第8章 主成分分析及R使用

武慧
wuh@hit.edu.cn

经济管理学院
哈尔滨工业大学（威海）

1

高维数据

居民消费水平评价

变量：人均食品支出、人均衣着商品支出、人均医疗保健支出、人均交通支出、人均居住支出、人均娱乐服务支出等

基于车联网的驾驶行为分析

变量：行驶/疲劳驾驶/早晚高峰/深夜出行/高速驾驶时长、平均/最大速度、速度标准差、加速度、行驶里程等

2

高维数据的挑战

- **维度灾难**：随着特征数量的增加，数据的稀疏性增大，计算成本、存储需求和分析难度都呈指数级增长。
- **数据可视化**：高维数据难以在2D或3D空间中进行直观展示和分析。
- **过拟合风险**：特征过多时，模型可能会过度拟合训练数据，无法泛化到新数据。

经典降维方法

变量太多怎么办？

降维

用少数几个新的变量代替原有变量，合并重复信息，但不损失重要信息

Reduce the fat in your data

4

经典降维方法

- 主成分分析（Principal Component Analysis，PCA）：通过将原始变量转换为原始变量的线性组合（主成分），在保留主要信息的基础上，达到简化和降维的目的
- 因子分析（Factor Analysis，FA）：通过研究众多变量之间的内部依赖关系，探求观测数据的基本结构，并用少数几个假想变量（因子）来表示原始数据

5

本章内容

1. 主成分分析的思想
2. 总体主成分分析
3. 样本主成分分析的R实现

6

1. 主成分分析的思想

7

引例

【例】下表记录了52位学生6门功课的考试分数的部分数据，依次表示数学、物理、化学、语文、历史、英语成绩：

```
library(readxl)
test_score <-
read_excel("test_score.csv")
View(test_score)
```

	Y1	Y2	Y3	Y4	Y5	Y6
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	78	84	75	62	71	64
5	66	71	67	52	65	57
6	83	100	79	41	67	50
7	86	94	97	51	63	55
8	67	84	53	58	66	56
9	69	56	67	75	94	80
10	77	90	80	68	66	60
11	84	67	75	60	70	63
12	62	67	83	71	85	77
13	91	74	97	62	71	66
14	82	70	83	68	77	85
15	66	61	77	62	73	64

