

Principal Components Analysis

主成分分析

肖磊，2026年5月7日

已学知识点 (Recap)

第 10 章 数据矩阵的因子分解

$$\mathcal{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = (\mathbf{x}_{[1]} \quad \mathbf{x}_{[2]} \quad \cdots \quad \mathbf{x}_{[p]})$$

10.1 几何的角度

- ▶ \mathcal{X} 的每一行 (观测点) \mathbf{x}_i^T 是 \mathbb{R}^p 中的一个向量. 所以 \mathcal{X} 是 \mathbb{R}^p 中的 n 个点.
- ▶ \mathcal{X} 的每一列 (变量) $\mathbf{x}_{[j]}$ 是 \mathbb{R}^n 中的一个向量. 所以 \mathcal{X} 是 \mathbb{R}^n 中的 p 个点.

已学知识点 (Recap)

第 10 章 数据矩阵的因子分解

10.2 拟合 p 维数据点

- ▶ 通过将每一个 p 维观测点投影到更低维度的空间上以图形方式表示.

$$\mathcal{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} = (\mathbf{x}_{[1]} \quad \mathbf{x}_{[2]} \quad \cdots \quad \mathbf{x}_{[p]})$$

- ▶ 第一个主因子轴 \mathbf{u}_1 确定了一条过原点的直线 F_1 . 这条直线通过使正交距离 $\sum_{i=1}^n \|\mathbf{x}_i - \mathbf{p}_{x_i}\|^2$

最小化而得到. 因子 \mathbf{u}_1 由矩阵 $\mathcal{X}^T \mathcal{X}$ 的最大特征值对应的特征向量确定. p 维观测点在直线上的坐标由 $z_1 = \mathcal{X} \mathbf{u}_1$ 给出.

- ▶ 第二个主因子轴 \mathbf{u}_2 由矩阵 $\mathcal{X}^T \mathcal{X}$ 的第二大特征值对应的特征向量确定. p 维观测点在平面上的坐标由 $z_1 = \mathcal{X} \mathbf{u}_1$ 与 $z_2 = \mathcal{X} \mathbf{u}_2$ 给出.

- ▶ 前 q 个主因子轴 $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_q$ 由矩阵 $\mathcal{X}^T \mathcal{X}$ 的前 q 个大特征值对应的特征向量确定. p 维观测点在 q 维子空间上的坐标由 $z_1 = \mathcal{X} \mathbf{u}_1, z_2 = \mathcal{X} \mathbf{u}_2, \dots, z_q = \mathcal{X} \mathbf{u}_q$ 给出.

已学知识点 (Recap)

第 10 章 数据矩阵的因子分解

10.3 拟合 n 维变量点

$$\mathcal{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} \\ = (\mathbf{x}_{[1]} \quad \mathbf{x}_{[2]} \quad \cdots \quad \mathbf{x}_{[p]})$$

- ▶ 通过将每一个 n 维变量点投影到更低维度的空间上以图形方式表示.

- ▶ 第一个主因子轴 \mathbf{v}_1 确定了一条过原点的直线 G_1 . 该直线通过使正交距离 $\sum_{j=1}^p \|\mathbf{x}_{[j]} - \mathbf{p}_{\mathbf{x}_{[j]}}\|^2$

最小化而得到. 因子 \mathbf{v}_1 由矩阵 $\mathcal{X}\mathcal{X}^T$ 的最大特征值对应的特征向量确定. n 维变量点在直线上的坐标由 $\mathbf{w}_1 = \mathcal{X}^T \mathbf{v}_1$ 给出.

- ▶ 第二个主因子轴 \mathbf{v}_2 由矩阵 $\mathcal{X}\mathcal{X}^T$ 的第二大特征值对应的特征向量确定. n 维变量点在平面上的坐标由 $\mathbf{w}_1 = \mathcal{X}^T \mathbf{v}_1$ 与 $\mathbf{w}_2 = \mathcal{X}^T \mathbf{v}_2$ 给出.

- ▶ 前 q 个主因子轴 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q$ 的方向由矩阵 $\mathcal{X}\mathcal{X}^T$ 的前 q 个大特征值对应的特征向量确定. n 维变量点在 q 维子空间上的坐标由 $\mathbf{w}_1 = \mathcal{X}^T \mathbf{v}_1, \mathbf{w}_2 = \mathcal{X}^T \mathbf{v}_2, \dots, \mathbf{w}_q = \mathcal{X}^T \mathbf{v}_q$ 给出.

已学知识点 (Recap)

第 10 章 数据矩阵的因子分解

10.4 两个子空间的关系

- ▶ 矩阵 $\mathcal{X}^T \mathcal{X}$ 与 $\mathcal{X} \mathcal{X}^T$ 有相同的非零特征值 $\lambda_1, \lambda_2, \dots, \lambda_r$, 其中 $r = \text{rank}(\mathcal{X})$.

$$\mathcal{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{[1]} & \mathbf{x}_{[2]} & \cdots & \mathbf{x}_{[p]} \end{pmatrix}$$

- ▶ 矩阵 $\mathcal{X}^T \mathcal{X}$ 的特征向量可以由矩阵 $\mathcal{X} \mathcal{X}^T$ 的特征向量计算得出, 反之亦然:

$$\mathbf{u}_k = \frac{1}{\sqrt{\lambda_k}} \mathcal{X}^T \mathbf{v}_k, \quad \mathbf{v}_k = \frac{1}{\sqrt{\lambda_k}} \mathcal{X} \mathbf{u}_k.$$

- ▶ \mathcal{X} 的变量 (列) 点在 q 维子空间中的坐标可以很容易地通过 $\mathbf{w}_k = \sqrt{\lambda_k} \mathbf{u}_k$ 计算得到.

已学知识点 (Recap)

第 10 章 数据矩阵的因子分解

10.5 实际计算

- ▶ 具体实施数据矩阵的因子分解时, 要计算矩阵 $\mathcal{X}^T \mathcal{X}$ 的特征值 $\lambda_1, \lambda_2, \dots, \lambda_p$ 以及对应的特征向量 $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$.

$$\begin{aligned}
 \mathcal{X}_{n \times p} &= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} \\
 &= (\mathbf{x}_{[1]} \quad \mathbf{x}_{[2]} \quad \cdots \quad \mathbf{x}_{[p]})
 \end{aligned}$$

通过绘制 $z_1 = \mathcal{X} \mathbf{u}_1$ 与 $z_2 = \mathcal{X} \mathbf{u}_2$ 的关系图 (如果需要的话, 还可以绘制与 $z_3 = \mathcal{X} \mathbf{u}_3$ 的关系图), 即可得到 n 个观测点的可视化表示. 通过绘制 $w_1 = \sqrt{\lambda_1} \mathbf{u}_1$ 与 $w_2 = \sqrt{\lambda_2} \mathbf{u}_2$ 的关系图 (如果需要的话, 还可以绘制与 $w_3 = \sqrt{\lambda_3} \mathbf{u}_3$ 的关系图), 即可得到 p 个变量点的可视化表示.

- ▶ 因子表示于 q 维子空间的质量可以用前 q 个因子的惯性百分比 τ_q 来评价, 其中

$$\tau_q = \frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_q + \lambda_{q+1} + \cdots + \lambda_p}.$$

已学知识点 (Recap)

第 11 章 主成分分析

- 标准线性组合

▶ 标准线性组合: $\delta^T X = \sum_{j=1}^p \delta_j X_j$, 其中 $\delta^T \delta = 1$.

▶ 使 $\text{Var}(\delta^T X)$ 最大的 $\delta = \gamma_1$, 其中 γ_1 是 $\Sigma = \text{Var}(X)$ 的最大特征值 λ_1 对应的单位特征向量. 称 $Y_1 = \gamma_1^T (X - \mu)$ 为第一主成分.

▶ 上述投影可以推广到高维. 主成分变换是线性变换 $Y = \Gamma^T (X - \mu)$, 其中 $\Sigma = \text{Var}(X) = \Gamma \Lambda \Gamma^T$, $\mu = \mathbb{E}(X)$. 称 Y_1, Y_2, \dots, Y_p 分别为第一主成分、第二主成分、...、第 p 主成分.

▶ 各主成分的均值为零, $\text{Var}(Y_j) = \lambda_j$, 且不相干. 对于 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 而言有

$\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \dots \geq \text{Var}(Y_p)$. 并且 $\sum_{j=1}^p \text{Var}(Y_j) = \text{tr}(\Sigma)$, $\prod_{j=1}^p \text{Var}(Y_j) = |\Sigma|$.

▶ 如果 $Y = a^T X$ 是与 X 的前 k 个主成分对应的标准线性组合, 则使 $Y = a^T X$ 的方差达到最大者必是 X 的第 $k+1$ 个主成分.

已学知识点 (Recap)

第 11 章 主成分分析

- 主成分的估计

- ▶ 主成分变换中各变量的尺度应大致相同.
- ▶ 实际进行主成分分析时, 我们用 $\bar{\mathbf{x}}$ 代替 $\boldsymbol{\mu}$, 用 \mathcal{S} 代替 Σ . 然后计算 \mathcal{S} 的特征值 $\ell_1, \ell_2, \dots, \ell_p$ 及其对应的单位特征向量 $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_p$. 通过作第一主成分关于第二主成分 (甚至第三主成分) 的图形即可获得主成分的图形表示.
- ▶ 特征向量 \mathbf{g}_i 的分量即是主成分在各初始变量上的权重.

已学知识点 (Recap)

第 11 章 主成分分析

● 主成分的解释

- ▶ 主成分的权重 (系数) 表明, 在初始坐标所表示的哪些方向上, 能实现对方差的最佳解释. 需要注意, 主成分不具备尺度不变性.

- ▶ 前 q 个主成分解释初始变量当中变化的能力为 $\psi_q = \frac{\lambda_1 + \dots + \lambda_q}{\lambda_1 + \dots + \lambda_p}$. 可以通过碎石图对其进行可视化.

- ▶ 主成分 Y_j 与初始变量 X_i 的相关系数为 $\rho_{X_i Y_j} = \gamma_{ij} \sqrt{\frac{\lambda_j}{\sigma_{X_i X_i}}}$. 对于数据矩阵则有 $r_{X_i Y_j}^2 = \frac{\ell_j g_{ij}^2}{s_{X_i X_i}}$. 这里的 $r_{X_i Y_j}^2$

可理解为 X_i 的方差可由 Y_j 解释的比例. $r_{X_i Y_1}$ 关于 $r_{X_i Y_2}$ 的散点图可用来表明哪些初始变量, 即那些靠近半径为 1 的圆周边缘的变量, 与主成分 (PC) 的相关性最强.

$$\sum_{j=1}^p r_{X_i Y_j}^2 = 1$$

主成分的解释 (Interpretation of the PCs)

- 例: 银行钞票数据集的主成分.

```
rm(list = ls(all = TRUE))  
graphics.off()
```

```
# 载入数据
```

```
library(mclust)  
data(banknote)  
x = banknote[, 2:7]  
n = nrow(x)
```

```
# 谱分解
```

```
e = eigen(cov(x) * (n-1) / n) # 协方差矩阵的谱分解
```

```
round(e$values, digits = 3) # 特征值
```

```
e1 = e$values / sum(e$values) # 各特征值的占比
```

```
round(e1, digits = 3)
```

```
round(e$vectors, digits = 3) # 特征向量
```

```
> round(e$values, digits = 3) # 特征值  
[1] 2.985 0.931 0.242 0.194 0.085 0.035
```

```
> round(e1, digits = 3)  
[1] 0.668 0.208 0.054 0.043 0.019 0.008
```

```
> round(e$vectors, digits = 3) # 特征向量  
      [,1] [,2] [,3] [,4] [,5] [,6]  
[1,] -0.044 0.011 -0.326 0.562 0.753 0.098  
[2,] 0.112 0.071 -0.259 0.455 -0.347 -0.767  
[3,] 0.139 0.066 -0.345 0.415 -0.535 0.632  
[4,] 0.768 -0.563 -0.218 -0.186 0.100 -0.022  
[5,] 0.202 0.659 -0.557 -0.451 0.102 -0.035  
[6,] -0.579 -0.489 -0.592 -0.258 -0.084 -0.046
```

主成分的解释 (Interpretation of the PCs)

- 例: 银行钞票数据集的主成分.

- ▶ 对经过中心化处理的银行数据 x , 我们有

$$y_1 = -0.044x_1 + 0.112x_2 + 0.139x_3 + 0.768x_4 + 0.202x_5 - 0.579x_6$$

→ 第一主成分主要反映底框 (x_4) 与对角线 (x_6) 之间的差异.

→ 第二主成分主要可以通过顶框 (x_5) 与底框 (x_4) 和对角线 (x_6) 之和的差值来描述.

$$y_2 = 0.011x_1 + 0.071x_2 + 0.066x_3 - 0.563x_4 + 0.659x_5 - 0.489x_6$$

$x_1 = \text{length}$
 $x_2 = \text{left height}$
 $x_3 = \text{right height}$
 $x_4 = \text{bottom frame}$
 $x_5 = \text{top frame}$
 $x_6 = \text{diagonal}$

主成分的权重告诉我们, 在以原始坐标表示的哪些方向上, 能够获得最佳的方差解释.

```

> round(e$vector, digits = 3) # 特征向量
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,] -0.044 0.011 -0.326 0.562 0.753 0.098
[2,] 0.112 0.071 -0.259 0.455 -0.347 -0.767
[3,] 0.139 0.066 -0.345 0.415 -0.535 0.632
[4,] 0.768 -0.563 -0.218 -0.186 0.100 -0.022
[5,] 0.202 0.659 -0.557 -0.451 0.102 -0.035
[6,] -0.579 -0.489 -0.592 -0.258 -0.084 -0.046
    
```

主成分的解释 (Interpretation of the PCs)

- 例: 银行钞票数据集的主成分.

- ▶ 对经过中心化处理的银行数据 x ,

前 q 个主成分对变异的解释程度可以用相对比例来衡量:

$$\Psi_q = \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} = \frac{\sum_{j=1}^q \text{Var}(Y_j)}{\sum_{j=1}^p \text{Var}(Y_j)}$$

- $x_1 = \text{length}$
- $x_2 = \text{left height}$
- $x_3 = \text{right height}$
- $x_4 = \text{bottom frame}$
- $x_5 = \text{top frame}$
- $x_6 = \text{diagonal}$

- ▶ 第一主成分 ($q = 1$) 解释了 66.8% 的变异.
- ▶ 前两个主成分 ($q = 2$) 解释了 87.6% 的变异.
- ▶ 前三个主成分 ($q = 3$) 解释了 93.0% 的变异.

```

e1 = e$values / sum(e$values) # 各特征值的占比
round(e1, digits = 3)
e2 = cumsum(e$values) / sum(e$values) # 各特征值的累积占比
round(e2, digits = 3)
  
```

```

> round(e1, digits = 3)
[1] 0.668 0.208 0.054 0.043 0.019 0.008
  
```

```

> round(e2, digits = 3)
[1] 0.668 0.876 0.930 0.973 0.992 1.000
  
```

- ▶ **注意:** 主成分不具有尺度不变性, 从相关矩阵导出的主成分与从协方差矩阵导出的主成分会给出不同的结果.

主成分的解释 (Interpretation of the PCs)

- **例: 银行钞票数据集的主成分.**
 - ▶ 对经过中心化处理的银行数据 x , 降维后进行可视化.

```

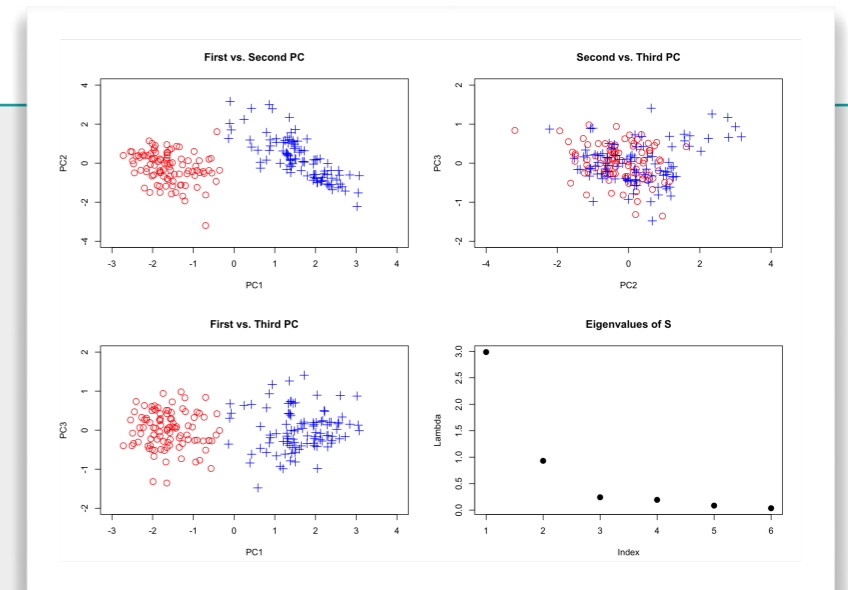
m = apply(as.matrix(x), 2, mean) # 样本均值向量
m
temp = as.matrix(x - matrix(m, n, ncol(x), byrow = T)) # 中心化数据
head(temp)
r0 = temp %*% e$vector # 主成分得分值
round(head(r0), digits = 3)

# 观测数据在第一、第二、第三主成分得分值的散点图, 碎石图
graphics.off()
par(mfrow = c(2, 2))
plot(r0[1:100, 1], r0[1:100, 2], xlim = c(-3, 4), ylim = c(-4, 4), xlab = 'PC1', ylab = 'PC2', col = 'red', cex = 1.5)
points(r0[101:200, 1], r0[101:200, 2], col = 'blue', cex = 1.5, pch = 3)
title(main = 'First vs. Second PC')

plot(r0[1:100, 2], r0[1:100, 3], xlim = c(-4, 4), ylim = c(-2, 2), xlab = 'PC2', ylab = 'PC3', col = 'red', cex = 1.5)
points(r0[101:200, 2], r0[101:200, 3], col = 'blue', cex = 1.5, pch = 3)
title(main = 'Second vs. Third PC')

plot(r0[1:100, 1], r0[1:100, 3], xlim = c(-3, 4), ylim = c(-2, 2), xlab = 'PC1', ylab = 'PC3', col = 'red', cex = 1.5)
points(r0[101:200, 1], r0[101:200, 3], col = 'blue', cex = 1.5, pch = 3)
title(main = 'First vs. Third PC')

plot(1:6, e$values, xlab = 'Index', ylab = 'Lambda', main = 'Eigenvalues of S', cex = 1.5, pch = 16)
    
```



主成分的解释 (Interpretation of the PCs)

- **例: 银行钞票数据集的主成分.**
 - ▶ 对经过中心化处理的银行数据 x , 变量点在前两个主成分上的可视化.

主成分与初始变量间的相关系数

```
r = cor(cbind(r0, x)) # 主成分与初始变量间的相关系数
round(r, digits = 3)
```

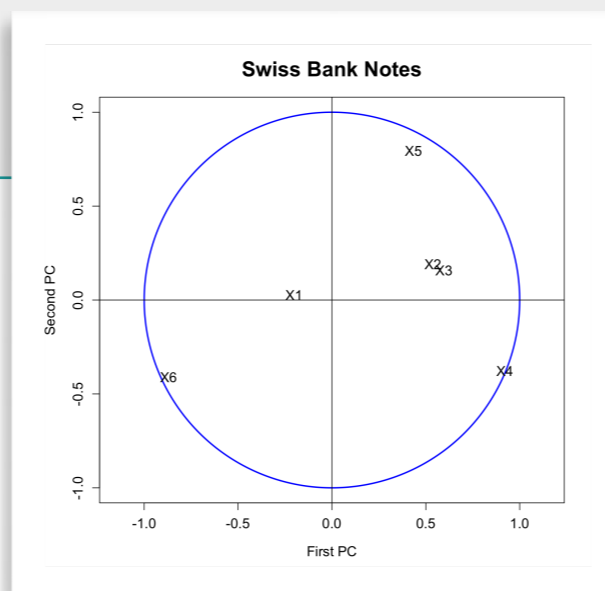
```
r1 = r[7:12, 1:2] # 初始变量与前两个主成分的相关系数
```

r1

初始变量与前两个主成分的相关系数的散点图

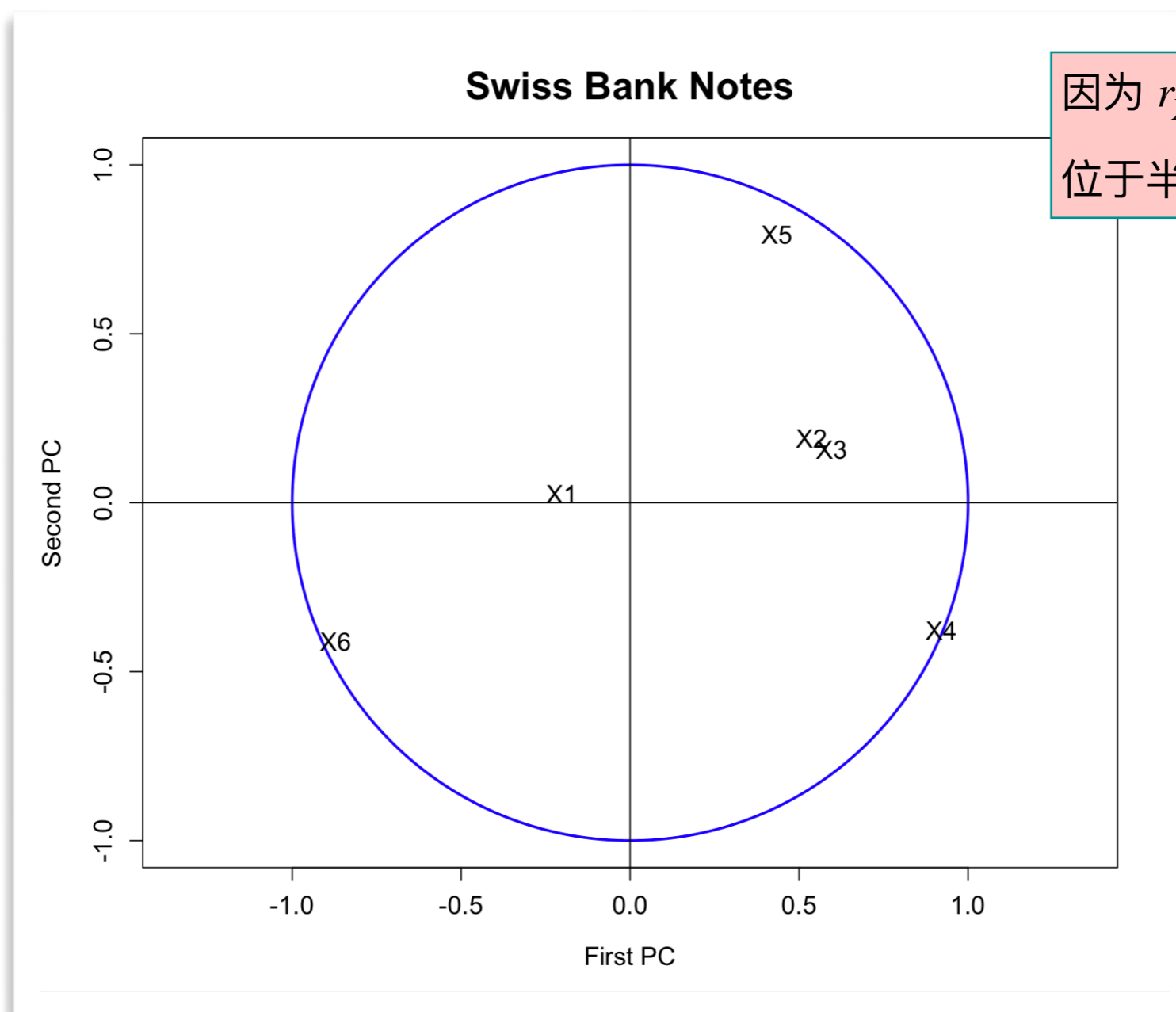
```
graphics.off()
ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab = "Second PC", main = "Swiss Bank Notes", cex.lab = 1.2, cex.axis = 1.2,
     cex.main = 1.8, lwd = 2, asp = 1)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6")
text(r1, label, cex = 1.2)
```

```
> round(r1, digits = 3)
      1      2
Length -0.201  0.028
Left   0.538  0.191
Right  0.597  0.159
Bottom 0.921 -0.377
Top    0.435  0.794
Diagonal -0.870 -0.410
```



主成分的解释 (Interpretation of the PCs)

- 例: 银行钞票数据集的主成分.



