + 0.408x_4^* + 0.337x_5^* + 0.427x_6^*$$

$$\hat{y}_2 = -0.365x_1^* - 0.397x_2^* + 0.397x_3^* - 0.365x_4^* + 0.569x_5^* + 0.308x_6^*$$

$$\hat{y}_3 = 0.092x_1^* + 0.613x_2^* - 0.279x_3^* - 0.705x_4^* + 0.164x_5^* + 0.119x_6^*$$

R中的主成分计算

在某中学随机抽取某年级 30 名学生，测量其身高 (X_1)、体重 (X_2)、胸围 (X_3) 和坐高 (X_4)，数据如表 9.1 所示. 试对这 30 名中学生身体四项指标数据做主成分分析.

```
student.pr <- princomp(student, cor = TRUE)
summary(student.pr, loadings=TRUE)
predict(student.pr)
```

```
> student.pr <- princomp(student, cor = TRUE)
> summary(student.pr, loadings=TRUE)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4
Standard deviation     1.8817805 0.55980636 0.28179594 0.25711844
Proportion of Variance 0.8852745 0.07834579 0.01985224 0.01652747
Cumulative Proportion  0.8852745 0.96362029 0.98347253 1.00000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
X1 -0.497  0.543 -0.450  0.506
X2 -0.515 -0.210 -0.462 -0.691
X3 -0.481 -0.725  0.175  0.461
X4 -0.507  0.368  0.744 -0.232
> predict(student.pr)
```

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间