8

主成分分析的思想

希望建立一个（或几个）度量学生考试表现的综合指标，让我们能尽量明显地区分学生。那么选什么指标合适呢？

平均数/总数？

尝试通过给各科赋予不同的权重，从而更好地区分学生。

9

主成分分析的思想

- 主成分分析：在所有可能的特征变量的线性组合模式中寻找一个（或几个）可以最大程度区分样本的线性组合/加权平均。
- 与线性判别分析的比较：

➢线性判别分析是寻求最大化两个或多个群体之间距离的线性组合。

➢在主成分分析中只有一个群体，目标是找到一个能使这个群体中个体差异达到最大的变量线性组合。

单选题 1分

主成分分析（PCA）的主要目的是：

A 提高数据的噪声

B 进行聚类分析

C 降低数据集的维度

D 增加数据集的维度

11

单选题 1分

主成分分析（PCA）的基本思想是什么？

A 通过增加样本数量来优化模型

B 使用非线性方法提高数据的可解释性

C 对数据进行完全的分类和回归建模

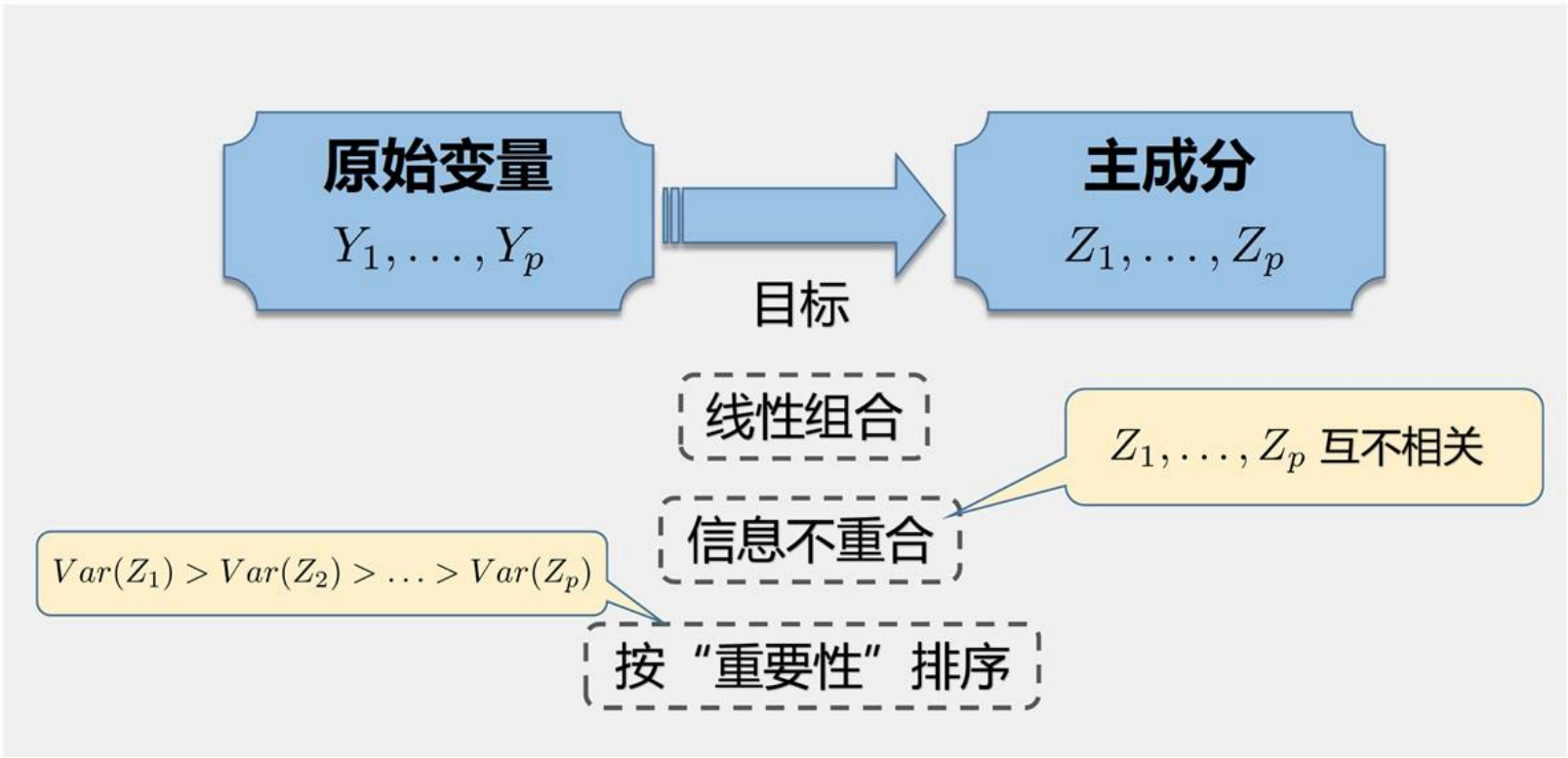
D 寻找一个或多个特征变量的线性组合来最大程度区分或代表样本

12

2. 总体主成分分析

13

总体主成分分析


原始变量
$Y_1, \ldots, Y_p$
目标
主成分
$Z_1, \ldots, Z_p$
线性组合
$Z_1, \ldots, Z_p$ 互不相关
信息不重合
$Var(Z_1) > Var(Z_2) > \ldots > Var(Z_p)$
按“重要性”排序

总体主成分分析

- 记原始变量$\boldsymbol{y}=\left(Y_1,Y_2,\cdots,Y_p\right)'$，其协方差矩阵记为$\boldsymbol{\Sigma}$。
- 主成分分析试图定义一组互不相关的变量（主成分），记为$Z_1,Z_2,\cdots,Z_p$，每一个主成分都是原始变量的线性组合：

$$\begin{aligned}
Z_1 &= \mathbf{a}_1'\mathbf{y} = a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1p}Y_p \\
Z_2 &= \mathbf{a}_2'\mathbf{y} = a_{21}Y_1 + a_{22}Y_2 + \ldots + a_{2p}Y_p \\
&\vdots \\
Z_p &= \mathbf{a}_p'\mathbf{y} = a_{p1}Y_1 + a_{p2}Y_2 + \ldots + a_{pp}Y_p
\end{aligned}$$