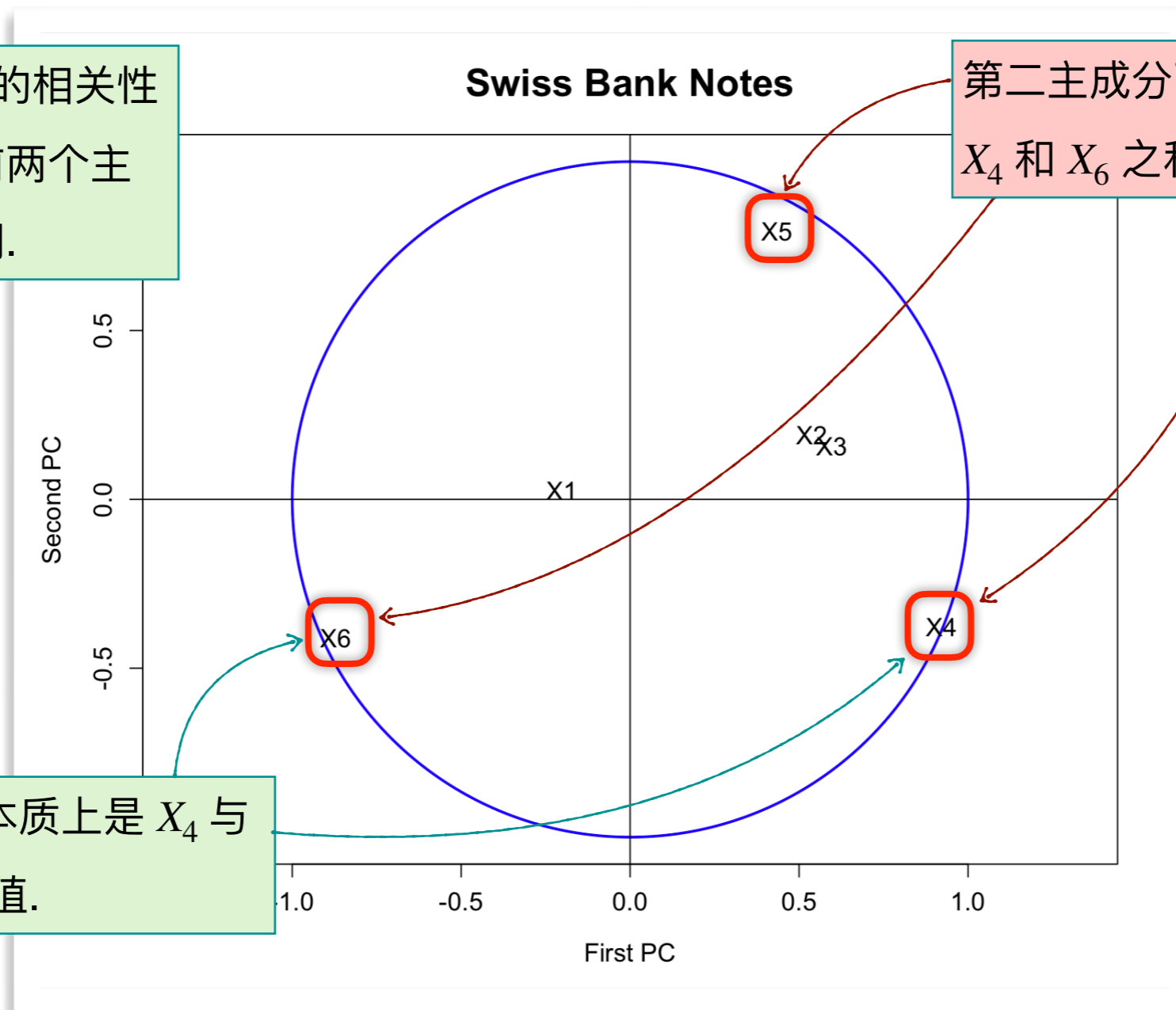
因为 $r_{X_i Y_1}^2 + r_{X_i Y_2}^2 \leq 1$, 散点总位于半径为 1 的单位圆内.

- ▶ 图形展示了哪些原始变量与主成分 Y_1 和 Y_2 的相关性最强.

主成分的解释 (Interpretation of the PCs)

- 例: 银行钞票数据集的主成分.

变量 X_4 , X_5 和 X_6 对应的相关性接近圆的边缘, 说明前两个主成分能很好地解释它们.



第二主成分可以很好地用 X_5 与 X_4 和 X_6 之和的差值来描述.

第一主成分本质上是 X_4 与 X_6 之间的差值.

- ▶ 图形展示了哪些原始变量与主成分 Y_1 和 Y_2 的相关性最强.

主成分的解释 (Interpretation of the PCs)

- 例: 银行钞票数据集的主成分.

- ▶ 初始变量 X_i 与前两个主成分的相关系数

```

# 前两个主成分对各个初始变量的解释比例
srs = array(0, dim = 6)
for (i in 1:6) srs[i] = r1[i, 1]^2 + r1[i, 2]^2
round(cbind(r1, ss = srs), digits = 3)
  
```

	$r_{X_i Y_1}$	$r_{X_i Y_2}$	$r_{X_i Y_1}^2 + r_{X_i Y_2}^2$
Length	-0.201	0.028	0.041
Left	0.538	0.191	0.326
Right	0.597	0.159	0.381
Bottom	0.921	-0.377	0.991
Top	0.435	0.794	0.820
Diagonal	-0.870	-0.410	0.925

前两个主成分解释 X_1 (以及 X_2, X_3) 的方差的比例相对较小.

主成分的渐进性质 (Asymptotic Properties of the PCs)

- 在实际应用中，主成分由样本数据求得. 下述定理给出了样本主成分的渐进分布.

定理 11.4 设 $\Sigma > 0$ 且有不同的特征值，再设 $\mathcal{S} \sim n^{-1}W_p(\Sigma, n-1)$ ，且有谱分解

$$\Sigma = \Gamma\Lambda\Gamma^T \text{ 与 } \mathcal{S} = \mathcal{G}\mathcal{L}\mathcal{G}^T. \text{ 则}$$

$$(a) \sqrt{n-1}(\boldsymbol{\ell} - \boldsymbol{\lambda}) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, 2\Lambda^2), \text{ 其中 } \boldsymbol{\ell} = (\ell_1, \ell_2, \dots, \ell_p)^T \text{ 与}$$

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p)^T \text{ 是 } \mathcal{L} \text{ 与 } \Lambda \text{ 的对角元素.}$$

$$(b) \sqrt{n-1}(\mathbf{g}_j - \boldsymbol{\gamma}_j) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \mathcal{V}_j), \text{ 其中 } \mathcal{V}_j = \lambda_j \sum_{k \neq j} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T.$$

(c) $\boldsymbol{\ell}$ 中的元素与矩阵 \mathcal{G} 中的元素渐近独立.

主成分的渐进性质 (Asymptotic Properties of the PCs)

- 例:** 如果 X_1, X_2, \dots, X_n 是来自 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的样本, 由于 $n\mathcal{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$, 根据定理 4.11 知

$$\sqrt{n-1} (\ell_j - \lambda_j) \xrightarrow{\mathcal{L}} N_1(0, 2\lambda_j^2), \quad j = 1, 2, \dots, p$$

▶ 设 $f(\ell_j) = \log(\ell_j)$, 则 $\left. \frac{d}{d\ell_j} f \right|_{\ell_j=\lambda_j} = \frac{1}{\lambda_j}$. ← $t = \ell_j, \boldsymbol{\mu} = \lambda_j, f = \log t, \mathcal{D} = \frac{1}{\lambda_j}$

定理 4.11 设 $\sqrt{n}(t - \boldsymbol{\mu}) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, $f = (f_1, f_2, \dots, f_q)^T : \mathbb{R}^p \rightarrow \mathbb{R}^q$ 为实值函数, 且在 $\boldsymbol{\mu} \in \mathbb{R}^p$ 可微, 则 $f(t)$ 渐近服从均值向量为 $f(\boldsymbol{\mu})$ 、协方差矩阵为 $\mathcal{D}^T \boldsymbol{\Sigma} \mathcal{D}$ 的正态分布, 即

$$\sqrt{n} [f(t) - f(\boldsymbol{\mu})] \xrightarrow{\mathcal{L}} N_q(\mathbf{0}, \mathcal{D}^T \boldsymbol{\Sigma} \mathcal{D}), \quad n \rightarrow \infty$$

其中

$$\mathcal{D} = \left(\frac{\partial f_j}{\partial t_i} \right) (t) \Big|_{t=\boldsymbol{\mu}}$$

是所有偏导数的 $(p \times q)$ 矩阵.

主成分的渐进性质 (Asymptotic Properties of the PCs)

- 例:** 如果 X_1, X_2, \dots, X_n 是来自 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的样本, 由于 $n\mathcal{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$, 根据定理 4.11 知

$$\sqrt{n-1} (\ell_j - \lambda_j) \xrightarrow{\mathcal{L}} N_1(0, 2\lambda_j^2), \quad j = 1, 2, \dots, p$$

▶ 设 $f(\ell_j) = \log(\ell_j)$, 则 $\left. \frac{d}{d\ell_j} f \right|_{\ell_j=\lambda_j} = \frac{1}{\lambda_j}$.

$$\sqrt{n-1} (\log \ell_j - \log \lambda_j) \xrightarrow{\mathcal{L}} N_1(0, 2)$$

$$\Rightarrow \sqrt{\frac{n-1}{2}} (\log \ell_j - \log \lambda_j) \xrightarrow{\mathcal{L}} N_1(0, 1)$$

$$\Rightarrow P\left(-z_{\alpha/2} \leq \sqrt{\frac{n-1}{2}} (\log \ell_j - \log \lambda_j) \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

$$\Rightarrow P\left(\log \ell_j - 1.96 \sqrt{\frac{2}{n-1}} \leq \log \lambda_j \leq \log \ell_j + 1.96 \sqrt{\frac{2}{n-1}}\right) \approx 0.95$$

定理 4.11 设 $\sqrt{n}(\mathbf{t} - \boldsymbol{\mu}) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, $f = (f_1, f_2, \dots, f_q)^T: \mathbb{R}^p \rightarrow \mathbb{R}^q$ 为实值函数, 且在 $\boldsymbol{\mu} \in \mathbb{R}^p$ 可微, 则 $f(\mathbf{t})$ 渐近服从均值向量为 $f(\boldsymbol{\mu})$ 、协方差矩阵为 $\mathcal{D}^T \boldsymbol{\Sigma} \mathcal{D}$ 的正态分布, 即

$$\sqrt{n} [f(\mathbf{t}) - f(\boldsymbol{\mu})] \xrightarrow{\mathcal{L}} N_q(\mathbf{0}, \mathcal{D}^T \boldsymbol{\Sigma} \mathcal{D}), \quad n \rightarrow \infty$$

其中

$$\mathcal{D} = \left(\frac{\partial f_j}{\partial t_i} \right) \bigg|_{\mathbf{t}=\boldsymbol{\mu}}$$

是所有偏导数的 $(p \times q)$ 矩阵.

$\rightarrow z_{0.025} = 1.96$

主成分的渐进性质 (Asymptotic Properties of the PCs)

- **例:** 如果 X_1, X_2, \dots, X_n 是来自 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的样本, 由于 $n\mathcal{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$, 根据定理 4.11 知

$$\sqrt{n-1} (\ell_j - \lambda_j) \xrightarrow{\mathcal{L}} N_1(0, 2\lambda_j^2), \quad j = 1, 2, \dots, p$$

- ▶ 在银行钞票数据的例子中, 我们有 $\ell_1 = 2.98$, 所以

$$P\left(\log(2.98) - 1.96 \sqrt{\frac{2}{200-1}} \leq \log \lambda_1 \leq \log(2.98) + 1.96 \sqrt{\frac{2}{200-1}}\right) \approx 0.95$$

$$\Rightarrow P[\lambda_1 \in (2.448, 3.62)] \approx 0.95$$

$$\Rightarrow P\left(\log \ell_j - 1.96 \sqrt{\frac{2}{n-1}} \leq \log \lambda_j \leq \log \ell_j + 1.96 \sqrt{\frac{2}{n-1}}\right) \approx 0.95$$

主成分的渐进性质 (Asymptotic Properties of the PCs)

- 前 q 个主成分解释的方差

- 前 q 个主成分所解释方差的比例由下式给出

$$\psi = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_q}{\lambda_1 + \lambda_2 + \dots + \lambda_p}$$

- 在实际应用中, 可由下式估计

$$\widehat{\psi} = \frac{\ell_1 + \ell_2 + \dots + \ell_q}{\ell_1 + \ell_2 + \dots + \ell_p}$$

定理 4.11 设 $\sqrt{n}(t - \mu) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \Sigma)$, $f = (f_1, f_2, \dots, f_q)^T: \mathbb{R}^p \rightarrow \mathbb{R}^q$ 为实值函数, 且在 $\mu \in \mathbb{R}^p$ 可微, 则 $f(t)$ 渐近服从均值向量为 $f(\mu)$ 、协方差矩阵为 $\mathcal{D}^T \Sigma \mathcal{D}$ 的正态分布, 即

$$\sqrt{n}[f(t) - f(\mu)] \xrightarrow{\mathcal{L}} N_q(\mathbf{0}, \mathcal{D}^T \Sigma \mathcal{D}), \quad n \rightarrow \infty$$

其中

$$\mathcal{D} = \left(\frac{\partial f_j}{\partial t_i} \right) (t) \Big|_{t=\mu}$$

是所有偏导数的 $(p \times q)$ 矩阵.

定理 11.4 设 $\Sigma > 0$ 且有不同的特征值, 再设 $\mathcal{S} \sim n^{-1}W_p(\Sigma, n-1)$, 且有谱分解

$$\Sigma = \Gamma \Lambda \Gamma^T \text{ 与 } \mathcal{S} = \mathcal{L} \mathcal{L}^T. \text{ 则}$$

(a) $\sqrt{n-1}(\ell - \lambda) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, 2\Lambda^2)$, 其中 $\ell = (\ell_1, \ell_2, \dots, \ell_p)^T$ 与

$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_p)^T$ 是 \mathcal{L} 与 Λ 的对角元素.

(b) $\sqrt{n-1}(\mathbf{g}_j - \gamma_j) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \mathcal{V}_j)$, 其中 $\mathcal{V}_j = \lambda_j \sum_{k \neq j} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \gamma_k \gamma_k^T$.

(c) ℓ 中的元素与矩阵 \mathcal{D} 中的元素渐近独立.

$$\Rightarrow \sqrt{n-1}(\ell - \lambda) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, 2\Lambda^2)$$

$$\sqrt{n-1}(\widehat{\psi} - \psi) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, 2\mathcal{D}^T \mathcal{V} \mathcal{D}) \quad \mathcal{V} = 2\Lambda^2$$

$$d_j = \frac{\partial \psi}{\partial \lambda_j} = \begin{cases} \frac{1 - \psi}{\text{tr}(\Sigma)}, & 1 \leq j \leq q \\ \frac{-\psi}{\text{tr}(\Sigma)}, & q + 1 \leq j \leq p \end{cases}$$

$$\mathcal{D} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{pmatrix}$$

$t = \ell, \mu = \lambda, f = \widehat{\psi}$

主成分的渐进性质 (Asymptotic Properties of the PCs)

- 前 q 个主成分解释的方差

定理 11.5

$$\sqrt{n-1} (\widehat{\psi} - \psi) \xrightarrow{\mathcal{L}} N_1(0, \omega^2)$$

其中

$$\begin{aligned} \omega^2 &= \mathcal{D}^T \mathcal{V} \mathcal{D} = \frac{2}{[\text{tr}(\Sigma)]^2} \left[(1 - \psi^2) (\lambda_1^2 + \dots + \lambda_q^2) + \psi^2 (\lambda_{q+1}^2 + \dots + \lambda_p^2) \right] \\ &= \frac{2\text{tr}(\Sigma^2)}{[\text{tr}(\Sigma)]^2} (\psi^2 - 2\beta\psi + \beta) \end{aligned}$$

以及

$$\beta = \frac{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_q^2}{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_p^2}$$

主成分的渐进性质 (Asymptotic Properties of the PCs)

- **例:** 对于瑞士银行纸币数据集, 已知第一主成分能解释 67% 的变异.
 - ▶ 检验 $H_0 : \psi = 0.75 \leftrightarrow H_1 : \psi \neq 0.75$.

$$\hat{\beta} = \frac{\ell_1^2}{\ell_1^2 + \ell_2^2 + \dots + \ell_p^2} = \frac{2.985^2}{2.985^2 + 0.931^2 + 0.242^2 + 0.194^2 + 0.085^2 + 0.035^2} = 0.902$$

$$\text{tr}(\mathcal{S}) = 4.472$$

$$\text{tr}(\mathcal{S}^2) = \sum_{j=1}^p \ell_j^2 = 9.883$$

$$\begin{aligned} \hat{\omega}^2 &= \frac{2\text{tr}(\mathcal{S}^2)}{[\text{tr}(\mathcal{S})]^2} \left(\hat{\psi}^2 - 2\hat{\beta}\hat{\psi} + \hat{\beta} \right) \\ &= \frac{2 \times 9.883}{4.472^2} (0.668^2 - 2 \times 0.902 \times 0.668 + 0.902) \\ &= 0.142 \end{aligned}$$

主成分的渐进性质 (Asymptotic Properties of the PCs)

- **例:** 对于瑞士银行纸币数据集, 已知第一主成分能解释 67% 的变异.

▶ 检验 $H_0: \psi = 0.75 \leftrightarrow H_1: \psi \neq 0.75$.

▶ ψ 的 $1 - \alpha = 0.95$ 置信区间为

$$\left(0.668 - 1.96\sqrt{\frac{0.142}{200-1}}, 0.668 + 1.96\sqrt{\frac{0.142}{200-1}} \right) = (0.615, 0.720)$$

▶ **结论:** 拒绝假设 $H_0: \psi = 0.75$.

定理 11.5

$$\sqrt{n-1}(\hat{\psi} - \psi) \xrightarrow{\mathcal{L}} N_1(0, \omega^2)$$

其中

$$\begin{aligned} \omega^2 &= \mathcal{D}^T \mathcal{V} \mathcal{D} = \frac{2}{[\text{tr}(\Sigma)]^2} \left[(1 - \psi^2) (\lambda_1^2 + \dots + \lambda_q^2) + \psi^2 (\lambda_{q+1}^2 + \dots + \lambda_p^2) \right] \\ &= \frac{2\text{tr}(\Sigma^2)}{[\text{tr}(\Sigma)]^2} (\psi^2 - 2\beta\psi + \beta) \end{aligned}$$

以及

$$\beta = \frac{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_q^2}{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_p^2}$$

归一化主成分分析 (Normalized Principal Components Analysis)

- 在某些情况下，原始变量在方差方面可能存在差异。
 - ▶ 当变量是在不同尺度 (如年、千克、美元等) 上进行测量时，这种情况尤为明显。
 - ▶ 在这种情况下，需要提供一种对数据所含信息的描述，该描述对于尺度的选择应具有稳健性。
 - ▶ 这可以通过对变量进行标准化来实现

$$\mathcal{X}_s = \mathcal{H} \mathcal{X} \mathcal{D}^{-1/2} \quad \mathcal{D} = \text{diag} (s_{X_1 X_1}, s_{X_2 X_2}, \dots, s_{X_p X_p})$$
$$\implies \bar{\mathbf{x}}_s = \mathbf{0}, \quad S_{\mathcal{X}_s} = \mathcal{R} \quad \mathcal{X} \text{ 的相关矩阵}$$

- ▶ 矩阵 \mathcal{X}_s 的主成分变换被称为归一化主成分 (NPCs).

归一化主成分分析 (Normalized Principal Components Analysis)

- 设 \mathcal{R} 的谱分解为

$$\mathcal{R} = \mathcal{G}_{\mathcal{R}} \mathcal{L}_{\mathcal{R}} \mathcal{G}_{\mathcal{R}}^T$$

$\mathcal{L}_{\mathcal{R}} = \text{diag}(\ell_1^{\mathcal{R}}, \ell_2^{\mathcal{R}}, \dots, \ell_p^{\mathcal{R}})$

\mathcal{R} 的特征值: $\ell_1^{\mathcal{R}} \geq \ell_2^{\mathcal{R}} \geq \dots \geq \ell_p^{\mathcal{R}}$

对应的特征向量: $(\mathbf{g}_1^{\mathcal{R}}, \mathbf{g}_2^{\mathcal{R}}, \dots, \mathbf{g}_p^{\mathcal{R}}) = \mathcal{G}_{\mathcal{R}}$

$$\Rightarrow \sum_{j=1}^p \ell_j^{\mathcal{R}} = \text{tr}(\mathcal{R}) = p$$

- 归一化主成分, 即 z_j , 为每个个体提供一种表征, 由以下方式得出

$$\mathcal{Z} = \mathcal{X}_{\mathcal{S}} \mathcal{G}_{\mathcal{R}} = (z_1, z_2, \dots, z_p)$$

$$\Rightarrow \bar{z} = \mathbf{0}$$

$$\mathcal{S}_{\mathcal{Z}} = \mathcal{G}_{\mathcal{R}}^T \mathcal{S}_{\mathcal{X}_{\mathcal{S}}} \mathcal{G}_{\mathcal{R}} = \mathcal{G}_{\mathcal{R}}^T \mathcal{R} \mathcal{G}_{\mathcal{R}} = \mathcal{L}_{\mathcal{R}}$$

归一化主成分分析 (Normalized Principal Components Analysis)



归一化主成分 (NPCs) 提供了与主成分 (PCs) 类似的视角, 但就个体的相对位置而言, NPC 给每个变量赋予相同的权重 (而在主成分分析中, 方差最大的变量获得最大权重).

- 计算 X_i 和 Z_j 之间的协方差与相关性很简单:

$$\mathcal{S}_{X_S, Z} = \frac{1}{n} \mathcal{X}_S^T \mathcal{Z} = \mathcal{G}_R \mathcal{L}_R$$

$$\mathcal{R}_{X_S, Z} = \mathcal{G}_R \mathcal{L}_R \mathcal{L}_R^{-1/2} = \mathcal{G}_R \mathcal{L}_R^{1/2}$$

- ▶ 原始变量 X_i 与归一化主成分 Z_j 之间的相关系数为

$$r_{X_i Z_j} = \sqrt{\ell_j} g_{R, ij}$$

$$\sum_{j=1}^p r_{X_i Z_j}^2 = 1$$

- ▶ 由此得到的归一化主成分, 即 Z_j , 可以从原始变量的角度进行解释, 并且可以评估每个主成分在解释变量 X_i 的变异方面所起的作用.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 事实证明，经验主成分 (无论是否归一化) 等同于通过将适当的数据矩阵分解为因子而得到的因子 (第10章).
 - ▶ 我们将证明，主成分就是代表经过**中心化**处理的数据矩阵行向量的因子.
 - ▶ 我们将证明，归一化主成分是代表**标准化**数据矩阵行的因子.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 假设我们想在更低维度的空间中得到观测值 (即 \mathcal{X} 的行) 和变量 (即 \mathcal{X} 的列) 的表示.
 - ▶ 由于原点在观测值空间中没有特殊的统计意义, 我们首先将原点平移到数据点的重心 \bar{x} 处.
 - ▶ 这与分析经过中心化的数据矩阵 $\mathcal{X}_C = \mathcal{H}\mathcal{X}$ 是一样的.
 - ▶ 现在所有变量的均值都为零; 因此, 第 10 章中使用的方法可应用于矩阵 \mathcal{X}_C .

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 假设我们想在更低维度的空间中得到观测值 (即 \mathcal{X} 的行) 和变量 (即 \mathcal{X} 的列) 的表示.

- 注意到矩阵 $\mathcal{X}_C^T \mathcal{X}_C$ 的谱分解与矩阵 \mathcal{S}_X 的谱分解相关,

$$\mathcal{X}_C^T \mathcal{X}_C = (\mathcal{H}\mathcal{X})^T \mathcal{H}\mathcal{X} = \mathcal{X}^T \mathcal{H}^T \mathcal{H}\mathcal{X} = n\mathcal{S}_X = n\mathcal{G}\mathcal{L}\mathcal{G}^T$$

\mathcal{S}_X 的谱分解

- 因子变量是通过将 \mathcal{X}_C 投影到 \mathcal{G} 上得到的

$$\mathcal{Y} = \mathcal{X}_C \mathcal{G} = (y_1, y_2, \dots, y_p)$$

$$\mathcal{Y} = (\mathcal{X} - \mathbf{1}_n \bar{\mathbf{x}}^T) \mathcal{G}$$

与上述得到的主成分相同

$$\Rightarrow \bar{\mathbf{y}} = \frac{1}{n} \mathcal{Y}^T \mathbf{1}_n = \frac{1}{n} \left[(\mathcal{X} - \mathbf{1}_n \bar{\mathbf{x}}^T) \mathcal{G} \right]^T \mathbf{1}_n = \mathcal{G}^T \frac{1}{n} (\mathcal{X}^T - \bar{\mathbf{x}} \mathbf{1}_n^T) \mathbf{1}_n$$

$$= \mathcal{G}^T \left(\frac{1}{n} \mathcal{X}^T \mathbf{1}_n - \frac{1}{n} \bar{\mathbf{x}} \mathbf{1}_n^T \mathbf{1}_n \right) = \mathcal{G}^T (\bar{\mathbf{x}} - \bar{\mathbf{x}}) = \mathbf{0}$$

$$\Rightarrow \mathcal{S}_Y = \frac{1}{n} \mathcal{Y}^T \mathcal{H} \mathcal{Y} = \frac{1}{n} (\mathcal{X}_C \mathcal{G})^T \mathcal{H} \mathcal{X}_C \mathcal{G} = \frac{1}{n} \mathcal{G}^T \mathcal{X}_C^T \mathcal{H} \mathcal{X}_C \mathcal{G} = \mathcal{G}^T \left(\frac{1}{n} \mathcal{X}_C^T \mathcal{X}_C \right) \mathcal{G}$$

$\mathcal{H} \mathcal{X}_C = \mathcal{X}_C$

$$= \mathcal{G}^T \left(\frac{1}{n} \mathcal{X}_C^T \mathcal{H} \mathcal{X}_C \right) \mathcal{G} = \mathcal{G}^T \mathcal{S}_X \mathcal{G} = \mathcal{G}^T \mathcal{G} \mathcal{L} \mathcal{G}^T \mathcal{G} = \mathcal{L} = \text{diag}(\ell_1, \ell_2, \dots, \ell_p)$$

\mathcal{I}_p

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 假设我们想在更低维度的空间中得到观测值 (即 \mathcal{X} 的行) 和变量 (即 \mathcal{X} 的列) 的表示.
 - ▶ 因此, 观测值在因子轴上的散点图以原点为中心.
 - ▶ 观测值的散点图在第一个方向 (第一主成分的方差为 ℓ_1) 上比在第二个方向 (第二主成分的方差为 ℓ_2) 上分布得更分散.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 变量的表示形式可以通过对偶关系得到.