- 主成分的方差和协方差：

$$var(Z_j) = \mathbf{a}_j'\boldsymbol{\Sigma}\mathbf{a}_j,\ cov(Z_j,Z_k) = \mathbf{a}_j'\boldsymbol{\Sigma}\mathbf{a}_k,\ j,k = 1,\ldots,p.$$

15

总体主成分分析

- 主成分按照“方差贡献度”依次导出：

第一主成分 $Z_1 = \mathbf{a}_1'\mathbf{y}$ ：在满足限制 $\mathbf{a}_1'\mathbf{a}_1 = 1$ 时，最大化方差 $var(\mathbf{a}_1'\mathbf{y})$

第二主成分 $Z_2 = \mathbf{a}_2'\mathbf{y}$ ：在满足限制 $\mathbf{a}_2'\mathbf{a}_2 = 1$ ，且 $cov(\mathbf{a}_1'\mathbf{y}, \mathbf{a}_2'\mathbf{y}) = 0$ 时，最大化方差 $var(\mathbf{a}_2'\mathbf{y})$

.....

第 j 主成分 $Z_j = \mathbf{a}_j'\mathbf{y}$ ：在满足限制 $\mathbf{a}_j'\mathbf{a}_j = 1$ ，且 $cov(\mathbf{a}_k'\mathbf{y}, \mathbf{a}_j'\mathbf{y}) = 0,\ k < j$ 时，最大化方差 $var(\mathbf{a}_j'\mathbf{y})$

第 p主成分 $Z_j = \mathbf{a}_j'\mathbf{y}$ ：在满足限制 $\mathbf{a}_p'\mathbf{a}_p = 1$时，最小化方差 $var(\mathbf{a}_p'\mathbf{y})$

通过这种方式，主成分分析能够依次提取出能够解释数据中最大方差的方向，并确保每个主成分之间是正交的，从而实现数据的降维和信息的提取。

16

定理：

记 $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$ 为协方差矩阵 $\mathbf{\Sigma}$ 的特征值-特征向量，$\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ 并且特征向量 $\mathbf{e}_1, \ldots, \mathbf{e}_p$ 是正交化特征向量。

则变量 $Y_1, \ldots, Y_p$ 的第 j 个主成分由下式给出：

$$Z_j = \mathbf{e}_j' \mathbf{y} = e_{j1} Y_1 + e_{j2} Y_2 + \ldots + e_{jp} Y_p, j = 1, \ldots, p,$$

这里有 $var(Z_j) = \mathbf{e}_j' \mathbf{\Sigma} \mathbf{e}_j = \lambda_j$

并且有 $cov(Z_j, Z_k) = \mathbf{e}_j' \mathbf{\Sigma} \mathbf{e}_k = 0$

进一步地，我们有：

$$\sum_{j=1}^{p} var(Z_j) = \sum_{j=1}^{p} var(Y_j)$$

主成分的个数

- 第k个主成分贡献的方差，占总体方差的比例可表示如下：

$$\frac{\lambda_k}{\sum_{j=1}^{p}\lambda_j},\ k=1,\ldots,p.$$

- 如果前面几个主成分可贡献总体的大部分方差/信息（如85%），那么这些主成分能够以较少的信息损失来代替原始变量。

Reduce the fat
in your data

单选题 1分

在主成分分析（PCA）中，第一主成分的提取目标是：

A 最小化方差

B 找到一个线性组合，使得方差最大化

C 确保主成分之间的相关性

D 提取原始数据的所有特征

19

单选题 1分

提取主成分的过程可以实现什么目的?

A 增加数据的维度

B 完全去除数据中的噪声

C 降低数据的维度并保留重要信息

D 保留所有原始变量的信息

20

单选题 1分

在PCA中，通过协方差矩阵的特征值的大小可以判断：

A 哪个主成分对数据的方差贡献最大

B 哪个特征是最重要的

C 哪个特征是最具代表性的

D 哪个特征的数值最小

21

3. 样本主成分分析的R实现

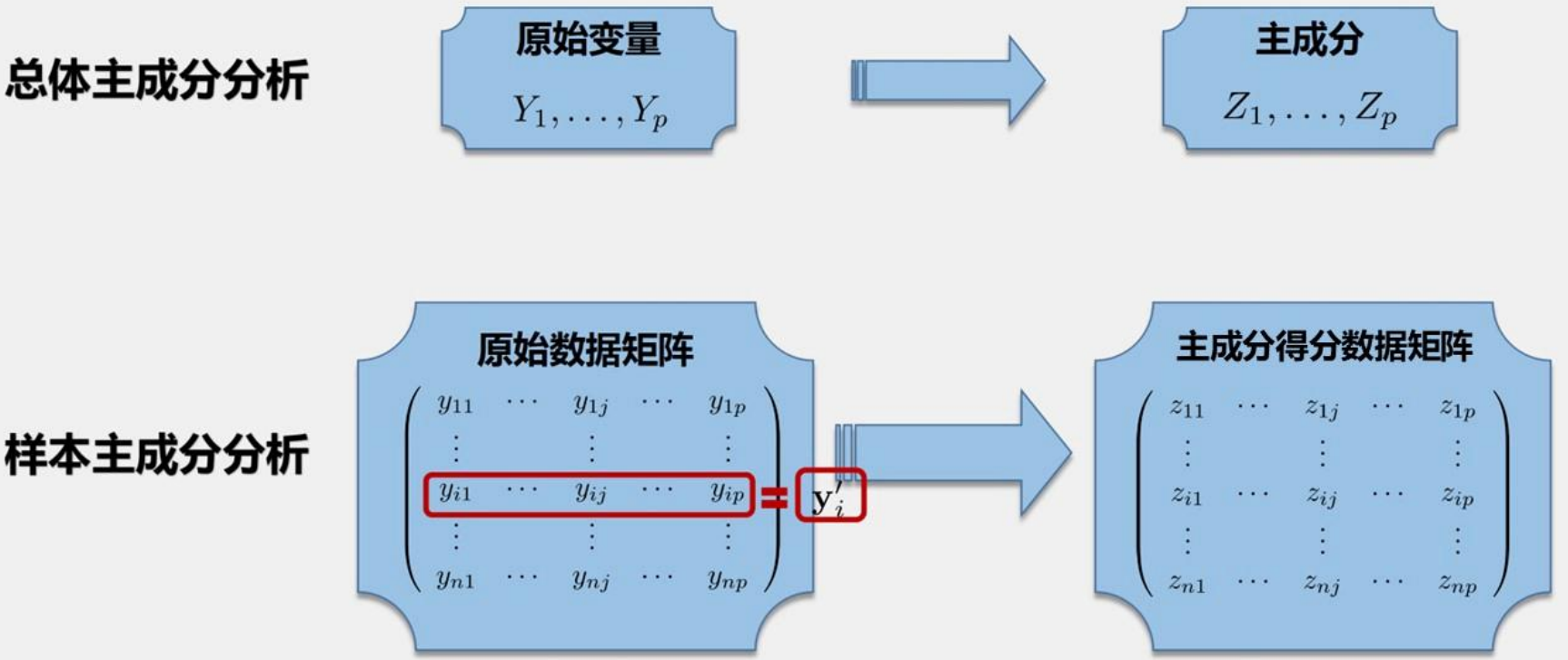
22

样本主成分分析案例

- 回到考试分数的案例，与其说用简单的平均数，我们更希望用一个更好的原始变量的线性组合来区分学生。
- 这就是样本层面的主成分分析。
- 样本主成分分析的目的是找到样本数据中最大的方差方向，并通过这些方向来描述数据的主要变化模式。得到的主成分表示数据中最重要的变化轴，主成分得分则表示每个样本在这些轴上的位置。

23

样本主成分分析


总体主成分分析
原始变量
$Y_1, \ldots, Y_p$
主成分
$Z_1, \ldots, Z_p$
样本主成分分析
原始数据矩阵
$\begin{pmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1p} \\ \vdots & & \vdots & & \vdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & & \vdots & & \vdots \\ y_{n1} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix} = \mathbf{y}_i'$
主成分得分数据矩阵
$\begin{pmatrix} z_{11} & \cdots & z_{1j} & \cdots & z_{1p} \\ \vdots & & \vdots & & \vdots \\ z_{i1} & \cdots & z_{ij} & \cdots & z_{ip} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \cdots & z_{nj} & \cdots & z_{np} \end{pmatrix}$

样本主成分分析的R实现

```
# 执行PCA
pca_result <- princomp(x = test_score,
     cor = TRUE)
# 查看PCA结果的概览
summary(pca_result)
```