定理 10.4 (对偶关系) 设 \mathcal{X} 的秩为 r . 对于 $k \leq r$, $\mathcal{X}^T \mathcal{X}$ 与 $\mathcal{X} \mathcal{X}^T$ 的第 k 个特征值 λ_k 相同, 对应的特征向量 (分别记为 \mathbf{u}_k 与 \mathbf{v}_k) 之间的关系如下

$$\mathbf{u}_k = \frac{1}{\sqrt{\lambda_k}} \mathcal{X}^T \mathbf{v}_k, \quad \mathbf{v}_k = \frac{1}{\sqrt{\lambda_k}} \mathcal{X} \mathbf{u}_k.$$

- ▶ \mathcal{X}_C 的列在 $\mathcal{X}_C \mathcal{X}_C^T$ 的特征向量 \mathbf{v}_k 上的投影为

$$\mathcal{X}_C^T \mathbf{v}_k = \frac{1}{\sqrt{n\ell_k}} \mathcal{X}_C^T \mathcal{X}_C \mathbf{g}_k = \frac{1}{\sqrt{n\ell_k}} (n\mathcal{S}_X) \mathbf{g}_k = \frac{n}{\sqrt{n\ell_k}} \mathcal{S}_X \mathbf{g}_k = \frac{n\ell_k}{\sqrt{n\ell_k}} \mathbf{g}_k = \sqrt{n\ell_k} \mathbf{g}_k$$

- ▶ 因此, 变量在前 p 个轴上的投影就是该矩阵的列

$$\mathcal{X}_C^T \mathcal{V} = \sqrt{n} \mathcal{G} \mathcal{L}^{1/2}$$

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 对于矩阵 \mathcal{X}_C 两列之间夹角，存在一种很好的统计学解释。

$$s_{X_j X_k} = \frac{1}{n} \mathbf{x}_{C[j]}^T \mathbf{x}_{C[k]} \implies \mathbf{x}_{C[j]}^T \mathbf{x}_{C[k]} = n s_{X_j X_k}$$

$$\implies \|\mathbf{x}_{C[j]}\|^2 = n s_{X_j X_j}$$

$$\mathcal{X}_C = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2j} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix}$$

$\mathbf{x}_{C[j]} \longleftarrow$ $\longrightarrow \mathbf{x}_{C[k]}$

- 用 θ_{jk} 表示两个变量 $\mathbf{x}_{C[j]}$ 与 $\mathbf{x}_{C[k]}$ 的夹角，则

$$\cos \theta_{jk} = \frac{\mathbf{x}_{C[j]}^T \mathbf{x}_{C[k]}}{\|\mathbf{x}_{C[j]}\| \cdot \|\mathbf{x}_{C[k]}\|} = r_{X_j X_k}$$

- 在矩阵 $\mathcal{X}_C^T \mathcal{V}$ 第一列的散点图中，变量的相对位置可以根据它们之间的相关性来解释，该图展示了原始数据集的相关结构。
- 显然，在评估相关性时，应该考虑所选因子轴所解释的方差百分比。

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 归一化主成分 (NPC) 也可被视为一种用于降维的因子方法.
 - ▶ 对变量进行标准化处理, 使得每个变量的均值为零、方差为一, 从而与变量的度量单位无关.
 - ▶ 对 \mathcal{X}_S 的因子分析可得出归一化主成分.
 - ▶ 矩阵 $\mathcal{X}_S^T \mathcal{X}_S$ 的谱分解与矩阵 \mathcal{R} 的谱分解相关.

$$\mathcal{X}_S^T \mathcal{X}_S = \mathcal{D}^{-1/2} \mathcal{X}^T \mathcal{H} \mathcal{X} \mathcal{D}^{-1/2} = n\mathcal{R} = n\mathcal{G}_{\mathcal{R}} \mathcal{L}_{\mathcal{R}} \mathcal{G}_{\mathcal{R}}^T$$

- ▶ 归一化主成分 \mathbf{z}_j 可被视为 \mathcal{X}_S 的行向量在 $\mathcal{G}_{\mathcal{R}}$ 上的投影.

$$\mathcal{F} = \mathcal{X}_S \mathcal{G}_{\mathcal{R}} = \left(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_p \right)$$

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 归一化主成分 (NPC) 也可被视为一种用于降维的因子方法.

- ▶ 变量的表示形式同样可以由下述矩阵的列给出

$$\mathbf{X}_S^T \mathbf{V}_R = \sqrt{n} \mathbf{G}_R \mathbf{L}_R^{1/2} \quad \longleftrightarrow \quad \mathbf{R}_{X_S, \mathcal{I}} = \mathbf{G}_R \mathbf{L}_R^{1/2}$$

- ▶ 在因子分析中, 变量的投影给出了归一化主成分 (NPC) \mathcal{I}_k 与原始变量 $\mathbf{x}_{[j]}$ 之间的相关性 (相差因子 \sqrt{n} , 该因子可能是坐标轴的尺度).

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 这意味着，通过同时查看绘制有变量的图形，可以对观测值的表征进行更深入的解读。

$$r_{X_j X_k} = \frac{1}{n} \mathbf{x}_{S[j]}^T \mathbf{x}_{S[k]} \implies \mathbf{x}_{S[j]}^T \mathbf{x}_{S[k]} = n r_{X_j X_k}$$

$$\implies \|\mathbf{x}_{S[j]}\|^2 = n$$

$$\mathcal{X}_S = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2j} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix}$$

$\mathbf{x}_{S[j]}$ ← → $\mathbf{x}_{S[k]}$

- 因此，在整个空间中，所有标准化变量 (\mathcal{X}_S 的列) 都包含在 \mathbb{R}^n 中的“球体”内，该球体以原点为中心，半径为 \sqrt{n} (图形的尺度)。
- 设 θ_{jk} 表示两个变量 $\mathbf{x}_{S[j]}$ 与 $\mathbf{x}_{S[k]}$ 之间的夹角，则 $\cos \theta_{jk} = r_{X_j X_k}$ 。
- 因此，当观察降维空间 (例如，前两个因子) 中变量的表示时，我们能从变量间夹角的角度，了解原始变量 X_i 之间的相关结构。
- 当然，这些子空间中表示的质量必须予以考虑，下一节将对此进行阐述。

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 表示的质量

- ▶ 表示质量的总体衡量指标由下式给出

$$\psi = \frac{\ell_1 + \ell_2 + \cdots + \ell_q}{\sum_{j=1}^p \ell_j}$$

- ▶ 在实际应用当中, q 往往取 1、2 或 3.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

● 表示的质量

- ▶ 检查每个个体是否能通过主成分得到良好呈现会很有用.
- ▶ 显然, 在投影空间中两组观测值的接近程度不一定与在完整的原始空间 \mathbb{R}^p 中的接近程度一致, 这可能会导致对图示的错误解读.
- ▶ 基于此, 计算个体 i 的表示与第 k 个主成分 (PC) 或归一化主成分 (NPC) 轴之间的夹角 ϑ_{ik} 是有意义的.

$$\cos \vartheta_{ik} = \frac{\mathbf{y}_i^T \mathbf{e}_k}{\|\mathbf{y}_i\| \|\mathbf{e}_k\|} = \frac{y_{ik}}{\|\mathbf{x}_{C[i]}\|}$$

$$\cos \zeta_{ik} = \frac{\mathbf{z}_i^T \mathbf{e}_k}{\|\mathbf{z}_i\| \|\mathbf{e}_k\|} = \frac{z_{ik}}{\|\mathbf{x}_{S[i]}\|}$$

- ▶ 如果个体 i 与第 k 个主成分轴对应的夹角较小, 即对于 $k = 1, 2, \dots, p$, $\cos^2 \vartheta_{ik}$ 接近 1, 那么个体 i 将在第 k 个主成分轴上得到体现.

$$\mathbf{e}_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 表示的质量

- ▶ 因为对每个 i 有

$$\sum_{k=1}^p \cos^2 \vartheta_{ik} = \frac{\mathbf{y}_i^T \mathbf{y}_i}{\mathbf{x}_{C[i]}^T \mathbf{x}_{C[i]}} = \frac{\mathbf{x}_{C[i]}^T \mathcal{G} \mathcal{G}^T \mathbf{x}_{C[i]}}{\mathbf{x}_{C[i]}^T \mathbf{x}_{C[i]}} = 1$$

- ▶ $\cos^2 \vartheta_{ik}$ 的值有时被称为第 k 轴对第 i 个观测值表示的相对贡献.
- ▶ 例如, 如果 $\cos^2 \vartheta_{i1} + \cos^2 \vartheta_{i2}$ 的值较大 (接近 1), 那么第 i 个观测值在前两个主轴所构成的平面上能得到很好的呈现, 这是因为它与该平面的对应夹角的余弦值接近 1.
- ▶ 我们已经知道, 变量表示的质量可以通过主成分 (PC) 对 X_i 方差的解释百分比来评估, 该百分比分别由 $r_{X_i X_j}^2$ 或 $r_{X_i Z_j}^2$ 给出.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

```
rm(list = ls(all = TRUE)) # clear all variables
```

```
graphics.off()
```

```
setwd("~/Desktop/2023_Applied Multivariate Statistical Analysis/R Codes with data/Data")
```

```
library(readr)
```

```
x = read_csv("food.csv")
```

```
x
```

```
> x
# A tibble: 12 × 8
  family bread veget fruit meat poult milk wine
  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 MA2    332  428  354  1437  526  247  427
2 EM2    293  559  388  1527  567  239  258
3 CA2    372  767  562  1948  927  235  433
4 MA3    406  563  341  1507  544  324  407
5 EM3    386  608  396  1501  558  319  363
6 CA3    438  843  689  2345  1148  243  341
7 MA4    534  660  367  1620  638  414  407
8 EM4    460  699  484  1856  762  400  416
9 CA4    385  789  621  2366  1149  304  282
10 MA5    655  776  423  1848  759  495  486
11 EM5    584  995  548  2056  893  518  319
12 CA5    515 1097  887  2630  1167  561  284
```

```
p = ncol(x)
```

```
n = nrow(x)
```

```
x = x[, 2:p]
```

```
x1 = sqrt((n - 1) * apply(x, 2, var)/n)
```

```
x2 = x - matrix(apply(as.matrix(x), 2, mean), nrow = n, ncol = p - 1, byrow = T) # centered data
```

```
x = as.matrix(x2/matrix(x1, nrow = n, ncol = p - 1, byrow = T)) # standardizes the data matrix
```

```
round(x, digits = 3)
```

```
> round(x, digits = 3)
      bread veget fruit meat poult milk wine
[1,] -1.118 -1.678 -0.955 -1.187 -1.160 -0.992 0.850
[2,] -1.498 -0.955 -0.740 -0.949 -0.988 -1.063 -1.609
[3,] -0.728 0.193 0.361 0.162 0.518 -1.099 0.937
[4,] -0.396 -0.933 -1.038 -1.002 -1.085 -0.305 0.559
[5,] -0.591 -0.685 -0.690 -1.018 -1.026 -0.350 -0.081
[6,] -0.084 0.613 1.164 1.209 1.443 -1.028 -0.401
[7,] 0.851 -0.398 -0.873 -0.704 -0.691 0.497 0.559
[8,] 0.130 -0.182 -0.133 -0.081 -0.172 0.372 0.690
[9,] -0.601 0.315 0.734 1.265 1.447 -0.484 -1.260
[10,] 2.031 0.243 -0.519 -0.102 -0.185 1.219 1.708
[11,] 1.339 1.452 0.272 0.447 0.376 1.425 -0.721
[12,] 0.666 2.015 2.417 1.962 1.523 1.808 -1.231
```

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

```
R = t(x) %*% x / n # correlation matrix
round(R, digits = 3)
```

```
e = eigen(R) # eigenvalues and eigenvectors of correlation matrix
e1 = e$values
round(e1, digits = 3) # eigenvalues
round(e1 / sum(e1), digits = 4) # proportions of each NPC
round(cumsum(e1) / sum(e1), digits = 4) # cumulative proportions of each NPC
```

```
e2 = e$vectors
round(e2, digits = 3) # eigenvectors
```

```
> round(R, digits = 3)
      bread veget fruit  meat  poult  milk  wine
bread 1.000  0.593  0.196  0.321  0.248  0.856  0.304
veget 0.593  1.000  0.856  0.881  0.827  0.663 -0.356
fruit 0.196  0.856  1.000  0.959  0.926  0.332 -0.486
meat  0.321  0.881  0.959  1.000  0.982  0.375 -0.437
poult 0.248  0.827  0.926  0.982  1.000  0.233 -0.400
milk  0.856  0.663  0.332  0.375  0.233  1.000  0.007
wine  0.304 -0.356 -0.486 -0.437 -0.400 0.007  1.000
```

```
> round(e1, digits = 3) # eigenvalues
[1] 4.333 1.830 0.631 0.128 0.058 0.019 0.001
> round(e1 / sum(e1), digits = 4) # proportions of each NPC
[1] 0.6190 0.2615 0.0901 0.0183 0.0082 0.0027 0.0001
> round(cumsum(e1) / sum(e1), digits = 4) # cumulative proportions of each NPCs
[1] 0.6190 0.8805 0.9706 0.9890 0.9972 0.9999 1.0000
```

```
> round(e2, digits = 3) # eigenvectors
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,] -0.240 0.622 0.011 -0.544 -0.036 0.508 -0.024
[2,] -0.466 0.098 0.062 -0.023 0.809 -0.301 0.156
[3,] -0.446 -0.205 -0.145 0.548 0.067 0.625 -0.205
[4,] -0.462 -0.141 -0.207 -0.053 -0.411 -0.093 0.737
[5,] -0.438 -0.197 -0.356 -0.324 -0.224 -0.350 -0.605
[6,] -0.281 0.523 0.444 0.450 -0.341 -0.332 -0.151
[7,] 0.206 0.479 -0.780 0.306 0.069 -0.138 0.045
```

$G_R =$

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.
 - ▶ 通过查看原始变量 X_i 与主成分之间的相关性, 对主成分的解读会最为透彻.
 - ▶ 由于前两个主成分解释了 88.1% 的方差, 因此我们只关注前两个主成分.

$$r_{X_i Z_j} = \sqrt{\ell_j} g_{R, ij}$$

correlations between original variables and NPCs

```

a = e2[, 1:2]
w = a * sqrt(matrix(e1[1:2], nrow(a), ncol(a), byrow = TRUE))
round(w, digits = 3)
    
```

sum of squared correlations between original variables and first two NPCs

```
round(as.matrix(apply(w^2, 1, sum)), digits = 3)
```

```
> round(w, digits = 3)
```

	[,1]	[,2]
[1,]	-0.499	0.842
[2,]	-0.970	0.133
[3,]	-0.929	-0.278
[4,]	-0.962	-0.191
[5,]	-0.911	-0.266
[6,]	-0.584	0.707
[7,]	0.428	0.648

```
> round(as.matrix(apply(w^2, 1, sum)), digits = 3)
```

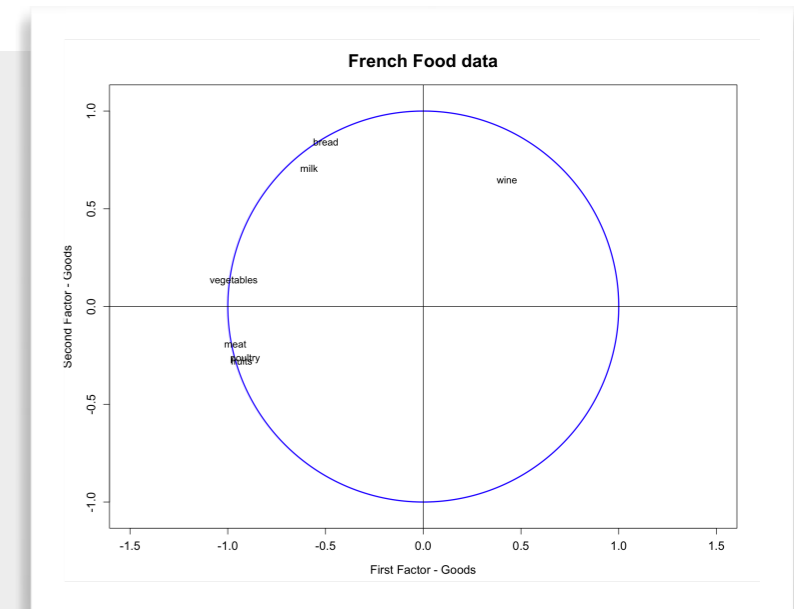
	[,1]
[1,]	0.957
[2,]	0.958
[3,]	0.941
[4,]	0.962
[5,]	0.901
[6,]	0.841
[7,]	0.603

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.
 - ▶ 变量的二维图形表示.

```
g = eigen(t(x) %*% x/n)
g1 = g$values
g2 = g$vectors
b = g2[, 1:2]
z = b * sqrt(matrix(g1[1:2], nrow(b), ncol(b), byrow = TRUE))

graphics.off()
ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))
plot(ucircle, type = "l", lty = "solid", col = "blue", xlim = c(-1.05, 1.05), ylim = c(-1.05, 1.05),
     xlab = "First Factor - Goods", ylab = "Second Factor - Goods", main = "French Food data",
     cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.8, lwd = 2, asp = 1)
abline(h = 0, v = 0)
label = c("bread", "vegetables", "fruits", "meat", "poultry", "milk", "wine")
text(z, label)
```

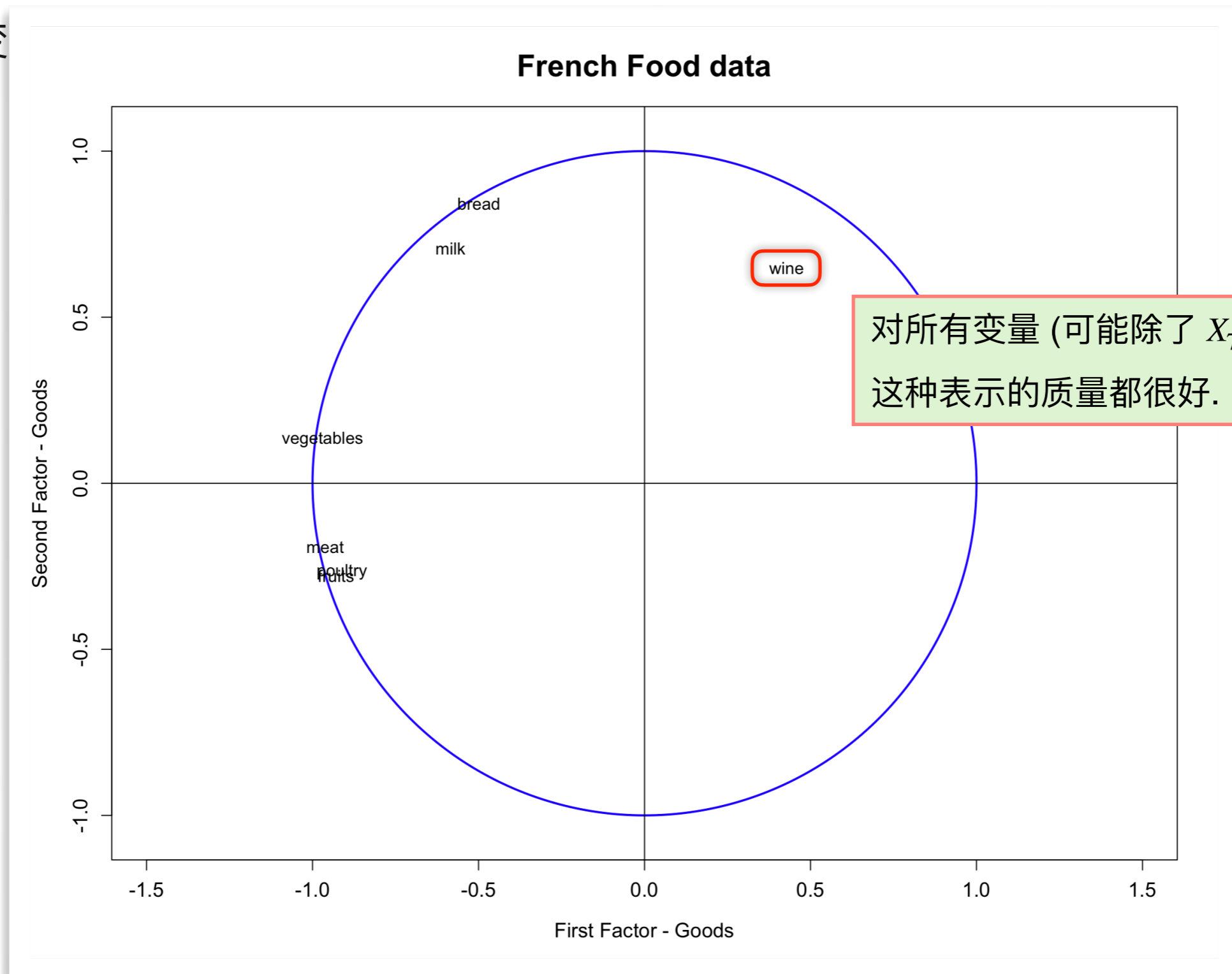


- ▶ 图形是变量在二维空间 \mathbb{R}^2 中的投影.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

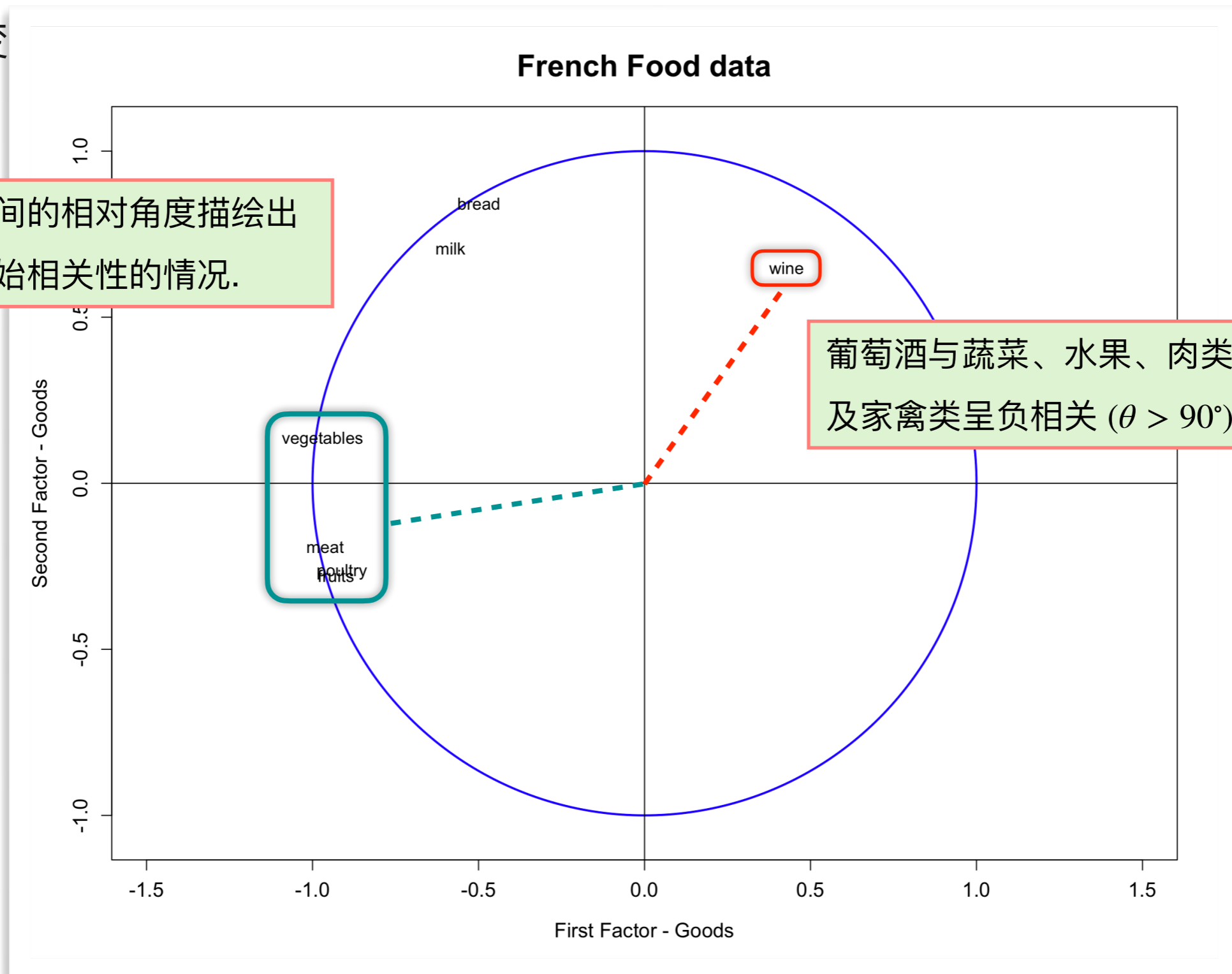
▶ 变



主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

▶ 变

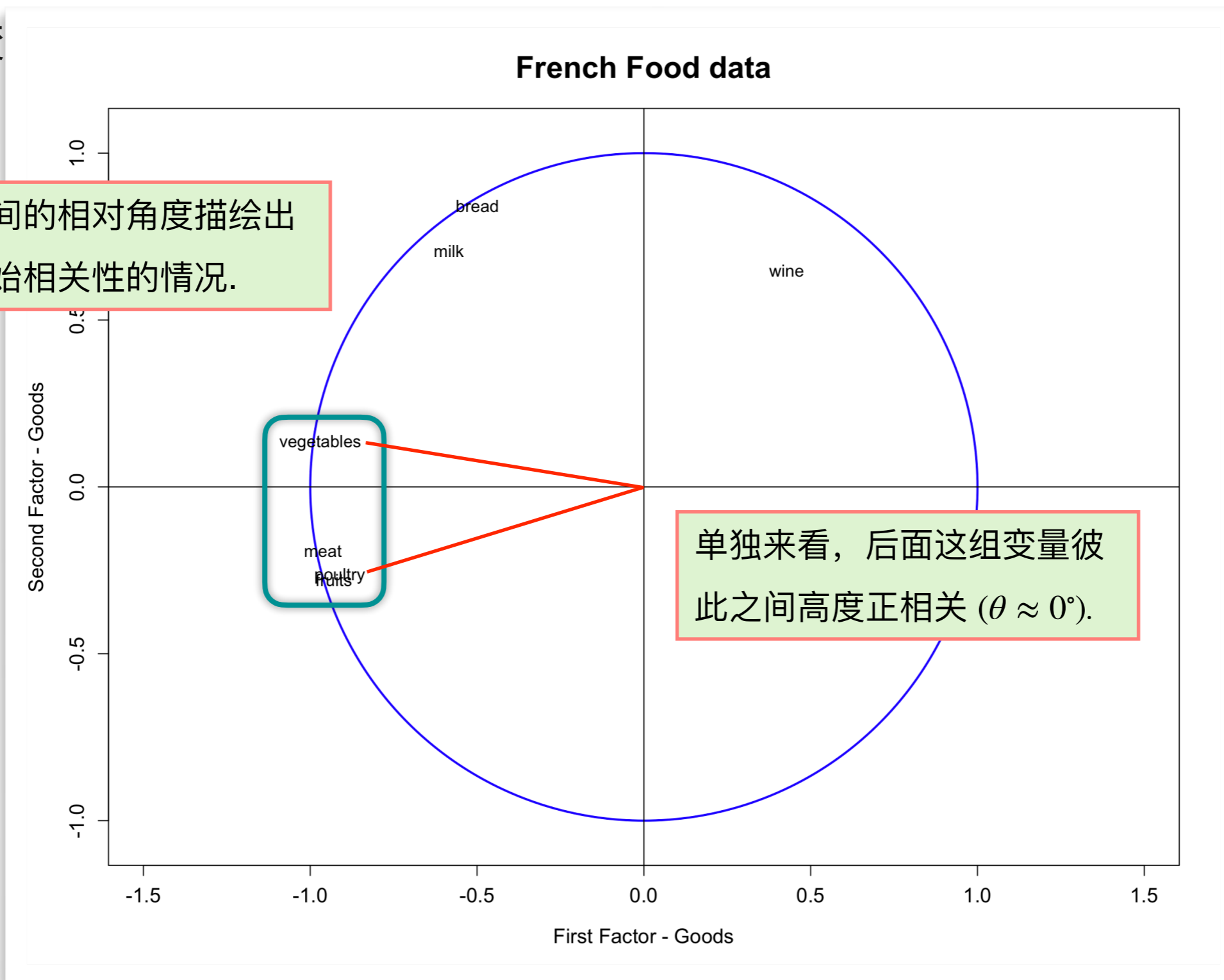


主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

▶ 变

它们之间的相对角度描绘出了其原始相关性的情况.



单独来看, 后面这组变量彼此之间高度正相关 ($\theta \approx 0^\circ$).

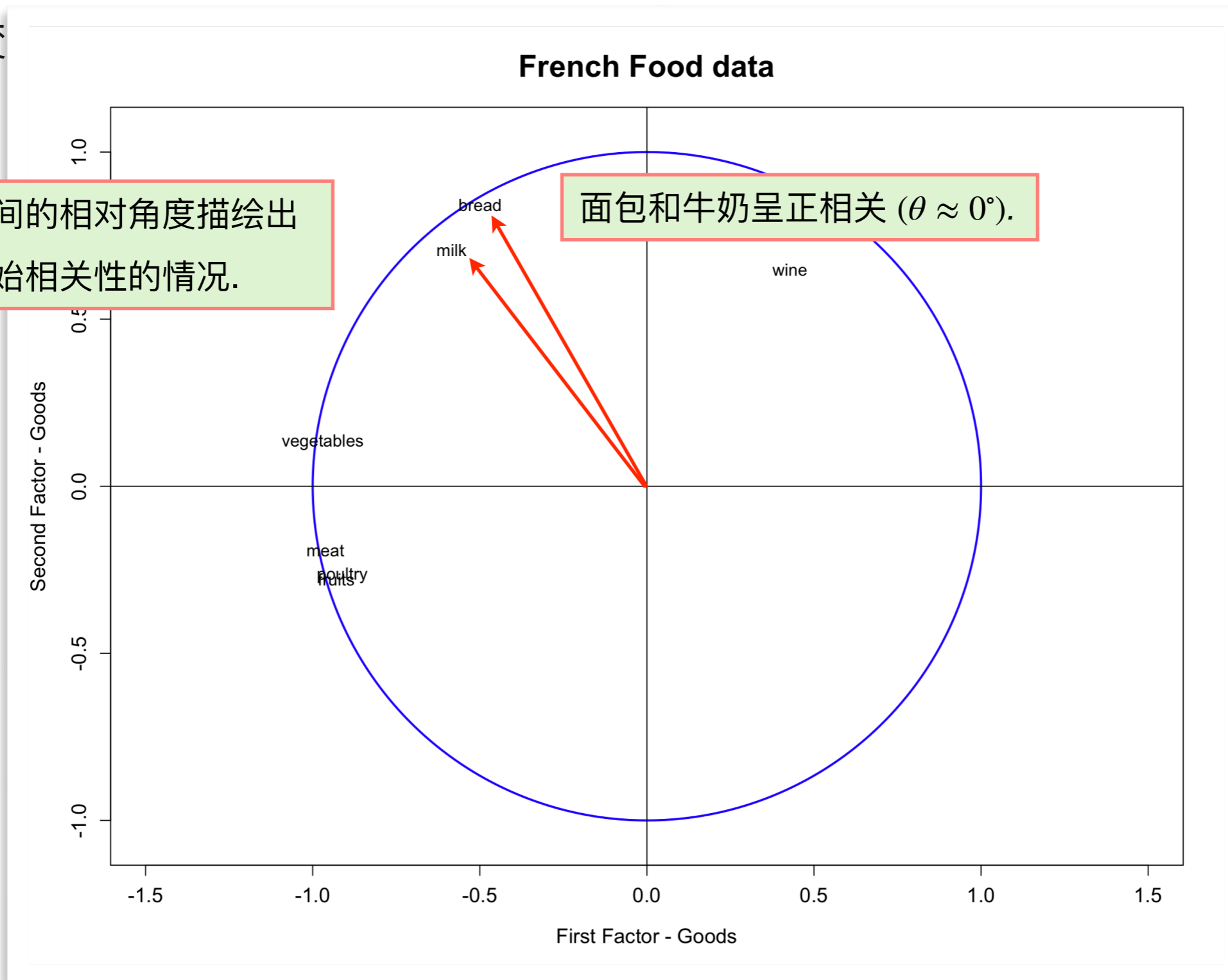
主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

▶ 变

它们之间的相对角度描绘出了其原始相关性的情况.

面包和牛奶呈正相关 ($\theta \approx 0^\circ$).

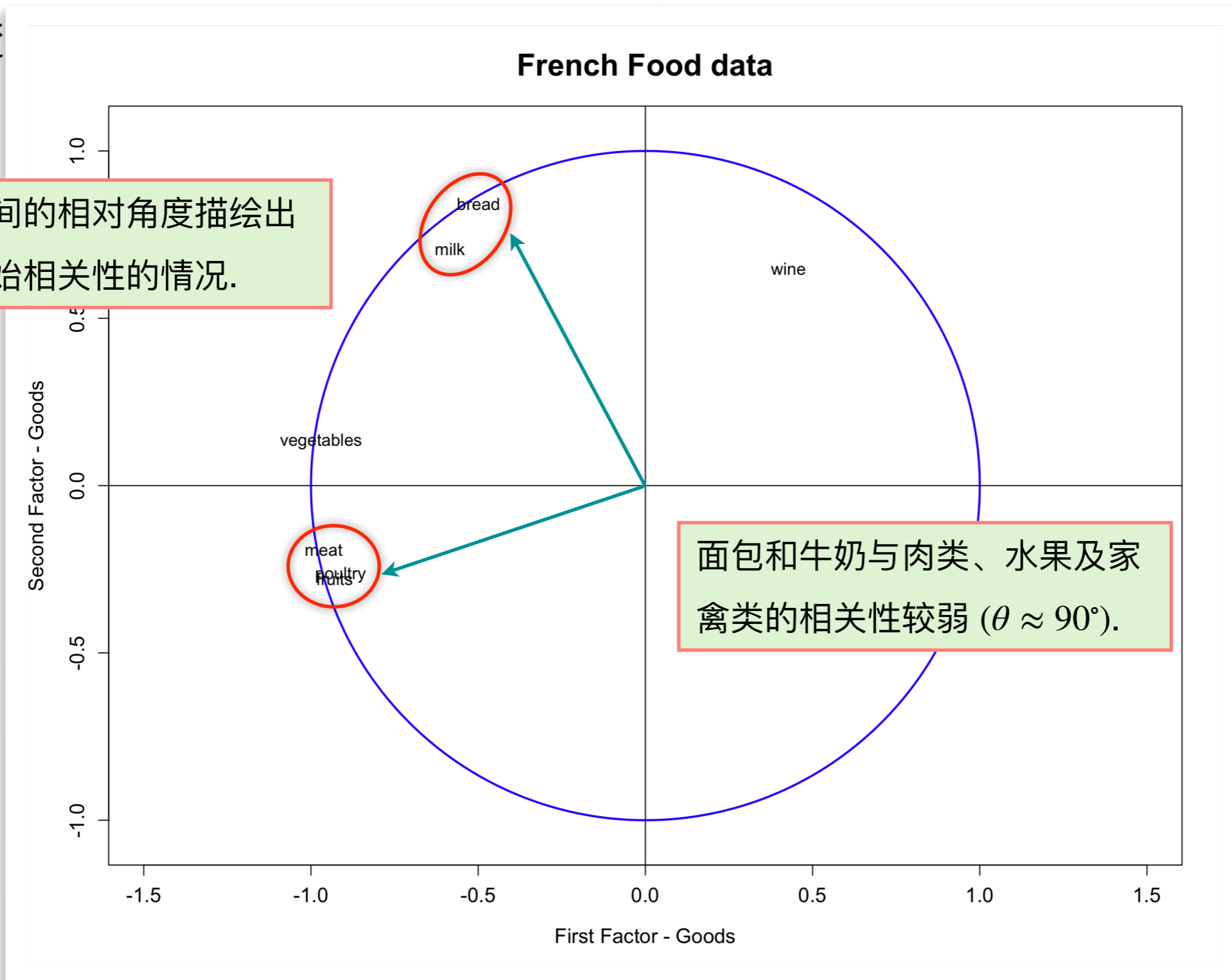


主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

▶ 变

它们之间的相对角度描绘出了其原始相关性的情况.



面包和牛奶与肉类、水果及家禽类的相关性较弱 ($\theta \approx 90^\circ$).

主成分作为一种因子方法 (Principal Components as a Factorial Method)

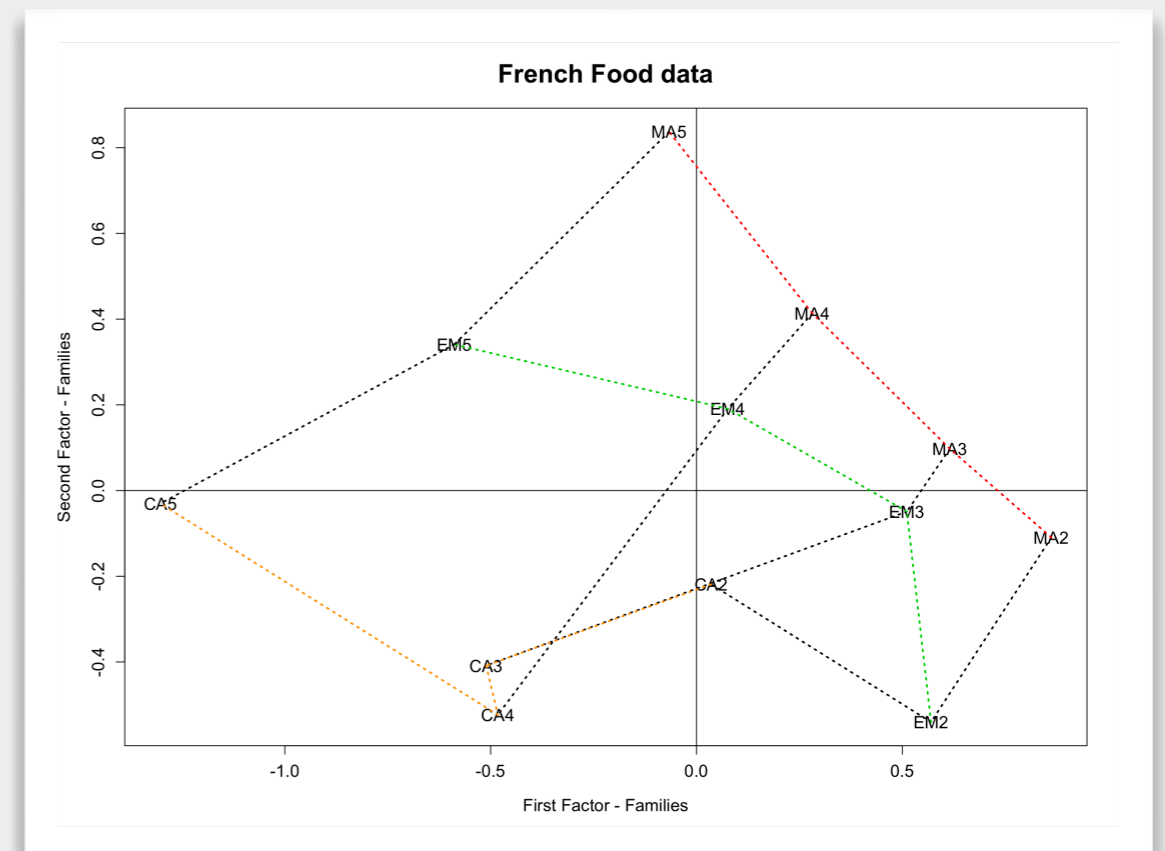
- 例: 法国食品支出数据集.

- ▶ 观测值的可视化.

```
e = eigen(x %*% t(x)/n)
e1 = e$values
e2 = e$vectors
a = e2[, 1:2]
w = -a * sqrt(matrix(e1[1:2], nrow(a), ncol(a), byrow = TRUE))

plot(w, type = "n", xlab = "First Factor - Families", ylab = "Second Factor - Families",
     main = "French Food data", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.8, lwd = 2)
text(w, c("MA2", "EM2", "CA2", "MA3", "EM3", "CA3", "MA4", "EM4", "CA4", "MA5", "EM5",
         "CA5"), cex = 1.2)
abline(h = 0, v = 0)
segments(w[1, 1], w[1, 2], w[2, 1], w[2, 2], lty = 3, lwd = 2)
segments(w[2, 1], w[2, 2], w[3, 1], w[3, 2], lty = 3, lwd = 2)
segments(w[1, 1], w[1, 2], w[4, 1], w[4, 2], lty = 3, lwd = 2, col = 'red')
segments(w[2, 1], w[2, 2], w[5, 1], w[5, 2], lty = 3, lwd = 2, col = 'green3')
segments(w[4, 1], w[4, 2], w[5, 1], w[5, 2], lty = 3, lwd = 2)
segments(w[5, 1], w[5, 2], w[6, 1], w[6, 2], lty = 3, lwd = 2)
segments(w[3, 1], w[3, 2], w[6, 1], w[6, 2], lty = 3, lwd = 2, col = 'orange')
segments(w[6, 1], w[6, 2], w[9, 1], w[9, 2], lty = 3, lwd = 2, col = 'orange')
segments(w[8, 1], w[8, 2], w[9, 1], w[9, 2], lty = 3, lwd = 2)
segments(w[5, 1], w[5, 2], w[8, 1], w[8, 2], lty = 3, lwd = 2, col = 'green3')
segments(w[7, 1], w[7, 2], w[8, 1], w[8, 2], lty = 3, lwd = 2)
segments(w[4, 1], w[4, 2], w[7, 1], w[7, 2], lty = 3, lwd = 2, col = 'red')
segments(w[7, 1], w[7, 2], w[10, 1], w[10, 2], lty = 3, lwd = 2, col = 'red')
segments(w[8, 1], w[8, 2], w[11, 1], w[11, 2], lty = 3, lwd = 2, col = 'green3')
segments(w[9, 1], w[9, 2], w[12, 1], w[12, 2], lty = 3, lwd = 2, col = 'orange')
segments(w[10, 1], w[10, 2], w[11, 1], w[11, 2], lty = 3, lwd = 2)
segments(w[11, 1], w[11, 2], w[12, 1], w[12, 2], lty = 3, lwd = 2)
```

图形是观测值在二维空间 \mathbb{R}^2 中的投影.

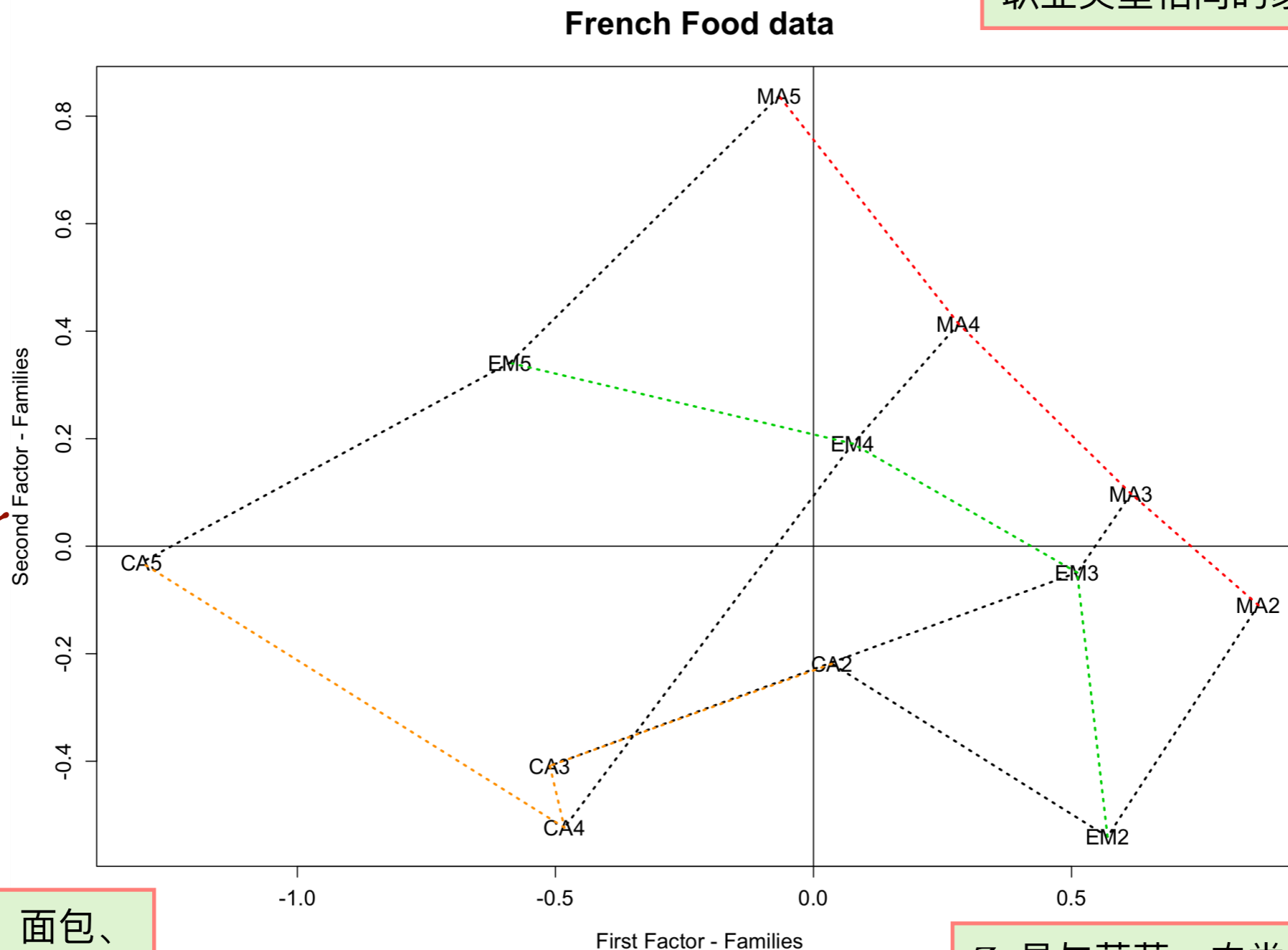


主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

▶ 观

绘制线条将规模相同的家庭以及职业类型相同的家庭连接起来.



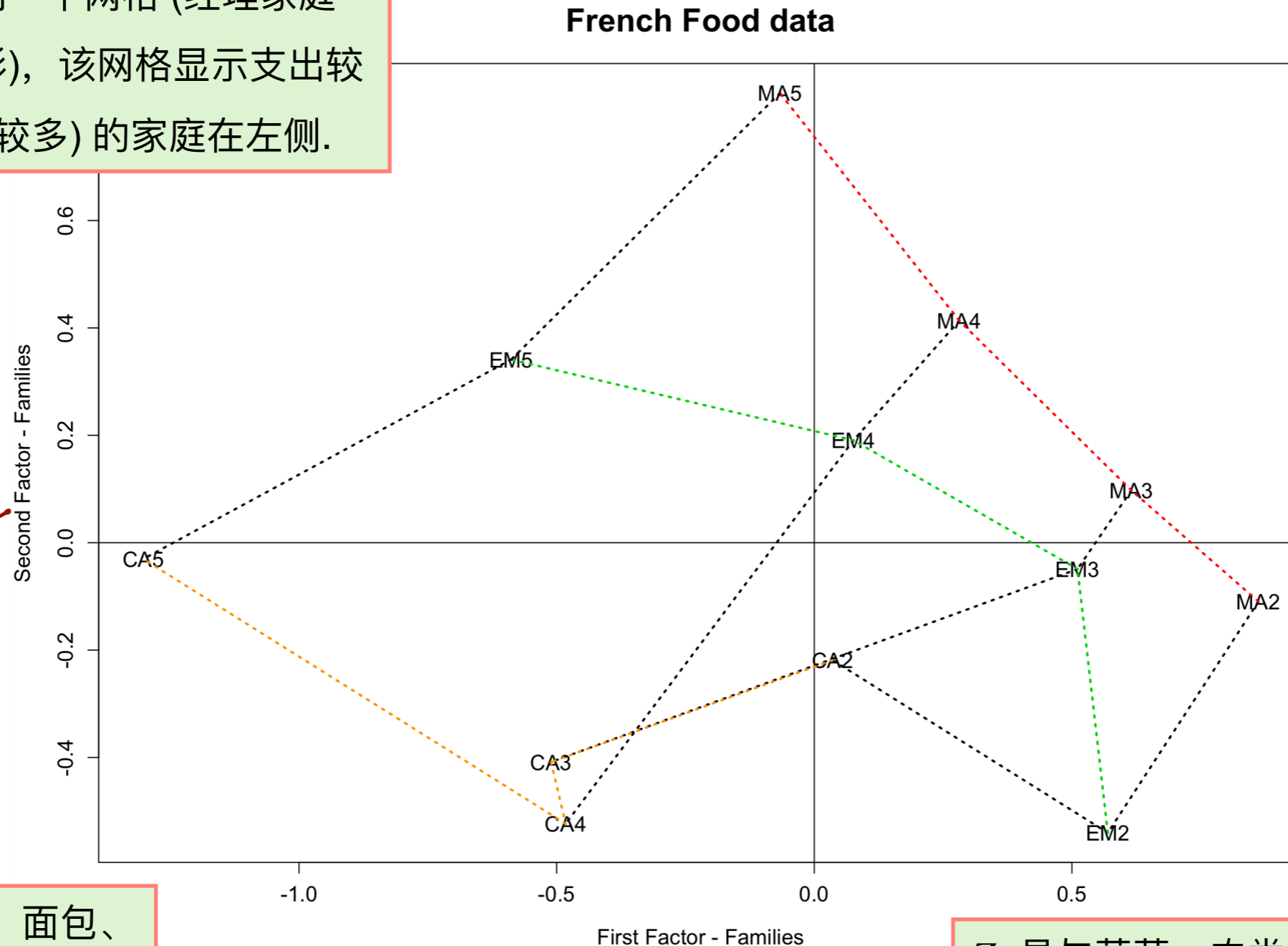
Z_2 是牛奶、面包、葡萄酒因子.

Z_1 是与蔬菜、肉类、家禽和水果相关的因子.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

- 例: 法国食品支出数据集.

可以清晰看到一个网格 (经理家庭使其略有变形), 该网格显示支出较高 (孩子数量较多) 的家庭在左侧.