- 参数说明：

➢x：一个数据框或矩阵，包含需要进行 PCA 的数据。

➢cor：布尔值，表示是否使用相关矩阵。如果为 TRUE，则使用相关矩阵（相当于数据标准化后的协方差矩阵）；如果为 FALSE（默认），则使用协方差矩阵。

25

样本主成分分析的R实现

```
> # 查看PCA结果的概览
> summary(pca_result)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.9261112 1.1236019 0.66395522 0.52009785 0.41172308 0.38309295
Proportion of Variance 0.6183174 0.2104135 0.07347275 0.04508363 0.02825265 0.02446003
Cumulative Proportion  0.6183174 0.8287309 0.90220369 0.94728732 0.97553997 1.00000000
```

- Standard deviation: 每个主成分的标准差。
- Proportion of Variance: 每个主成分所解释的方差比例。
- Cumulative Proportion: 累计解释的方差比例。

样本主成分分析的R实现

```
# 查看每个主成分的载荷（loadings）
pca_result$loadings
```

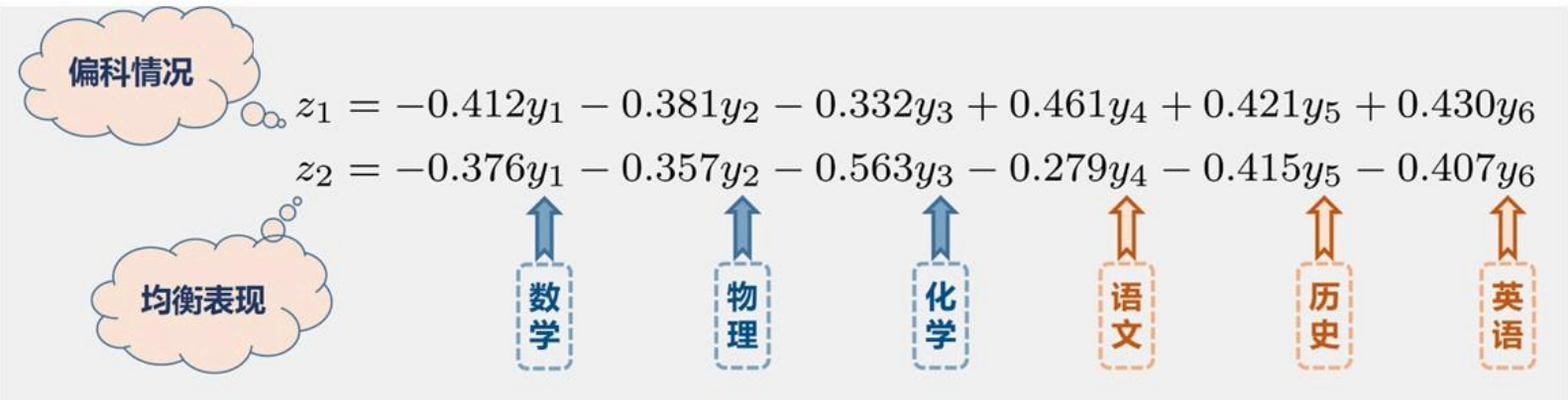
```
> pca_result$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Y1  0.412  0.376  0.216  0.788         0.145
Y2  0.381  0.357 -0.806 -0.118 -0.212 -0.141
Y3  0.332  0.563  0.467 -0.588
Y4 -0.461  0.279               -0.599  0.590
Y5 -0.421  0.415 -0.250         0.738  0.205
Y6 -0.430  0.407  0.146  0.134 -0.222 -0.749
```

- 主成分的载荷（loadings）是指原始变量在每个主成分上的权重或贡献程度，表示每个主成分是由哪些原始变量构成的，以及这些变量对主成分的影响大小。

样本主成分分析的R实现

- 主成分分析的目的是通过将数据从高维空间映射到一个低维空间来减少维度，同时尽可能保留数据的方差。这个低维空间中的每个维度就是主成分。
- 每个主成分实际上是原始变量的一个加权组合（线性组合），这些权重就是主成分的载荷。



样本主成分分析的R实现

- 在(Y_1, Y_2, Y_3)和(Y_4, Y_5, Y_6)两组中，z_1包含相反的符号和相似的量级，反映了文理科的偏科现象。典型学生为6, 7, 45, 30, 49号：

```
> test_score[c(6, 7, 45, 30, 49),]
   Y1  Y2 Y3  Y4 Y5  Y6
6  83 100 79  41 67  50
7  86  94 97  51 63  55
45 99 100 99  53 63  60
30 64  61 49 100 99  95
49 52  62 65 100 96 100
```