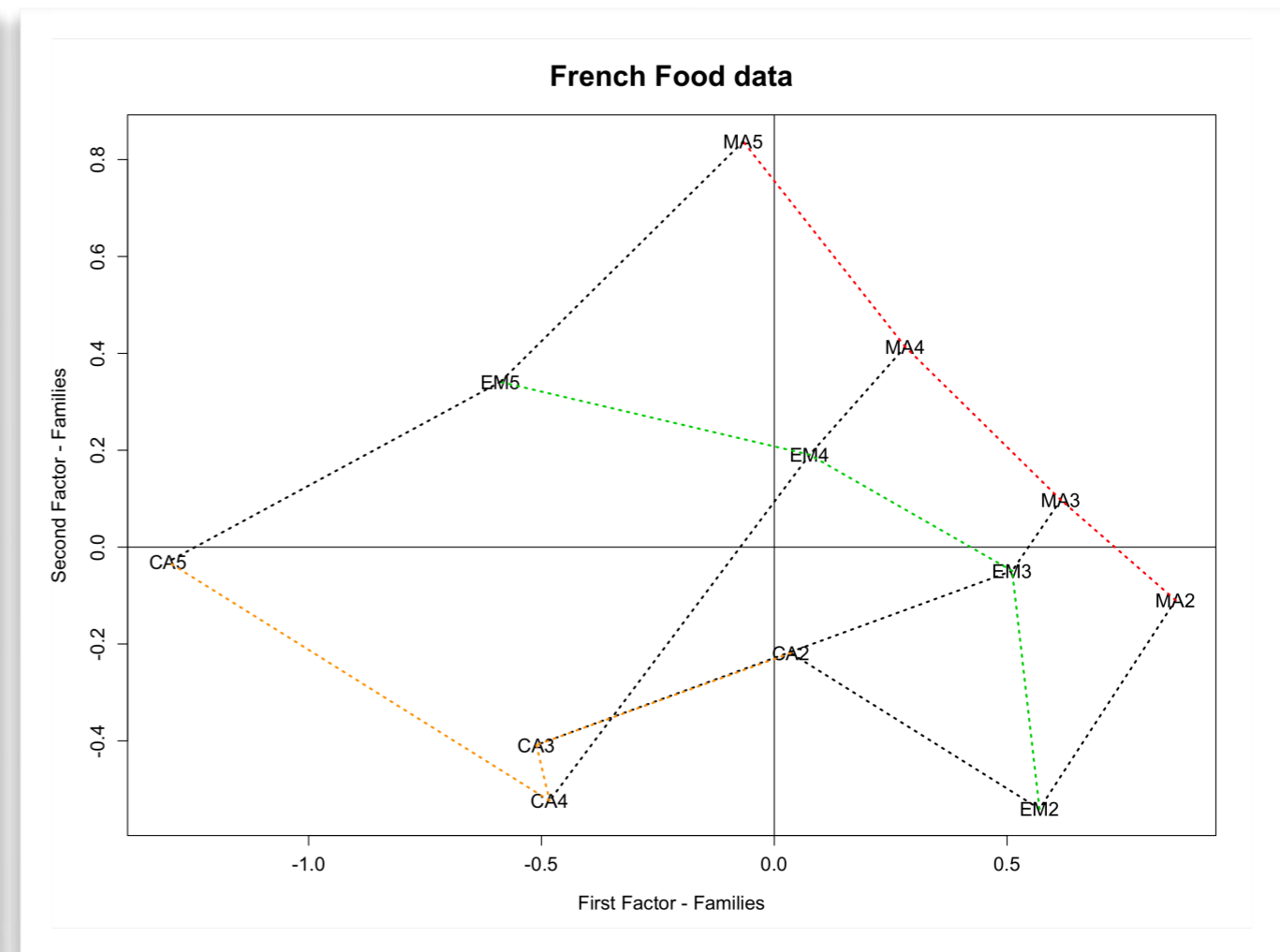
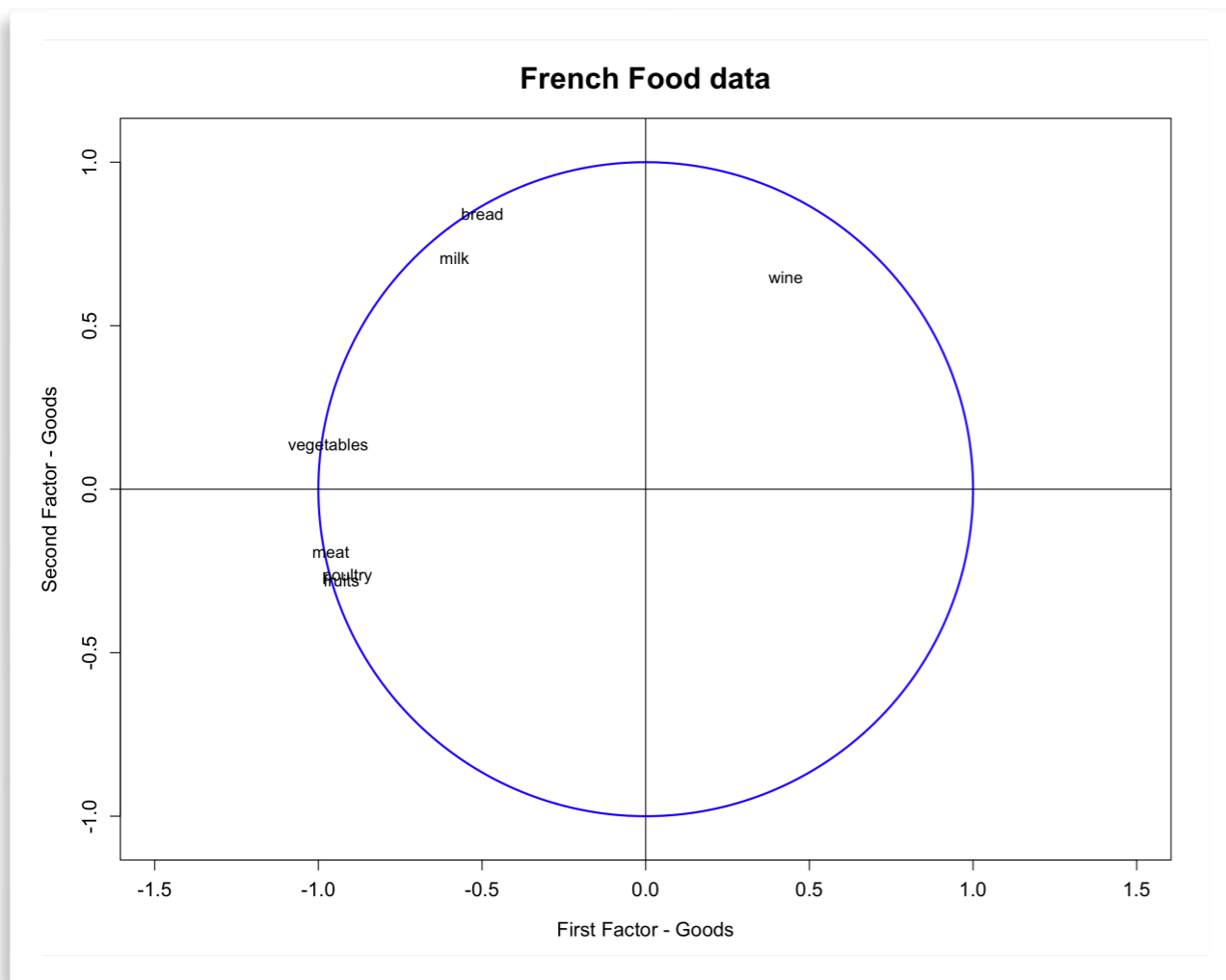


Z_2 是牛奶、面包、葡萄酒因子.

Z_1 是与蔬菜、肉类、家禽和水果相关的因子.

主成分作为一种因子方法 (Principal Components as a Factorial Method)

● 例: 法国食品支出数据集.



- ▶ 综合考虑这两张图，就能说明是哪些类型的支出导致了食品支出的相似性。
- ▶ 面包、牛奶和葡萄酒的支出，在体力劳动者和雇员之间情况相近。
- ▶ 经理家庭的特点是在蔬菜、水果、肉类和禽类上的支出较高。

Boston Housing

- 波士顿住房数据集的 NPCA.

X_1 : 人均犯罪率

X_2 : 划定用于大片住宅用地的比例

X_3 : 非零售商业用地的比例

X_4 : 查尔斯河 (与河相邻为 1, 否则为 0)

X_5 : 一氧化氮浓度

X_6 : 每套住宅的平均房间数

X_7 : 1940年之前建造的自住房屋比例

X_8 : 到波士顿五个就业中心的加权距离

X_9 : 辐射状公路可达性指数

X_{10} : 每 10000 美元房产的全额税率

X_{11} : 学生与教师的比例

X_{12} : $1000(B - 0.63)^2 I(B < 0.63)$ 其中 B 是非洲裔美国人的比例

X_{13} : 人口中较低社会地位群体的占比

X_{14} : 自有住房的中位价值 (单位: 1000 美元)

由于 X_4 是一个离散的 0 - 1 变量, 我们将其剔除.

Boston Housing

- 波士顿住房数据集的 NPCA.

```
library(MASS)
data = Boston
head(data)
```

```
> head(data)
      crim  zn   indus      nox      rm      age      dis      rad      tax  ptratio  black  lstat  medv
1 -5.064036 1.8 0.8372475 -0.6198967 1.883275 3.432567 1.408545 0.0000000 5.690359 0.4548647 3.9690 2.231591 3.178054
2 -3.600502 0.0 1.9558605 -0.7571525 1.859574 5.529585 1.602836 0.6931472 5.488938 1.2364504 3.9690 3.023243 3.072693
3 -3.601235 0.0 1.9558605 -0.7571525 1.971996 2.918119 1.602836 0.6931472 5.488938 1.2364504 3.9283 2.007486 3.546740
4 -3.430523 0.0 0.7793249 -0.7808861 1.945624 1.419592 1.802073 1.0986123 5.402677 1.7722408 3.9463 1.714643 3.508556
5 -2.672924 0.0 0.7793249 -0.7808861 1.966693 2.162710 1.802073 1.0986123 5.402677 1.7722408 3.9690 2.308679 3.589059
6 -3.511570 0.0 0.7793249 -0.7808861 1.860975 2.639947 1.802073 1.0986123 5.402677 1.7722408 3.9412 2.282542 3.356897
```

Boston Housing

- 波士顿住房数据集的 NPCA.

- ▶ 由于大多数变量呈现出不对称性, 因此进行以下变换:

$$\widetilde{X}_1 = \log X_1,$$

$$\widetilde{X}_8 = \log X_8$$

$$\widetilde{X}_2 = \frac{X_2}{10},$$

$$\widetilde{X}_9 = \log X_9$$

$$\widetilde{X}_3 = \log X_3,$$

$$\widetilde{X}_{10} = \log X_{10}$$

$$\widetilde{X}_4 = X_4,$$

$$\widetilde{X}_{11} = \frac{e^{0.4 \times X_{11}}}{100}$$

$$\widetilde{X}_5 = \log X_5,$$

$$\widetilde{X}_{12} = \frac{X_{12}}{100}$$

$$\widetilde{X}_6 = \log X_6,$$

$$\widetilde{X}_{13} = \sqrt{X_{13}}$$

$$\widetilde{X}_7 = \frac{X_7^{2.5}}{10000},$$

$$\widetilde{X}_{14} = \log X_{14}$$

Boston Housing

- 波士顿住房数据集的 NPCA.

```
# transform data
```

```
xt = data
xt[, 1] = log(data[, 1])
xt[, 2] = data[, 2]/10
xt[, 3] = log(data[, 3])
xt[, 5] = log(data[, 5])
xt[, 6] = log(data[, 6])
xt[, 7] = (data[, 7]^(2.5))/10000
xt[, 8] = log(data[, 8])
xt[, 9] = log(data[, 9])
xt[, 10] = log(data[, 10])
xt[, 11] = exp(0.4 * data[, 11])/1000
xt[, 12] = data[, 12]/100
xt[, 13] = sqrt(data[, 13])
xt[, 14] = log(data[, 14])
data = xt[, -4]
head(data)
```

```
> head(data)
```

	crim	zn	indus	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	-5.064036	1.8	0.8372475	-0.6198967	1.883275	3.432567	1.408545	0.0000000	5.690359	0.4548647	3.9690	2.231591	3.178054
2	-3.600502	0.0	1.9558605	-0.7571525	1.859574	5.529585	1.602836	0.6931472	5.488938	1.2364504	3.9690	3.023243	3.072693
3	-3.601235	0.0	1.9558605	-0.7571525	1.971996	2.918119	1.602836	0.6931472	5.488938	1.2364504	3.9283	2.007486	3.546740
4	-3.430523	0.0	0.7793249	-0.7808861	1.945624	1.419592	1.802073	1.0986123	5.402677	1.7722408	3.9463	1.714643	3.508556
5	-2.672924	0.0	0.7793249	-0.7808861	1.966693	2.162710	1.802073	1.0986123	5.402677	1.7722408	3.9690	2.308679	3.589059
6	-3.511570	0.0	0.7793249	-0.7808861	1.860975	2.639947	1.802073	1.0986123	5.402677	1.7722408	3.9412	2.282542	3.356897

Boston Housing

- 波士顿住房数据集的 NPCA.

```
n1 = nrow(data)
```

```
n2 = ncol(data)
```

```
# standardizes the data
```

```
x = (data - matrix(apply(data, 2, mean), n1, n2, byrow = T)) / matrix(sqrt((n1 - 1) * apply(data, 2, var)/n1), n1, n2, byrow = T)
```

```
round(head(x), digits = 4)
```

```
> round(head(x), digits = 4)
```

	crim	zn	indus	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	-1.9832	0.2848	-1.7043	-0.0490	0.4581	-0.4570	0.4091	-2.1370	-0.6087	-1.2440	0.4411	-1.2027	0.3515
2	-1.3056	-0.4877	-0.2632	-0.7309	0.2469	0.1316	0.7696	-1.3439	-1.1174	-0.6705	0.4411	-0.4000	0.0935
3	-1.3060	-0.4877	-0.2632	-0.7309	1.2488	-0.6014	0.7696	-1.3439	-1.1174	-0.6705	0.3964	-1.4299	1.2544
4	-1.2269	-0.4877	-1.7790	-0.8489	1.0138	-1.0219	1.1392	-0.8800	-1.3353	-0.2773	0.4162	-1.7269	1.1609
5	-0.8762	-0.4877	-1.7790	-0.8489	1.2015	-0.8134	1.1392	-0.8800	-1.3353	-0.2773	0.4411	-1.1245	1.3580
6	-1.2645	-0.4877	-1.7790	-0.8489	0.2594	-0.6794	1.1392	-0.8800	-1.3353	-0.2773	0.4106	-1.1510	0.7895

Boston Housing

- 波士顿住房数据集的 NPCA.

```
# spectral decomposition
```

```
eig = eigen((n1 - 1) * cov(x)/n1)
```

```
e = eig$values # eigenvalues
```

```
perc = e/sum(e) # explained variance
```

```
cum = cumsum(e)/sum(e) # cumulative explained percentages
```

```
evpc = data.frame(eigenvalues = e, percent = perc, cumpercent = cum)
```

```
round(evpc, digits = 4)
```

```
> round(evpc, digits = 4)
  eigenvalues percent cumpercent
1      7.2852  0.5604    0.5604
2      1.3517  0.1040    0.6644
3      1.1266  0.0867    0.7510
4      0.7802  0.0600    0.8111
5      0.6359  0.0489    0.8600
6      0.5290  0.0407    0.9007
7      0.3397  0.0261    0.9268
8      0.2628  0.0202    0.9470
9      0.1936  0.0149    0.9619
10     0.1547  0.0119    0.9738
11     0.1405  0.0108    0.9846
12     0.1100  0.0085    0.9931
13     0.0900  0.0069    1.0000
```

第一主成分解释了总方差的56%.

前三个主成分加起来能解释超过 75% 的方差.

这些结果表明, 研究 2 个, 最多 3 个主成分就足够了.

Boston Housing

- 波士顿住房数据集的 NPCA.

v = eig\$vectors

round(v, digits = 4) # eigenvectors

```
> round(v, digits = 4) # eigenvectors
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
[1,]	0.3363	0.1933	0.1373	0.0459	0.1084	-0.0882	0.0861	0.2047	0.0154	-0.2098	0.3262	-0.2388	0.7430
[2,]	-0.2371	-0.0251	0.4765	-0.0981	0.4405	0.5190	-0.3205	-0.1946	0.0851	-0.2847	0.0774	-0.0877	-0.0244
[3,]	0.3179	0.0352	-0.1738	0.0753	-0.0466	-0.2477	-0.1635	-0.7175	0.4110	-0.2527	0.1132	-0.0663	-0.0824
[4,]	0.3237	0.2057	-0.1677	-0.1356	0.1037	0.1222	-0.1289	-0.0560	-0.5291	-0.1539	0.3923	0.5251	-0.1669
[5,]	-0.1891	0.6052	0.0819	0.0961	-0.2884	0.3207	0.5312	-0.2899	-0.0475	-0.1109	-0.1150	0.0290	0.0350
[6,]	0.2964	0.1338	-0.2778	-0.1141	0.0235	0.5206	0.0298	0.2278	0.5005	0.3668	0.2521	-0.0748	-0.1599
[7,]	-0.3060	-0.2498	0.2810	0.1206	-0.0558	-0.1563	0.2993	-0.0452	0.2982	0.1296	0.5670	0.4502	0.0376
[8,]	0.2790	0.2457	0.3584	0.2227	0.2489	-0.2888	0.2165	0.3307	0.1270	-0.2297	-0.0102	-0.0967	-0.5511
[9,]	0.3006	0.1415	0.3459	0.1959	0.2685	-0.0474	-0.0666	-0.2473	-0.0411	0.6317	-0.3017	0.2612	0.1778
[10,]	0.2102	-0.2294	0.1411	0.6836	-0.4683	0.3196	-0.2405	0.1023	-0.0565	-0.1361	-0.0158	0.0646	-0.0073
[11,]	-0.1818	-0.0895	-0.4870	0.5618	0.5757	0.0831	0.2277	-0.0606	-0.0827	-0.0349	0.0168	-0.0374	0.0495
[12,]	0.2963	-0.3658	-0.0236	-0.2109	0.0963	0.1796	0.3610	0.0659	0.2089	-0.3542	-0.4397	0.4157	0.1481
[13,]	-0.2729	0.4438	-0.1646	0.1096	0.0181	-0.1465	-0.4272	0.2746	0.3613	-0.1633	-0.1939	0.4360	0.1573

Boston Housing

- 波士顿住房数据集的 NPCA.

```
# principal components - projections of the individuals
```

```
xv = as.matrix(x) %*% v
```

```
round(head(xv), digits = 4)
```

```
> round(head(xv), digits = 4)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
1	-3.2125	-0.1160	-0.9491	-1.0005	-0.1491	0.6016	-0.2317	-0.1843	-1.1462	1.2367	-0.0798	0.3718	-0.2497
2	-1.9628	-0.6111	-1.2270	-0.4652	-0.6344	0.0433	0.4074	-0.2832	0.3274	0.5174	0.1874	-0.0152	-0.2738
3	-2.9835	0.7931	-1.0865	0.0342	-1.0445	-0.3756	0.0398	-0.4871	0.1210	0.3143	0.1144	0.1485	-0.0939
4	-3.5896	0.4982	-0.4317	0.4026	-1.0863	-0.4097	0.2406	0.7798	-0.5793	0.4407	0.2567	0.1180	-0.1727
5	-3.3252	0.5725	-0.4849	0.3214	-1.0216	-0.1905	0.5156	0.9371	-0.2835	0.1764	0.0995	0.3595	0.1826
6	-3.0851	-0.2948	-0.5435	0.1239	-0.8193	-0.3126	0.2121	1.0052	-0.3861	0.5148	0.2363	0.1571	-0.2552

```
# correlations of the first 3 PC's with the original variables
```

```
corr = cor(x, xv)[, 1:3]
```

```
colnames(corr) = c('PC1', 'PC2', 'PC3')
```

```
round(corr, digits = 4)
```

```
> round(corr, digits = 4)
```

	PC1	PC2	PC3
crim	0.9076	0.2247	0.1457
zn	-0.6399	-0.0292	0.5058
indus	0.8580	0.0409	-0.1845
nox	0.8737	0.2391	-0.1780
rm	-0.5104	0.7037	0.0869
age	0.7999	0.1556	-0.2949
dis	-0.8259	-0.2904	0.2982
rad	0.7531	0.2857	0.3804
tax	0.8114	0.1645	0.3672
ptratio	0.5674	-0.2667	0.1498
black	-0.4906	-0.1041	-0.5170
lstat	0.7996	-0.4253	-0.0251
medv	-0.7366	0.5160	-0.1747

```
# proportions of variables explained by first 3 PC's
```

```
cumr2 = apply(corr^2, 1, sum)
```

```
round(as.matrix(cumr2), digits = 4)
```

```
> round(as.matrix(cumr2), digits = 4)
```

	[,1]
crim	0.8955
zn	0.6661
indus	0.7719
nox	0.8521
rm	0.7632
age	0.7510
dis	0.8554
rad	0.7935
tax	0.8203
ptratio	0.4155
black	0.5188
lstat	0.8209
medv	0.8394

Boston Housing

- 波士顿住房数据集的 NPCA.

```
x1 = as.matrix(x - matrix(mean(as.matrix(x)), nrow(x), ncol(x), byrow = T))
```

```
# projections of individuals into PCs
```

```
r1 = x1 %*% v
```

```
# correlation matrix of PCs and original variables
```

```
r = cor(cbind(r1, x))
```

```
round(r, digits = 2)
```

```
> round(r, digits = 2)
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	crim	zn	indus	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.91	-0.64	0.86	0.87	-0.51	0.80	-0.83	0.75	0.81	0.57	-0.49	0.80	-0.74
2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	-0.03	0.04	0.24	0.70	0.16	-0.29	0.29	0.16	-0.27	-0.10	-0.43	0.52
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.51	-0.18	-0.18	0.09	-0.29	0.30	0.38	0.37	0.15	-0.52	-0.03	-0.17
4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	-0.09	0.07	-0.12	0.08	-0.10	0.11	0.20	0.17	0.60	0.50	-0.19	0.10
5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.35	-0.04	0.08	-0.23	0.02	-0.04	0.20	0.21	-0.37	0.46	0.08	0.01
6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.06	0.38	-0.18	0.09	0.23	0.38	-0.11	-0.21	-0.03	0.23	0.06	0.13	-0.11
7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	-0.19	-0.10	-0.08	0.31	0.02	0.17	0.13	-0.04	-0.14	0.13	0.21	-0.25
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.10	-0.10	-0.37	-0.03	-0.15	0.12	-0.02	0.17	-0.13	0.05	-0.03	0.03	0.14
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.01	0.04	0.18	-0.23	-0.02	0.22	0.13	0.06	-0.02	-0.02	-0.04	0.09	0.16
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	-0.08	-0.11	-0.10	-0.06	-0.04	0.14	0.05	-0.09	0.25	-0.05	-0.01	-0.14	-0.06
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.12	0.03	0.04	0.15	-0.04	0.09	0.21	0.00	-0.11	-0.01	0.01	-0.16	-0.07
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	-0.08	-0.03	-0.02	0.17	0.01	-0.02	0.15	-0.03	0.09	0.02	-0.01	0.14	0.14
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.22	-0.01	-0.02	-0.05	0.01	-0.05	0.01	-0.17	0.05	0.00	0.01	0.04	0.05
crim	0.91	0.22	0.15	0.04	0.09	-0.06	0.05	0.10	0.01	-0.08	0.12	-0.08	0.22	1.00	-0.52	0.74	0.81	-0.32	0.70	-0.74	0.84	0.81	0.45	-0.48	0.62	-0.57
zn	-0.64	-0.03	0.51	-0.09	0.35	0.38	-0.19	-0.10	0.04	-0.11	0.03	-0.03	-0.01	-0.52	1.00	-0.66	-0.57	0.31	-0.53	0.59	-0.35	-0.31	-0.35	0.18	-0.45	0.36
indus	0.86	0.04	-0.18	0.07	-0.04	-0.18	-0.10	-0.37	0.18	-0.10	0.04	-0.02	-0.02	0.74	-0.66	1.00	0.75	-0.43	0.66	-0.73	0.58	0.66	0.45	-0.33	0.62	-0.55
nox	0.87	0.24	-0.18	-0.12	0.08	0.09	-0.08	-0.03	-0.23	-0.06	0.15	0.17	-0.05	0.81	-0.57	0.75	1.00	-0.32	0.78	-0.86	0.61	0.67	0.34	-0.38	0.61	-0.52
rm	-0.51	0.70	0.09	0.08	-0.23	0.23	0.31	-0.15	-0.02	-0.04	-0.04	0.01	0.01	-0.32	0.31	-0.43	-0.32	1.00	-0.28	0.28	-0.21	-0.31	-0.32	0.13	-0.64	0.61
age	0.80	0.16	-0.29	-0.10	0.02	0.38	0.02	0.12	0.22	0.14	0.09	-0.02	-0.05	0.70	-0.53	0.66	0.78	-0.28	1.00	-0.80	0.47	0.54	0.38	-0.29	0.64	-0.48
dis	-0.83	-0.29	0.30	0.11	-0.04	-0.11	0.17	-0.02	0.13	0.05	0.21	0.15	0.01	-0.74	0.59	-0.73	-0.86	0.28	-0.80	1.00	-0.54	-0.60	-0.32	0.32	-0.56	0.41
rad	0.75	0.29	0.38	0.20	0.20	-0.21	0.13	0.17	0.06	-0.09	0.00	-0.03	-0.17	0.84	-0.35	0.58	0.61	-0.21	0.47	-0.54	1.00	0.82	0.40	-0.41	0.46	-0.43
tax	0.81	0.16	0.37	0.17	0.21	-0.03	-0.04	-0.13	-0.02	0.25	-0.11	0.09	0.05	0.81	-0.31	0.66	0.67	-0.31	0.54	-0.60	0.82	1.00	0.48	-0.43	0.53	-0.56
ptratio	0.57	-0.27	0.15	0.60	-0.37	0.23	-0.14	0.05	-0.02	-0.05	-0.01	0.02	0.00	0.45	-0.35	0.45	0.34	-0.32	0.38	-0.32	0.40	0.48	1.00	-0.20	0.43	-0.51
black	-0.49	-0.10	-0.52	0.50	0.46	0.06	0.13	-0.03	-0.04	-0.01	0.01	-0.01	0.01	-0.48	0.18	-0.33	-0.38	0.13	-0.29	0.32	-0.41	-0.43	-0.20	1.00	-0.36	0.40
lstat	0.80	-0.43	-0.03	-0.19	0.08	0.13	0.21	0.03	0.09	-0.14	-0.16	0.14	0.04	0.62	-0.45	0.62	0.61	-0.64	0.64	-0.56	0.46	0.53	0.43	-0.36	1.00	-0.83
medv	-0.74	0.52	-0.17	0.10	0.01	-0.11	-0.25	0.14	0.16	-0.06	-0.07	0.14	0.05	-0.57	0.36	-0.55	-0.52	0.61	-0.48	0.41	-0.43	-0.56	-0.51	0.40	-0.83	1.00

Boston Housing

- 波士顿住房数据集的 NPCA.

```

graphics.off()
par(mfrow = c(2, 2))
ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab = "Second PC",
     main = "Boston Housing", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd = 2, asp = 1)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X5", "X6", "X7", "X8", "X9", "X10", "X11", "X12", "X13", "X14")
text(corr[, 1], corr[, 2], label)
  
```

```

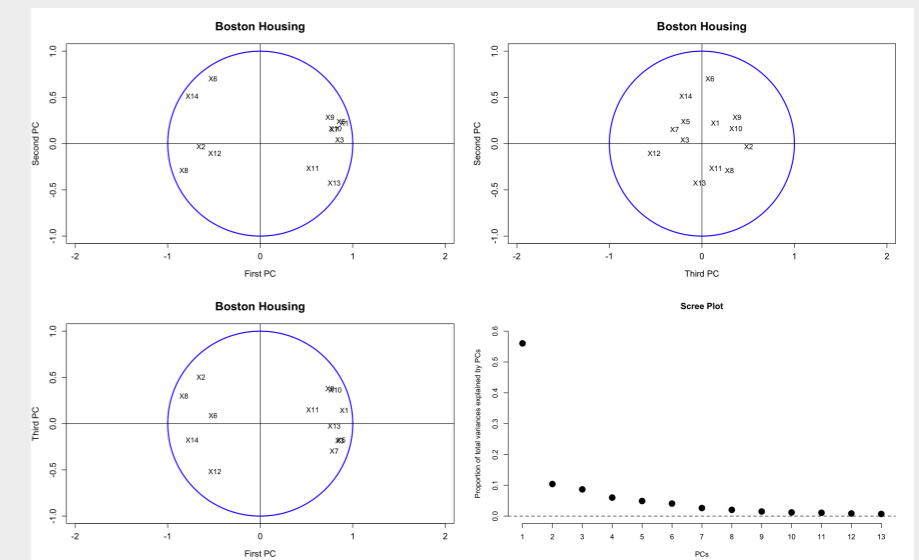
ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "Third PC", ylab = "Second PC",
     main = "Boston Housing", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd = 2, asp = 1)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X5", "X6", "X7", "X8", "X9", "X10", "X11", "X12", "X13", "X14")
text(corr[, 3], corr[, 2], label)
  
```

```

ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab = "Third PC",
     main = "Boston Housing", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd = 2, asp = 1)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X5", "X6", "X7", "X8", "X9", "X10", "X11", "X12", "X13", "X14")
text(corr[, 1], corr[, 3], label)
  
```

```

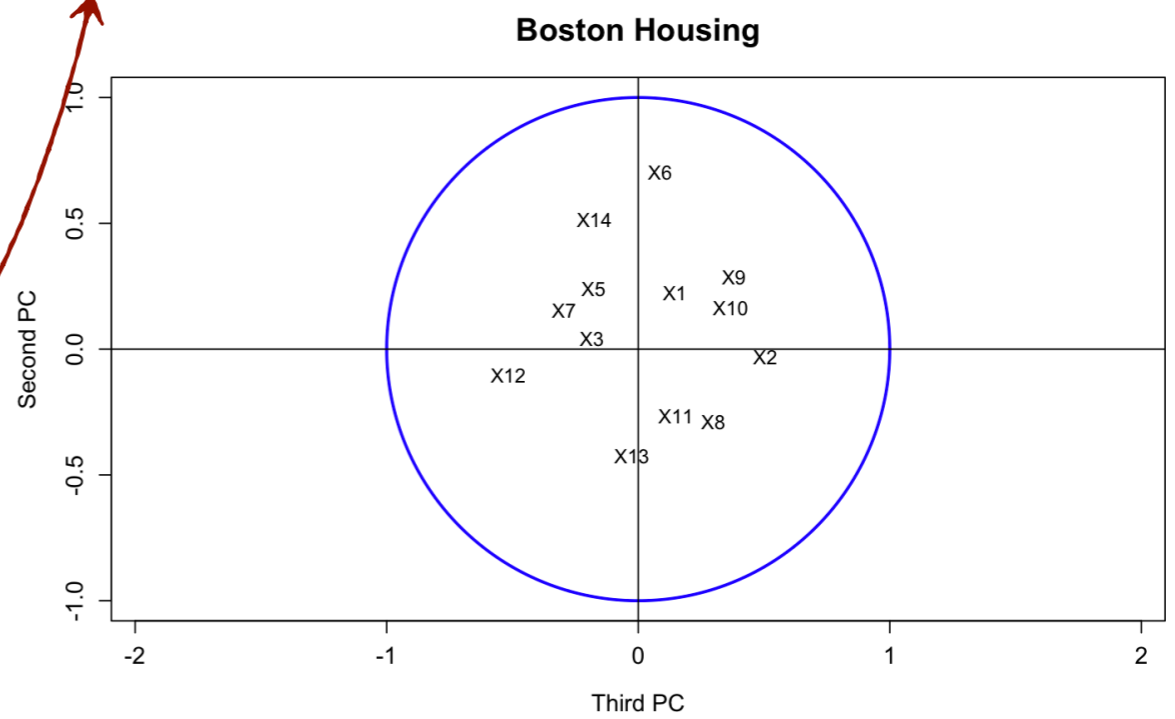
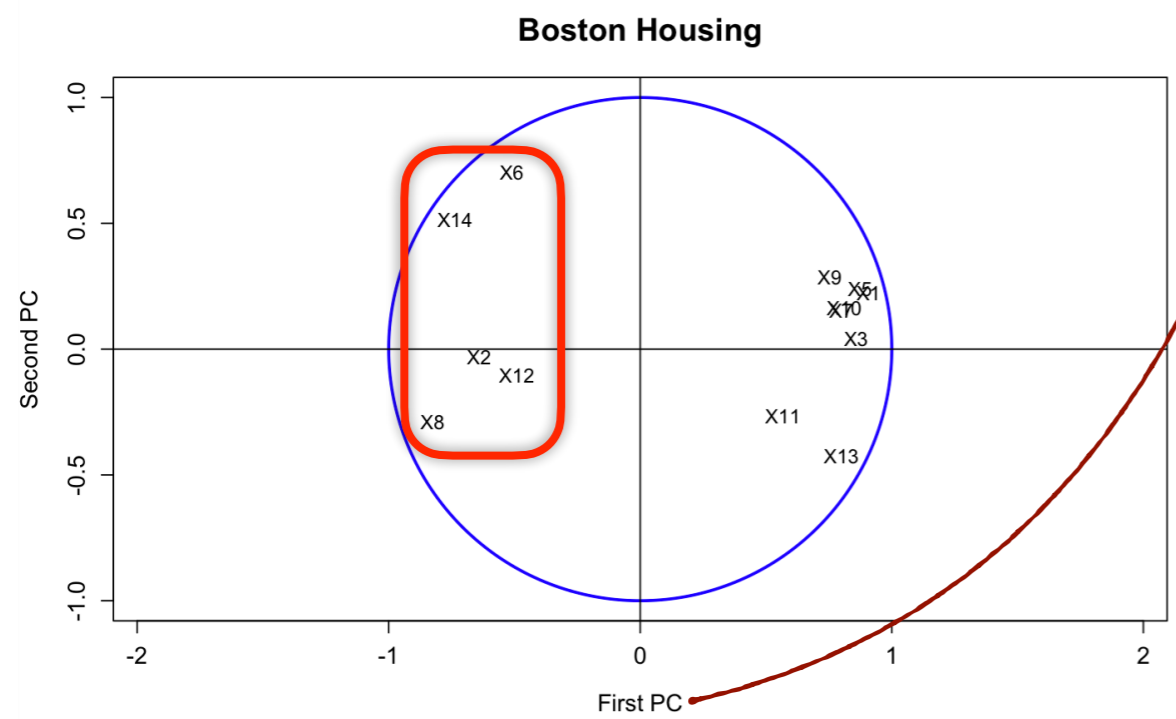
plot(perc, xlab = 'PCs', ylab = 'Proportion of total variances explained by PCs', main = 'Scree Plot', pch = 16,
     cex = 2, axes = FALSE, ylim = c(0, 0.6))
axis(1, at = 1:13)
axis(2, at = c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6))
abline(h=0, lty=2)
  
```



Boston Housing

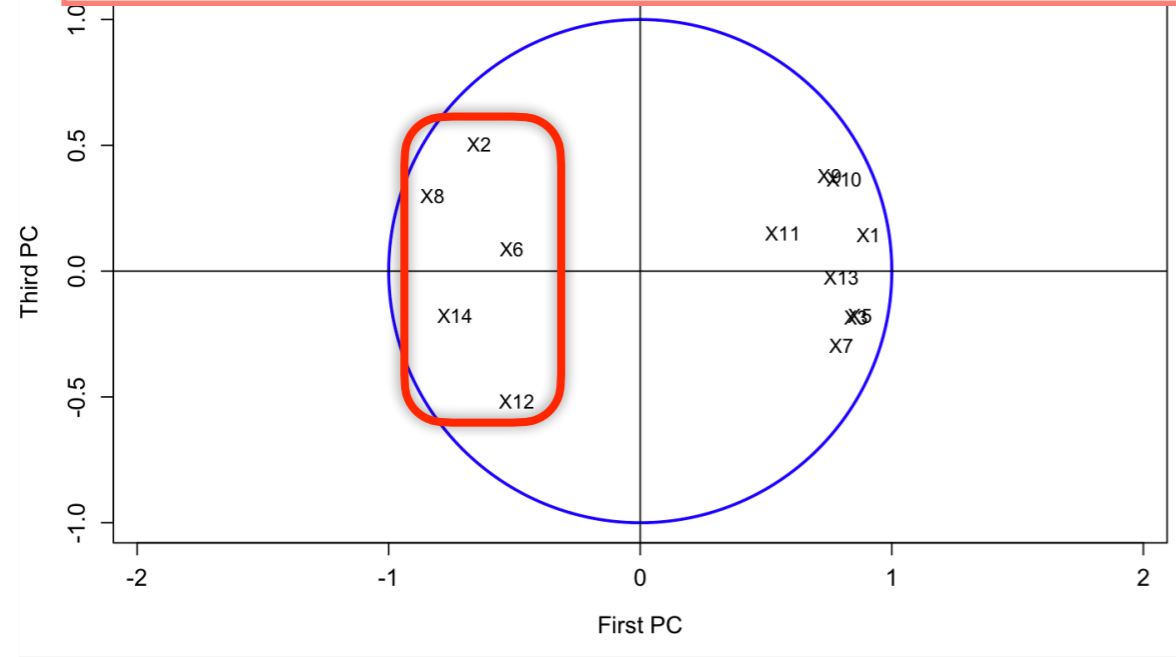
● 波士顿住房数据集的 NPCA.

生活质量与住房指标.

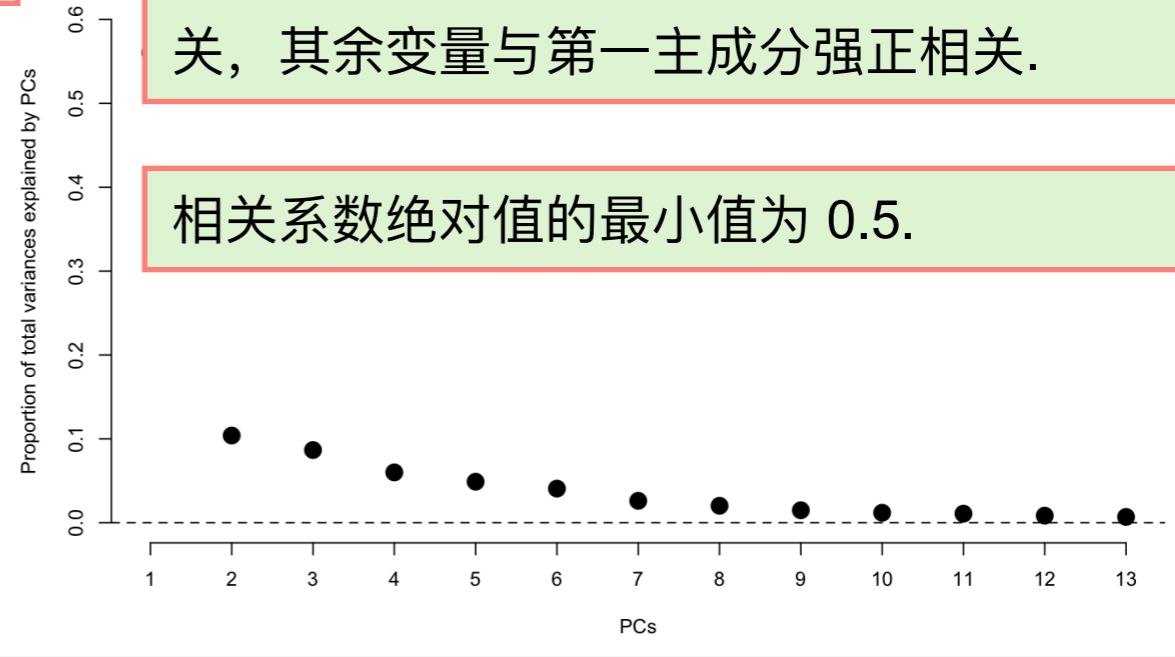


与第一主成分的相关性呈现出非常清晰的模式.

变量 $X_2, X_6, X_8, X_{12}, X_{14}$ 与第一主成分强负相关, 其余变量与第一主成分强正相关.

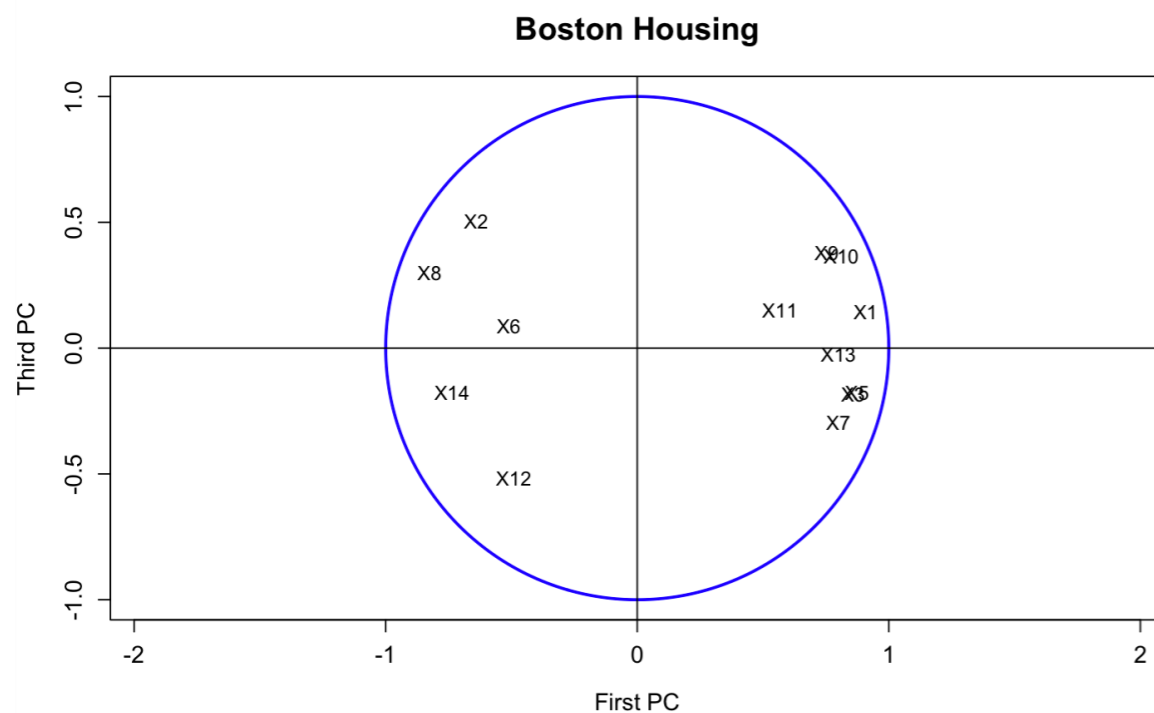
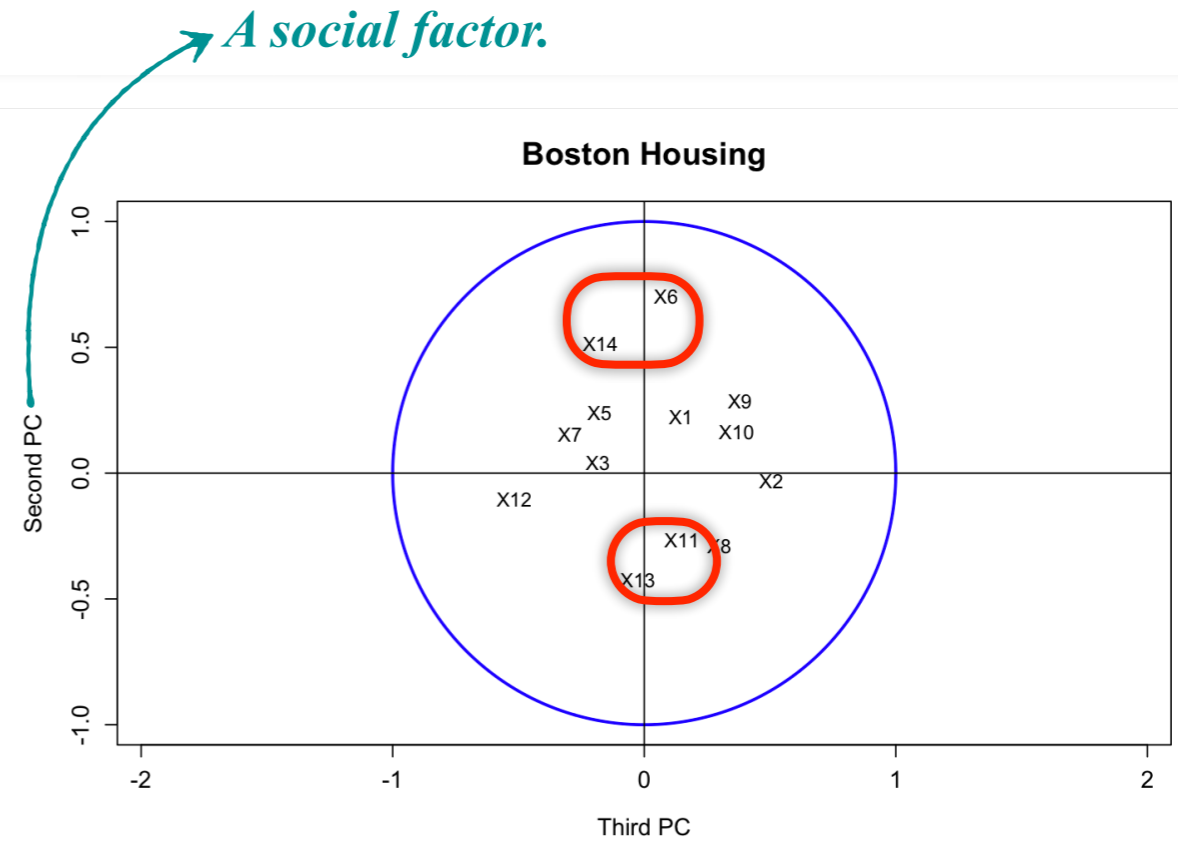
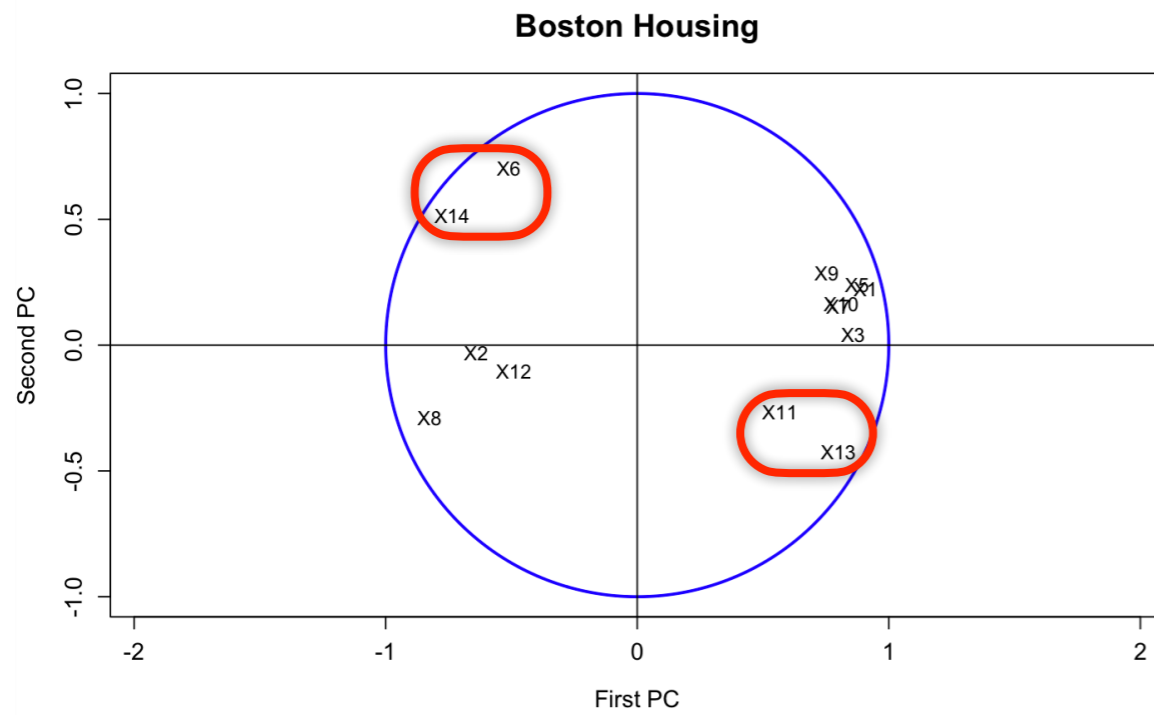


相关系数绝对值的最小值为 0.5.



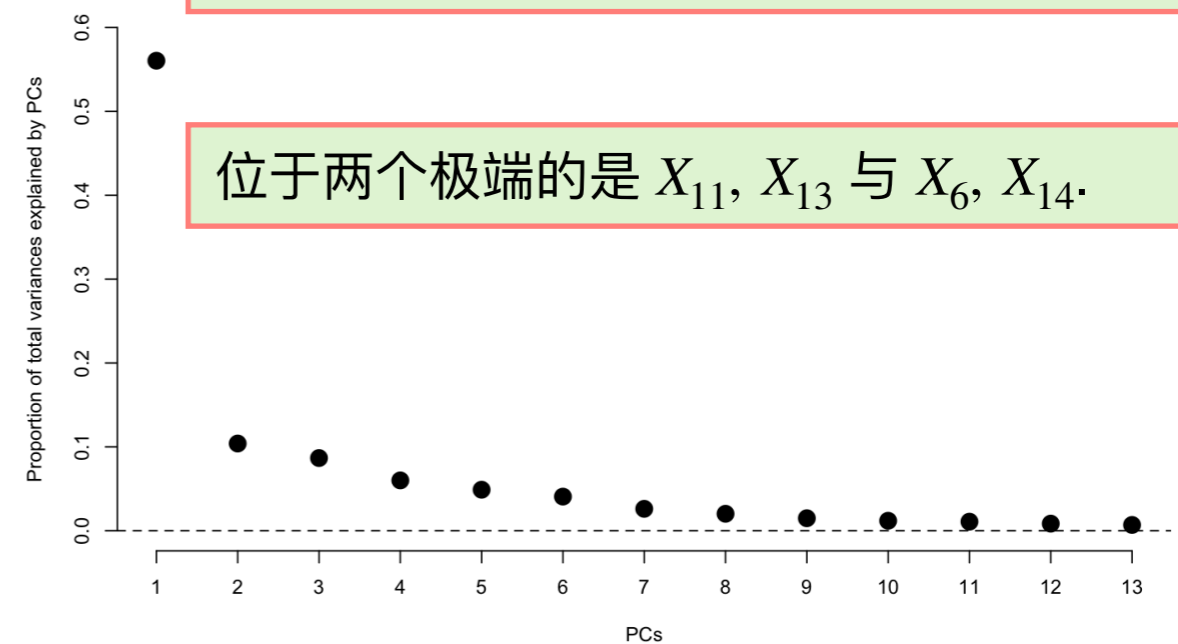
Boston Housing

● 波士顿住房数据集的 NPCA.



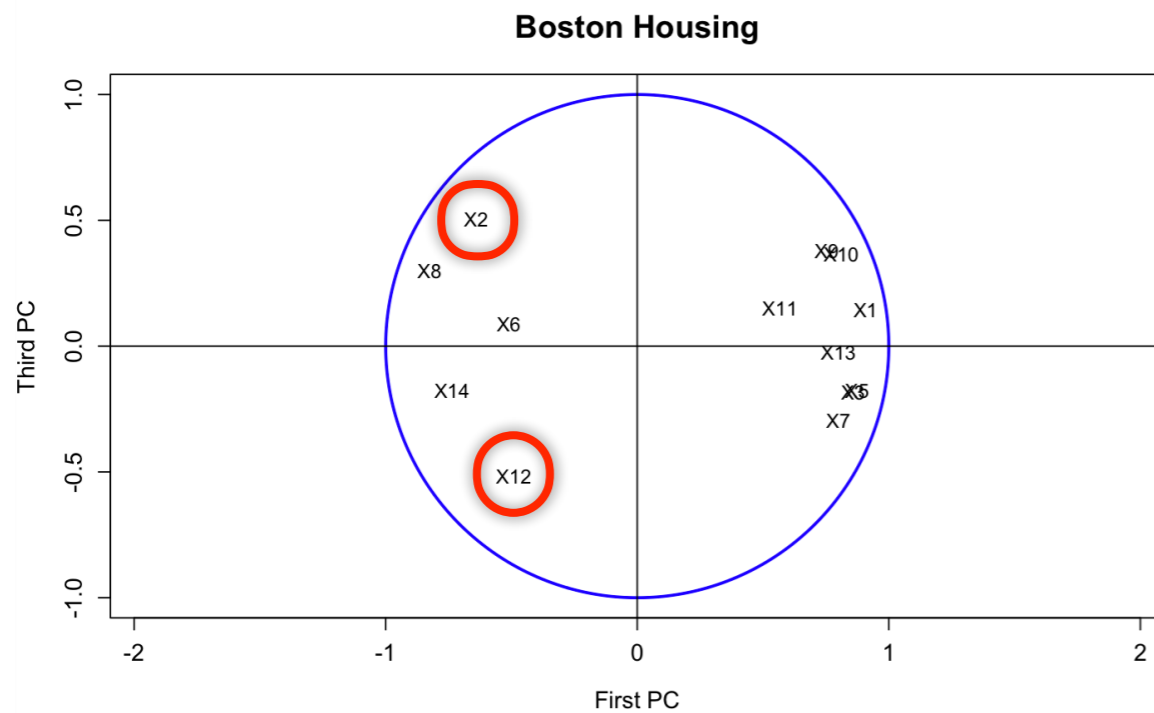
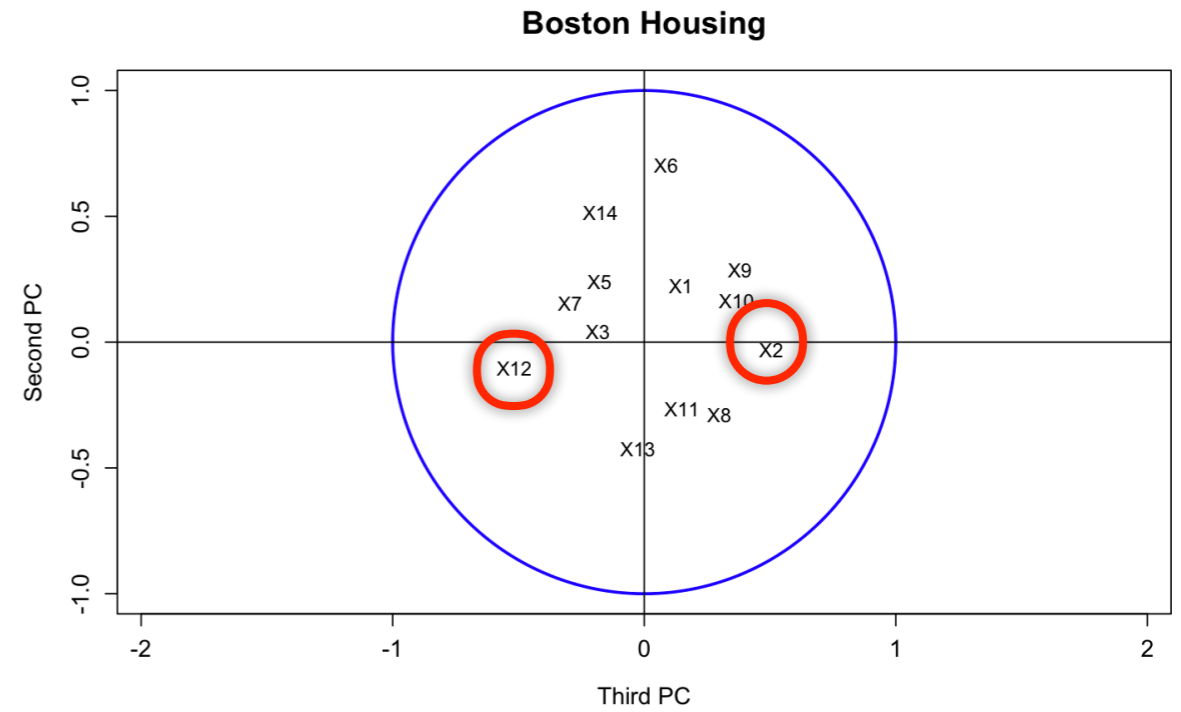
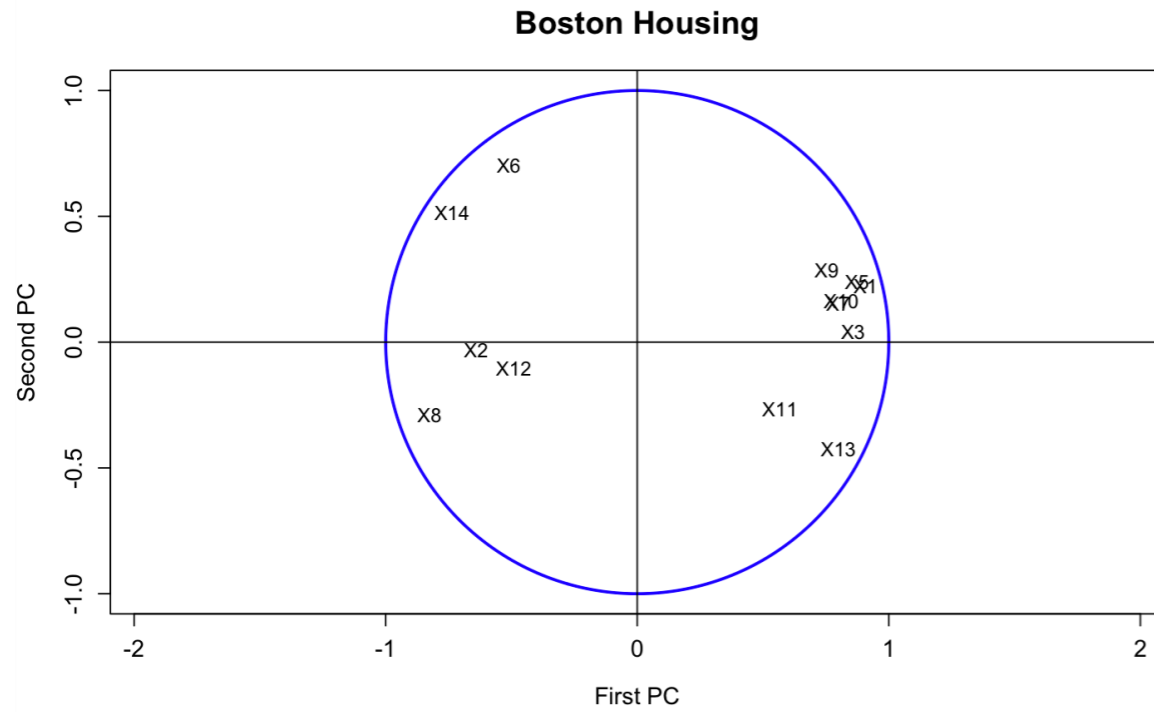
第二主成分仅解释了总方差的 10%.

位于两个极端的是 X_{11} , X_{13} 与 X_6 , X_{14} .

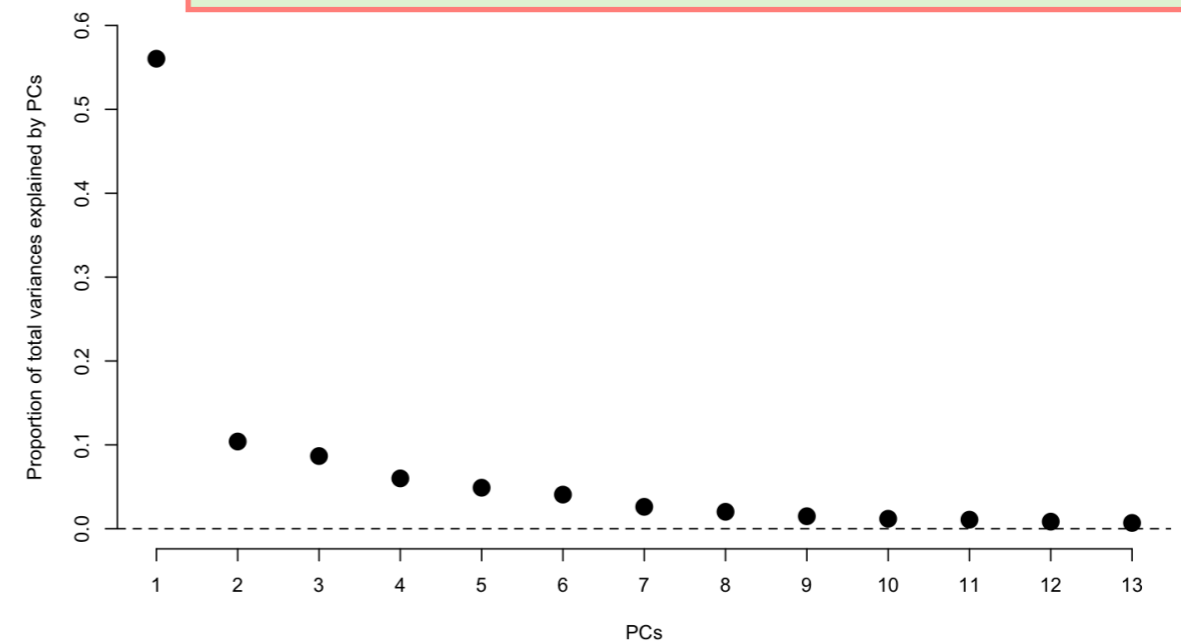


Boston Housing

● 波士顿住房数据集的 NPCA.



第三主成分主要由 X_2 和 X_{12} 之间的差异主导.



Boston Housing

- 波士顿住房数据集的 NPCA.
 - ▶ 观测数据投影到前两个主成分上的图形.