- 在z_2中，所有变量的符号相同，反映了文理科的均衡表现。典型学生为26, 33, 8号：

```
> test_score[c(26,33,8),]
   Y1 Y2  Y3 Y4 Y5 Y6
26 87 84 100 74 81 76
33 86 78  92 87 87 77
8  67 84  53 58 66 56
```

样本主成分分析的R实现

- 提取这几位同学对应的主成分得分（数据点在每个主成分上的投影。），进一步分析：

```
> samplePC<-(round(test_PCA$scores,3))[c(6,7,45,30,49,26,33,8),]
> rownames(samplePC)<-c(6,7,45,30,49,26,33,8)
> samplePC
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
6  -3.518  0.820 -1.072 -0.156 -0.763 -0.166
7  -3.516 -0.104  0.101 -0.574 -0.011  0.080
45 -3.975 -1.054  0.147  0.349  0.252 -0.049
30  4.490 -0.693 -0.620  0.832 -0.054  0.029
49  4.622 -0.997 -0.236 -0.724  0.289 -0.465
26 -0.841 -2.117  0.544 -0.156 -0.070  0.192
33  0.345 -2.187  0.414  0.225  0.018  0.854
8  -0.982  2.326 -1.292 -0.017  0.093 -0.052
```

理科 >文科

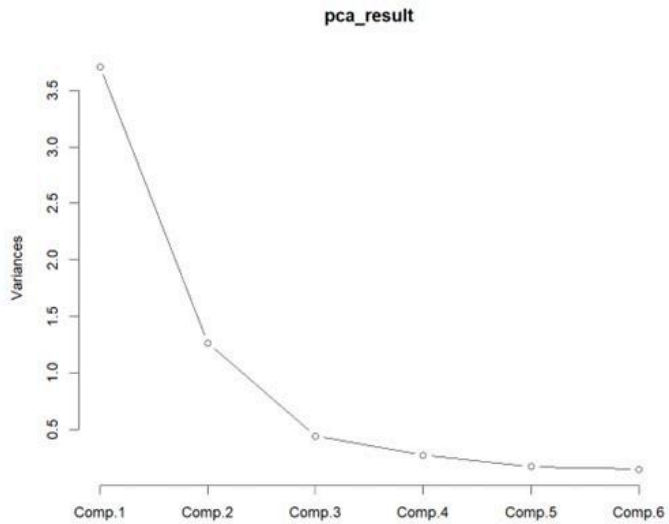
文科 >理科

总成绩高

总成绩低

样本主成分分析的R实现

```
# 绘制Scree Plot
screeplot(pca_result,
     type = "lines")
```




```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.9261112 1.1236019 0.66395522 0.52009785 0.41172308 0.38309295
Proportion of Variance 0.6183174 0.2104135 0.07347275 0.04508363 0.02825265 0.02446003
Cumulative Proportion  0.6183174 0.8287309 0.90220369 0.94728732 0.97553997 1.00000000
```

样本主成分分析的R实现

```
# 提取前三个主成分
pca_reduced <- pca_result$scores[, 1:3]
# 查看降维后的数据
head(pca_reduced)
```

```
> # 查看降维后的数据
> head(pca_reduced)
          Comp.1      Comp.2       Comp.3
[1,] -1.8458505 -0.09869937  0.501399041
[2,]  1.3829375 -0.93264526 -0.027536570
[3,] -0.0444749 -2.80445313 -0.220442506
[4,]  1.2555721 -0.40584893 -0.350765626
[5,]  1.1390256 -2.26874265 -0.004069094
[6,]  3.5175343 -0.81974747 -1.071937337
```

32

本章小结

- **主成分分析的思想**：寻找一个（或几个）可以最大程度区分（或代表）样本的特征变量线性组合。
- **总体主成分分析**：主成分的提取过程是通过最大化方差并确保主成分之间的正交性来实现的。通过对总体的协方差矩阵进行特征值分解，找出主成分方向，使得这些方向尽可能多地保留总体数据的方差。
- **样本主成分分析的R实现**：通过对样本数据的协方差矩阵（或相关矩阵）进行特征值分解，估计出样本数据的主成分方向。这里，协方差矩阵是基于样本数据估计的，而不是总体的真实协方差矩阵。（princomp）

33