```
# price x[, 13] greater or less than average, 15 is >= average, 17 is < average, stored in h1
```

```
h = x[, 13]
```

```
h[h >= mean(h)] = 15
```

```
h1 = h
```

```
h1[h1 < mean(x[, 13])] = 17
```

```
hr = h1
```

```
# mark more expensive houses with red square
```

```
hr[hr == 15] = "red"
```

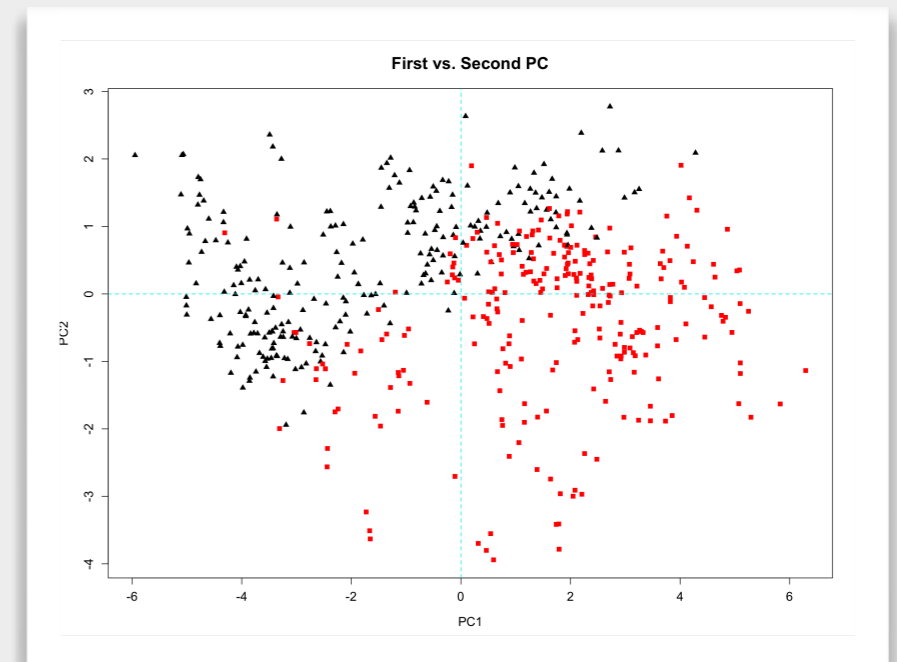
```
# mark cheaper houses with black triangle
```

```
hr[hr == 17] = "black"
```

```
# correction for signs of eigenvectors
```

```
xv = xv * (-1)
```

```
plot(xv[, 1], xv[, 2], pch = h1, col = hr, xlab = "PC1", ylab = "PC2", main = "First vs. Second PC",  
     cex.axis = 1.2, cex.lab = 1.2, cex.main = 1.6)
```

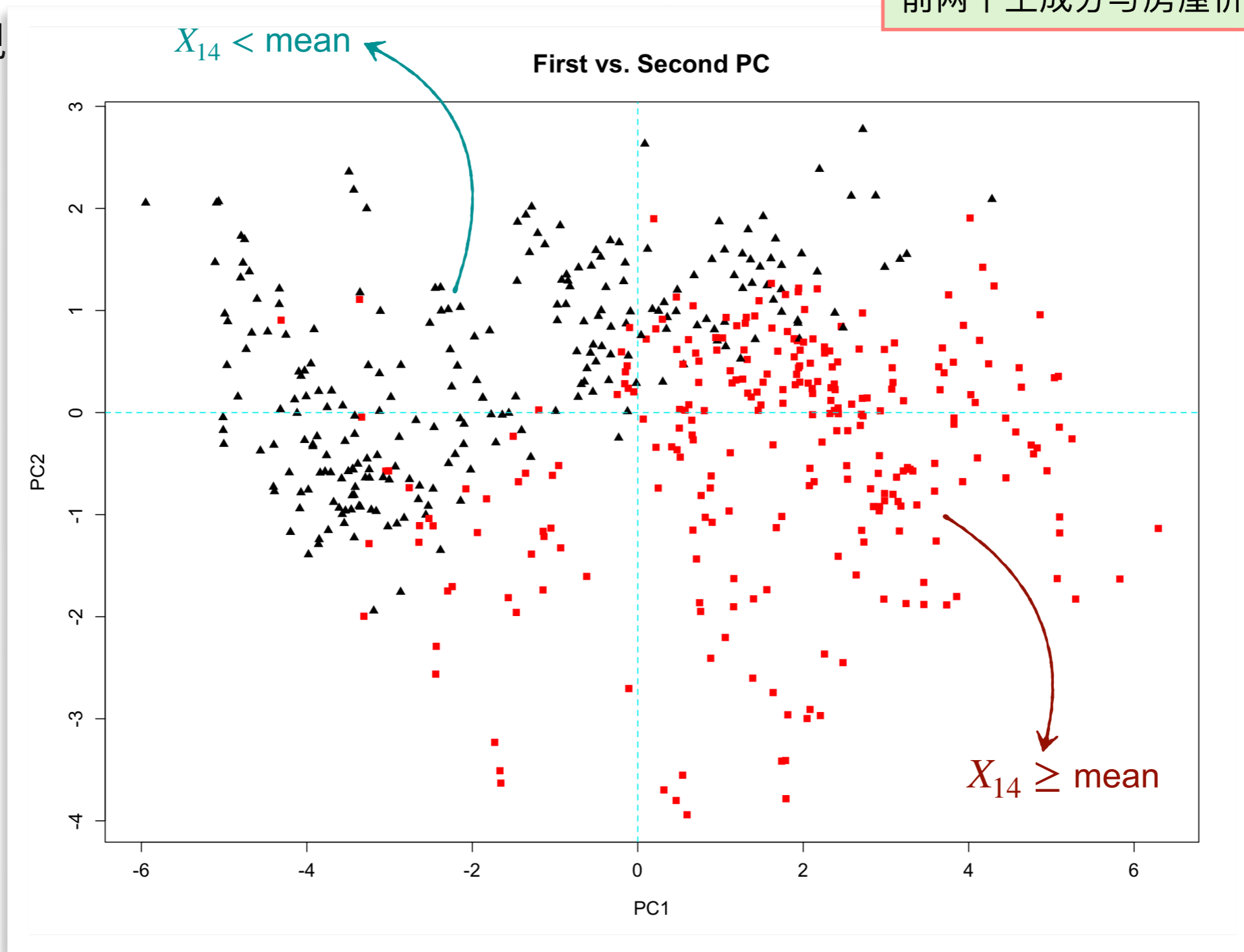


Boston Housing

- 波士顿住房数据集的 NPCA.

前两个主成分与房屋价值相关.

▶ 观



Boston Housing

- 波士顿住房数据集的 NPCA.
 - ▶ 观测数据投影到前两个主成分上的图形.

```
# houses close to the Charles River are indicated with red squares
```

```
col = xt[, 4]
```

```
col[col == 1] = "red"
```

```
col[col == 0] = "black"
```

```
h2 = xt[, 4]
```

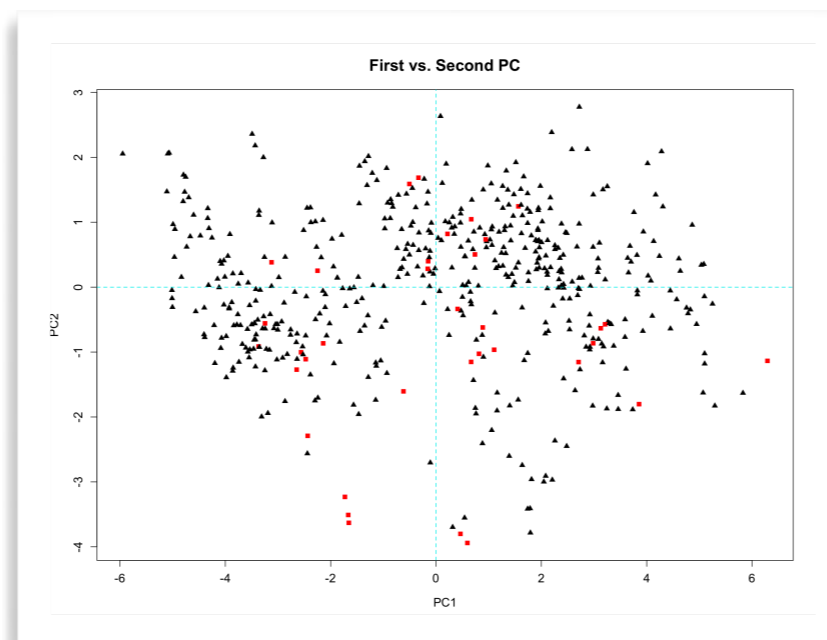
```
h2[h2 == 1] = 15
```

```
h2[h2 == 0] = 17
```

```
graphics.off()
```

```
plot(xv[, 1], xv[, 2], pch = h2, col = col, xlab = "PC1", ylab = "PC2", main = "First vs. Second PC",  
     cex.axis = 1.2, cex.lab = 1.2, cex.main = 1.6)
```

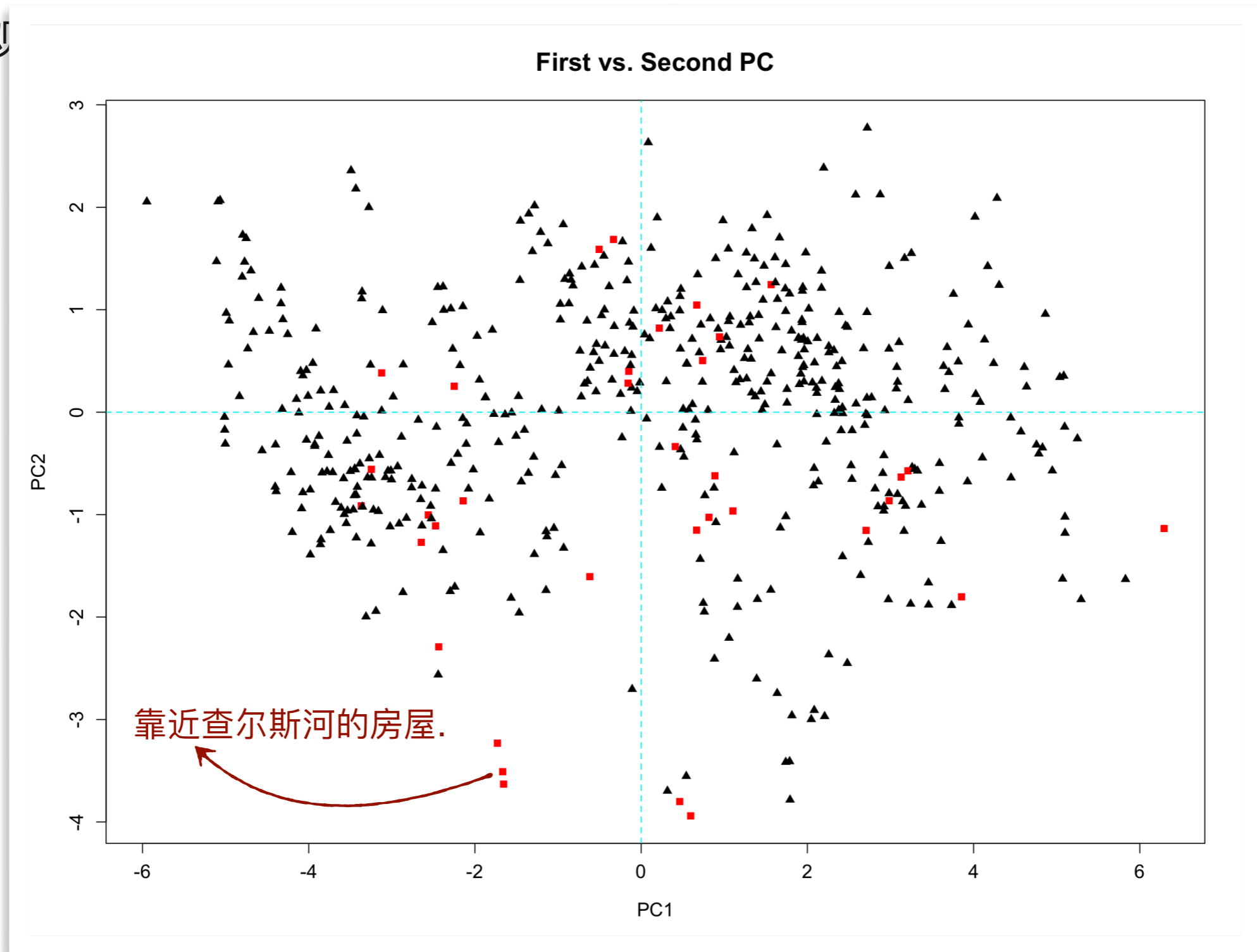
```
abline(h=0, v=0, lty=2, col = 'cyan2')
```



Boston Housing

- 波士顿住房数据集的 NPCA.

▶ 观



79 家美国公司

- 考虑 79 家美国公司的数据集.

```

rm(list = ls(all = TRUE)) # clear all variables
graphics.off()
setwd("~/Desktop/2025_Multivariate Statistical Analysis/R_Code/Data")
library(readr)
x = read_table("uscomp2.txt", col_name = FALSE)
colnames(x) = c("Company", "A", "S", "MV", "P", "CF", "E", "Sector")
attach(x)
Sector = as.character(Sector)
Sector[1:2] = "H"
Sector[3:17] = "E"
Sector[18:34] = "F"
Sector[35:42] = "H"
Sector[43:52] = "M"
Sector[53:63] = "*"
Sector[64:73] = "R"
Sector[74:79] = "*"
detach(x)
head(x)
  
```

	assets (X_1)		market value (X_3)		cash flow (X_5)		
Company	A	S	MV	P	CF	E	Sector
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1 BellAtlantic	19788	9084	10636	1093.	2577.	79.4	Communication
2 ContinentalTelecom	5074	2557	1892	240.	578.	21.9	Communication
3 AmericanElectricPower	13621	4848	4572	485	899.	23.4	Energy
4 BrooklynUnionGas	1117	1038	478	59.7	91.7	3.8	Energy
5 CentralIllinoisPublicService	1633	701	679	74.3	136.	2.8	Energy
6 ClevelandElectricIlluminating	5651	1254	2002	311.	408.	6.2	Energy

sales (X_2) profits (X_4) employees (X_6)

79 家美国公司

- 考虑 79 家美国公司的数据集。
 - ▶ 通过减去均值并除以标准差，首先对数据进行标准化处理。

```
# standardizes the data
```

```
x = as.matrix(x[, 2:7])
```

```
n = nrow(x) # number of rows
```

```
x = (x - matrix(apply(x, 2, mean), n, 6, byrow = T))/matrix(sqrt((n - 1) * apply(x, 2, var)/n), n, 6, byrow = T)
```

```
round(head(x), digits = 4)
```

```
> round(head(x), digits = 4)
      A      S      MV      P      CF      E
[1,] 1.5219 0.7041 0.6558 1.1151 1.8164 0.6522
[2,] -0.0952 -0.2327 -0.1227 0.0380 0.1481 -0.2449
[3,] 0.8441 0.0961 0.1159 0.3475 0.4157 -0.2215
[4,] -0.5301 -0.4507 -0.2486 -0.1896 -0.2582 -0.5273
[5,] -0.4734 -0.4991 -0.2307 -0.1712 -0.2213 -0.5429
[6,] -0.0318 -0.4197 -0.1129 0.1274 0.0058 -0.4899
```

79 家美国公司

- 考虑 79 家美国公司的数据集.
 - ▶ 计算特征值和对应的特征向量.

```
# spectral decomposition
```

```
eig = eigen((n - 1) * cov(x)/n)
```

```
e = eig$values
```

```
round(as.matrix(e), digits = 3)
```

```
v = eig$vectors
```

```
round(v, digits = 3)
```

eigenvalues

```
> round(as.matrix(e), digits = 3)
      [,1]
[1,] 5.039
[2,] 0.517
[3,] 0.359
[4,] 0.050
[5,] 0.029
[6,] 0.007
```

corresponding eigenvectors

```
> round(v, digits = 3)
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,] -0.340 0.849 -0.339 -0.205 -0.077 0.006
[2,] -0.423 0.170 0.379 0.783 0.006 0.186
[3,] -0.434 -0.190 -0.192 -0.071 0.844 -0.149
[4,] -0.420 -0.364 -0.324 -0.156 -0.261 0.703
[5,] -0.428 -0.285 -0.267 0.121 -0.452 -0.667
[6,] -0.397 -0.010 0.726 -0.548 -0.098 -0.065
```

79 家美国公司

- 考虑 79 家美国公司的数据集。
 - ▶ 前两个主成分的图形表示如下

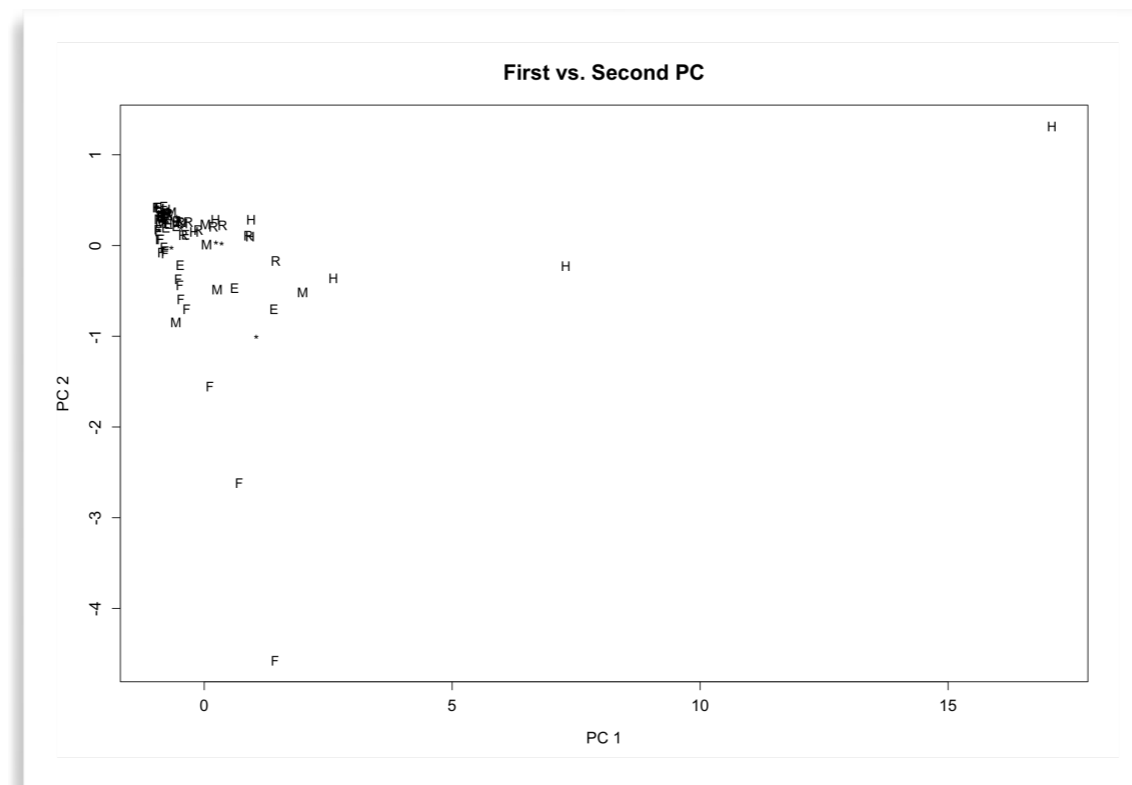
```
# principal components
```

```
x = as.matrix(x) %*% v
```

```
graphics.off()
```

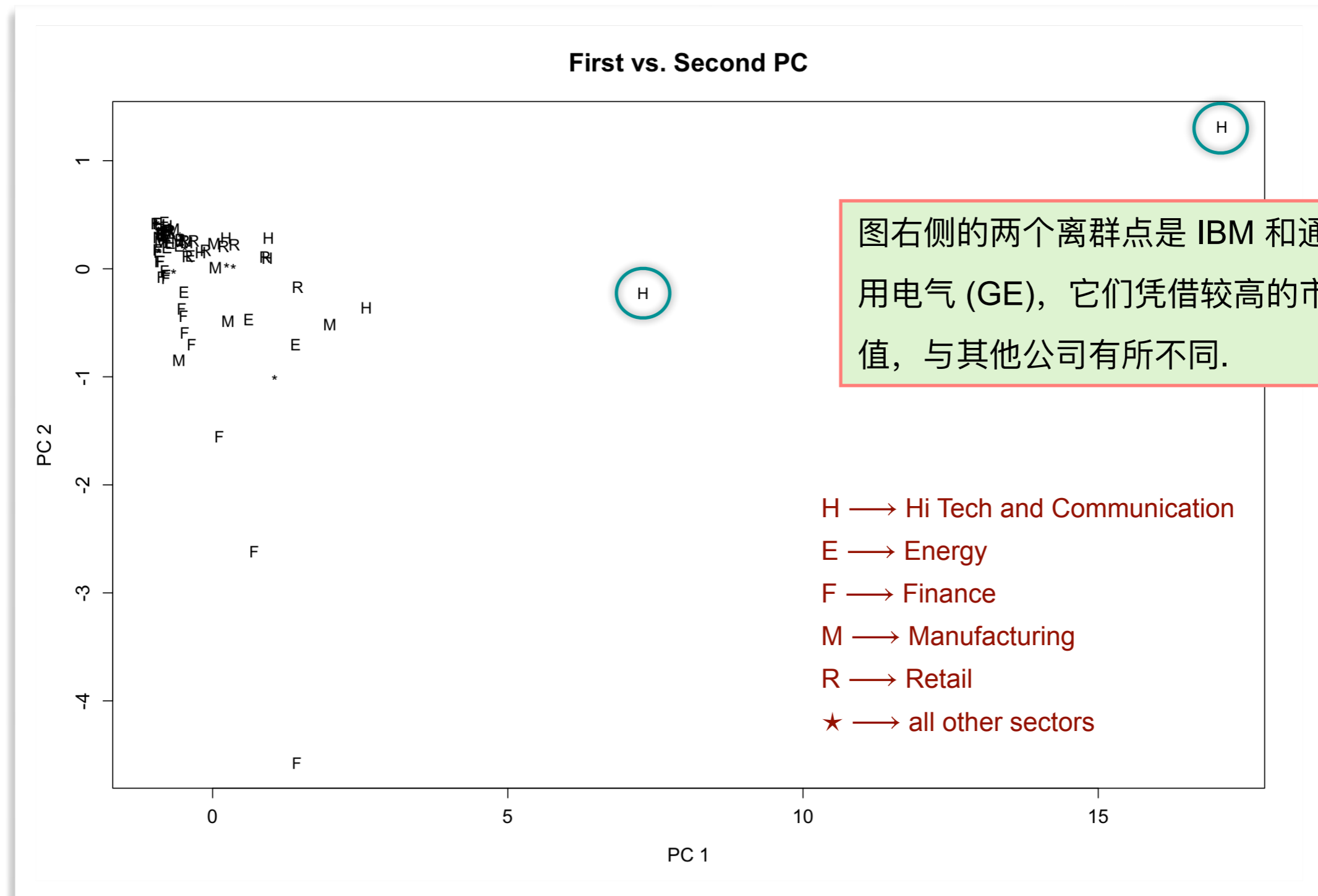
```
plot(cbind(-x[, 1], -x[, 2]), type = "n", xlab = "PC 1", ylab = "PC 2", main = "First vs. Second PC",  
      cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6)
```

```
text(cbind(-x[, 1], -x[, 2]), Sector)
```



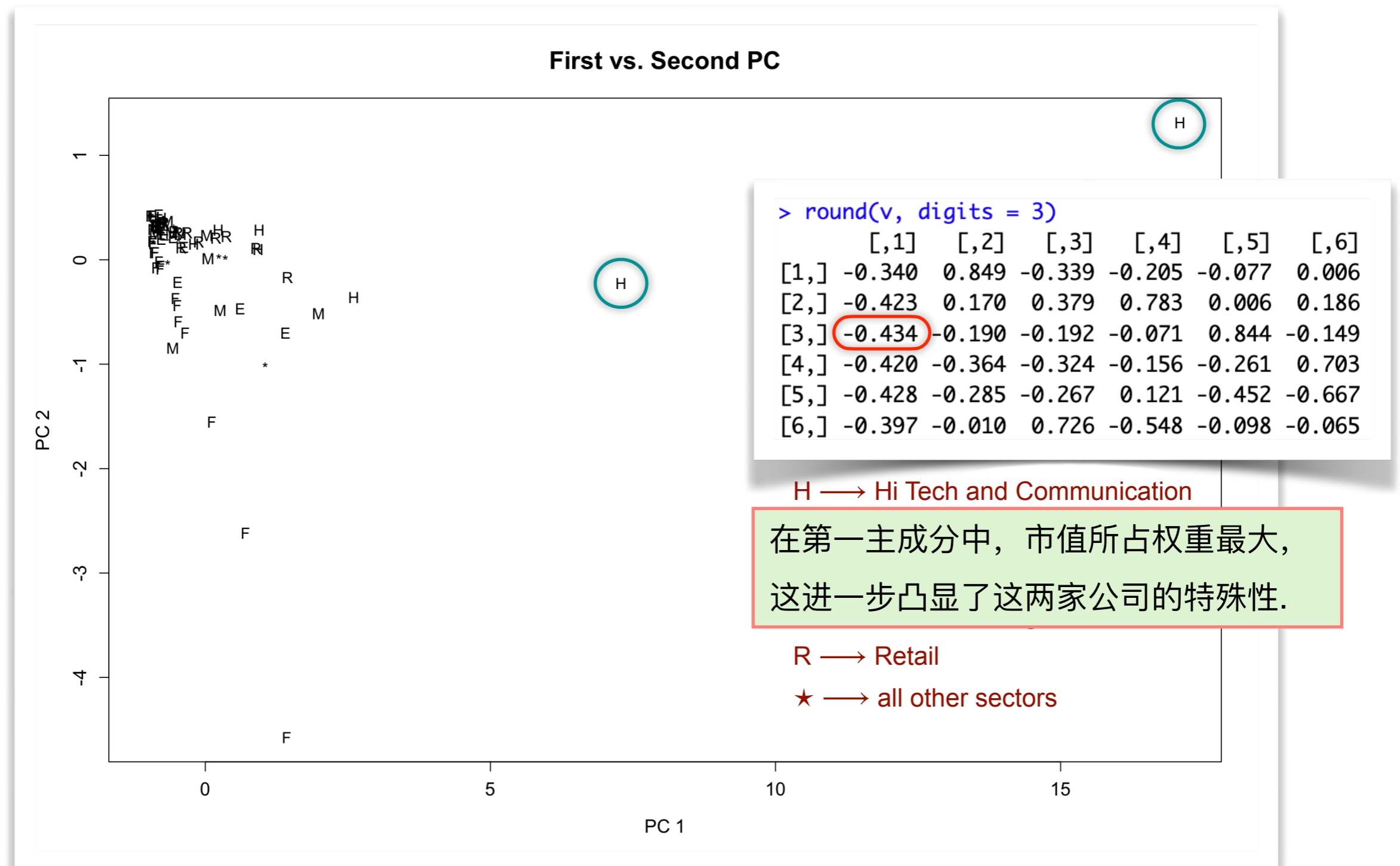
79 家美国公司

- 考虑 79 家美国公司的数据集.



79 家美国公司

- 考虑 79 家美国公司的数据集.



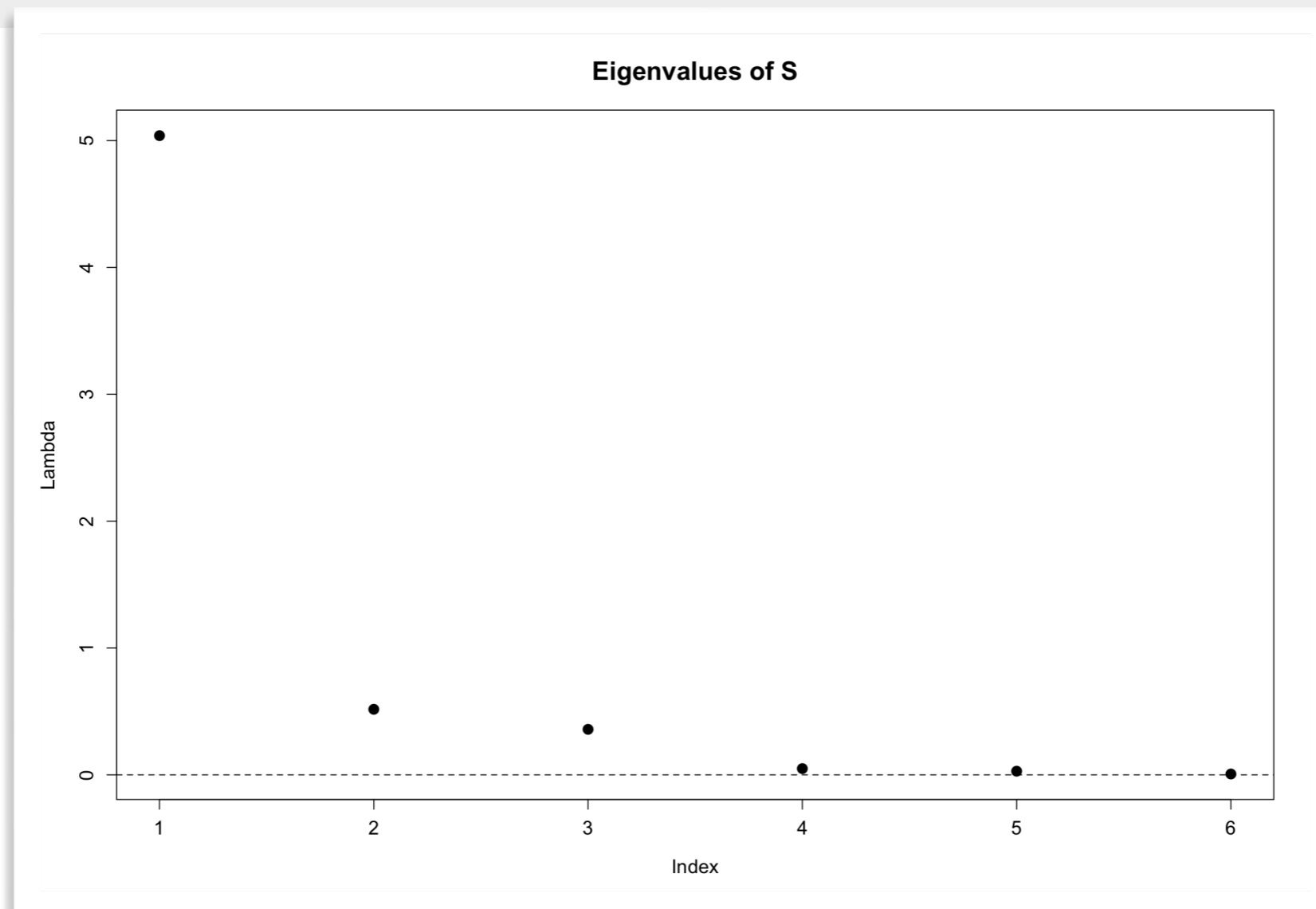
79 家美国公司

- 考虑 79 家美国公司的数据集。
 - ▶ 碎石图.

```
# scree plot
```

```
plot(e, xlab = "Index", ylab = "Lambda", main = "Eigenvalues of S", cex.lab = 1.2,  
      cex.axis = 1.2, cex.main = 1.6, pch = 16, cex = 1.5)
```

```
abline(h = 0, lty = 2)
```



79 家美国公司

- 考虑 79 家美国公司的数据集。
 - ▶ 如果将IBM和通用电气从数据集中剔除，将会呈现出截然不同的情形。

```
rm(list = ls(all = TRUE)) # clear all variables
graphics.off()
setwd("~/Desktop/2023_Applied Multivariate Statistical Analysis/R Codes with data/Data")
library(readr)
x = read_table("uscomp2.txt", col_name = FALSE)
x = rbind(x[1:37, ], x[39, ], x[41:79, ])
colnames(x) = c("Company", "A", "S", "MV", "P", "CF", "E", "Sector")
attach(x)
Sector = as.character(Sector)
Sector[1:2] = "H"
Sector[3:17] = "E"
Sector[18:34] = "F"
Sector[35:40] = "H"
Sector[41:50] = "M"
Sector[51:61] = "*"
Sector[62:71] = "R"
Sector[72:77] = "*"
detach(x)
```

79 家美国公司

- 考虑 79 家美国公司的数据集。
 - ▶ 此时的特征值和对应的特征向量为

```
x = x[, 2:7]
n = nrow(x)
# standardizes the data
x = (x - matrix(apply(x, 2, mean), n, 6, byrow = T))/matrix(sqrt((n - 1) * apply(x, 2, var)/n), n, 6, byrow = T)
# spectral decomposition
eig = eigen((n - 1) * cov(x)/n)
e = eig$values
round(as.matrix(e), digits = 3)
v = eig$vectors
round(v, digits = 3)
```

eigenvalues

```
> round(as.matrix(e), digits = 3)
      [,1]
[1,] 3.191
[2,] 1.535
[3,] 0.791
[4,] 0.292
[5,] 0.149
[6,] 0.041
```

corresponding eigenvectors

```
> round(v, digits = 3)
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,] -0.263 0.408 0.800 0.067 -0.333 -0.099
[2,] -0.438 0.407 -0.162 0.509 0.441 0.403
[3,] -0.500 0.003 0.035 -0.801 0.264 0.190
[4,] -0.331 -0.623 0.080 0.192 -0.426 0.526
[5,] -0.443 -0.450 0.123 0.238 0.335 -0.646
[6,] -0.427 0.277 -0.558 -0.021 -0.575 -0.313
```

79 家美国公司

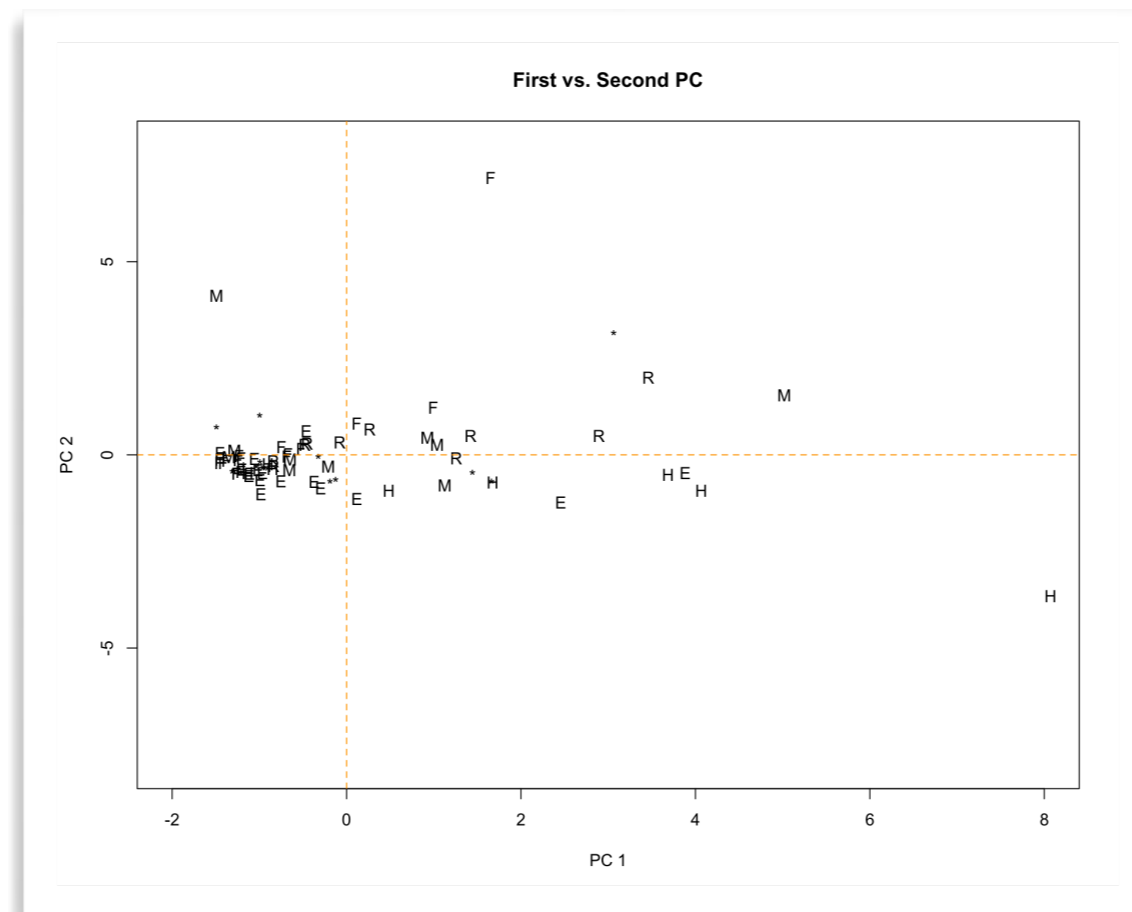
- 考虑 79 家美国公司的数据集。
 - ▶ 美国公司数据 (不含 IBM 和通用电气) 的主成分。

```
x = as.matrix(x) %*% v # principal components
```

```
plot(-x[, 1], x[, 2], xlim = c(-2, 8), ylim = c(-8, 8), type = "n", xlab = "PC 1", ylab = "PC 2", main = "First vs. Second PC")
```

```
text(-x[, 1], x[, 2], Sector)
```

```
abline(h = 0, v = 0, lty = 2, col = 'orange')
```

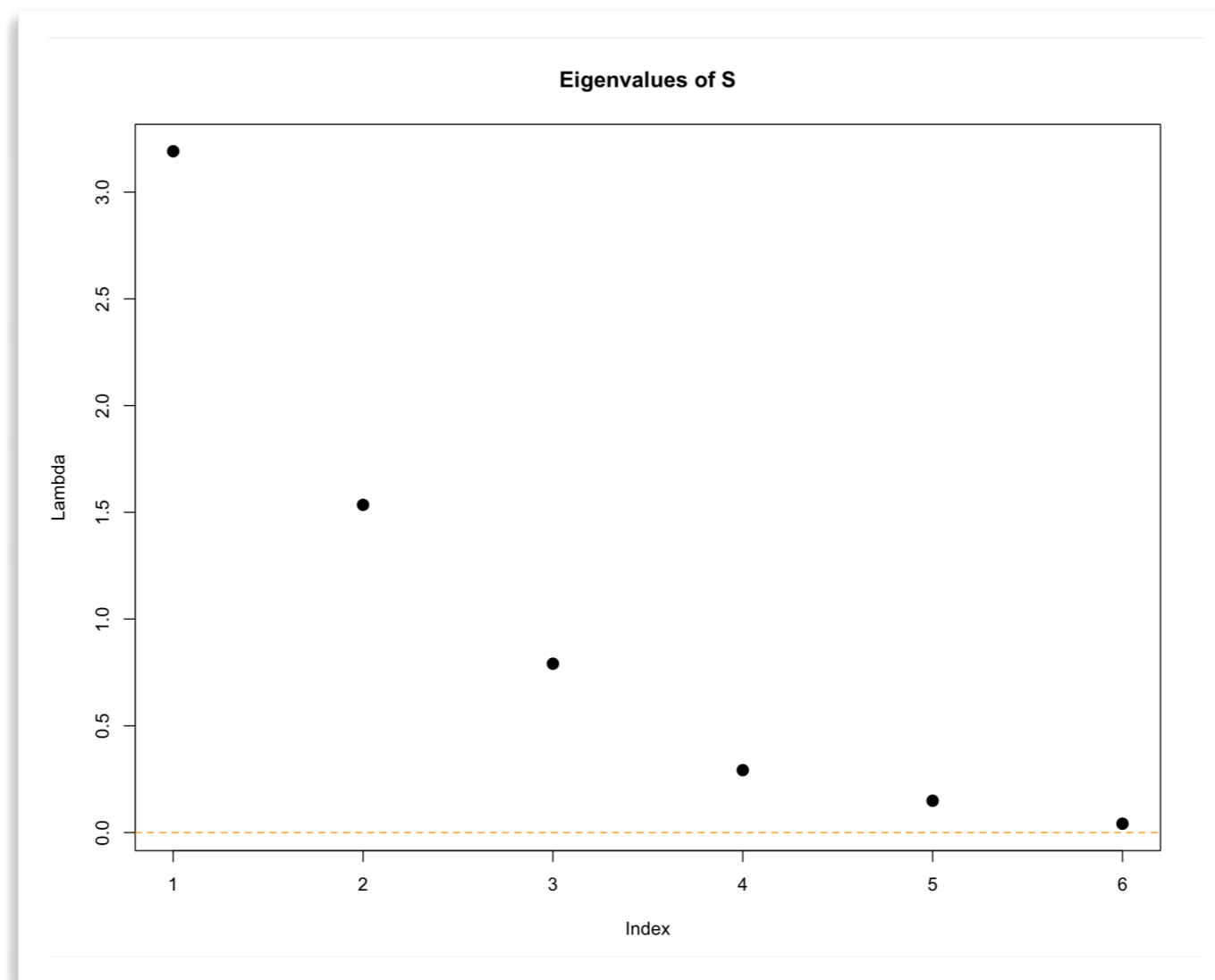


79 家美国公司

- 考虑 79 家美国公司的数据集。
 - ▶ 碎石图.

scree plot

```
plot(e, xlab = "Index", ylab = "Lambda", main = "Eigenvalues of S", pch = 16, cex = 1.5)  
abline(h = 0, lty = 2, col = 'orange')
```



79 家美国公司

- 考虑 79 家美国公司的数据集.
 - ▶ 特征值与解释方差的比例.

```
rm(list = ls(all = TRUE)) # clear all variables
graphics.off()
setwd("~/Desktop/2023_Applied Multivariate Statistical Analysis/R Codes with da
library(readr)
x = read_table("uscomp2.txt", col_name = FALSE)
x = rbind(x[1:37, ], x[39, ], x[41:79, ])
x = x[, 2:7]
n = nrow(x)
x = scale(x) # standardizes the data matrix
e = eigen((n - 1) * cov(x)/n) # calculates eigenvalues and eigenvectors
e1 = e$values
evprop = data.frame(eigenvalues = e1, proportion = e1/sum(e1), cumprop = cumsum(e1)/sum(e1))
round(evprop, digits = 3)
```

```
> round(evprop, digits = 3)
  eigenvalues proportion cumprop
1      3.150      0.532   0.532
2      1.515      0.256   0.788
3      0.781      0.132   0.920
4      0.289      0.049   0.968
5      0.147      0.025   0.993
6      0.041      0.007   1.000
```

79 家美国公司

- 考虑 79 家美国公司的数据集.
 - ▶ 变量与主成分的相关系数.

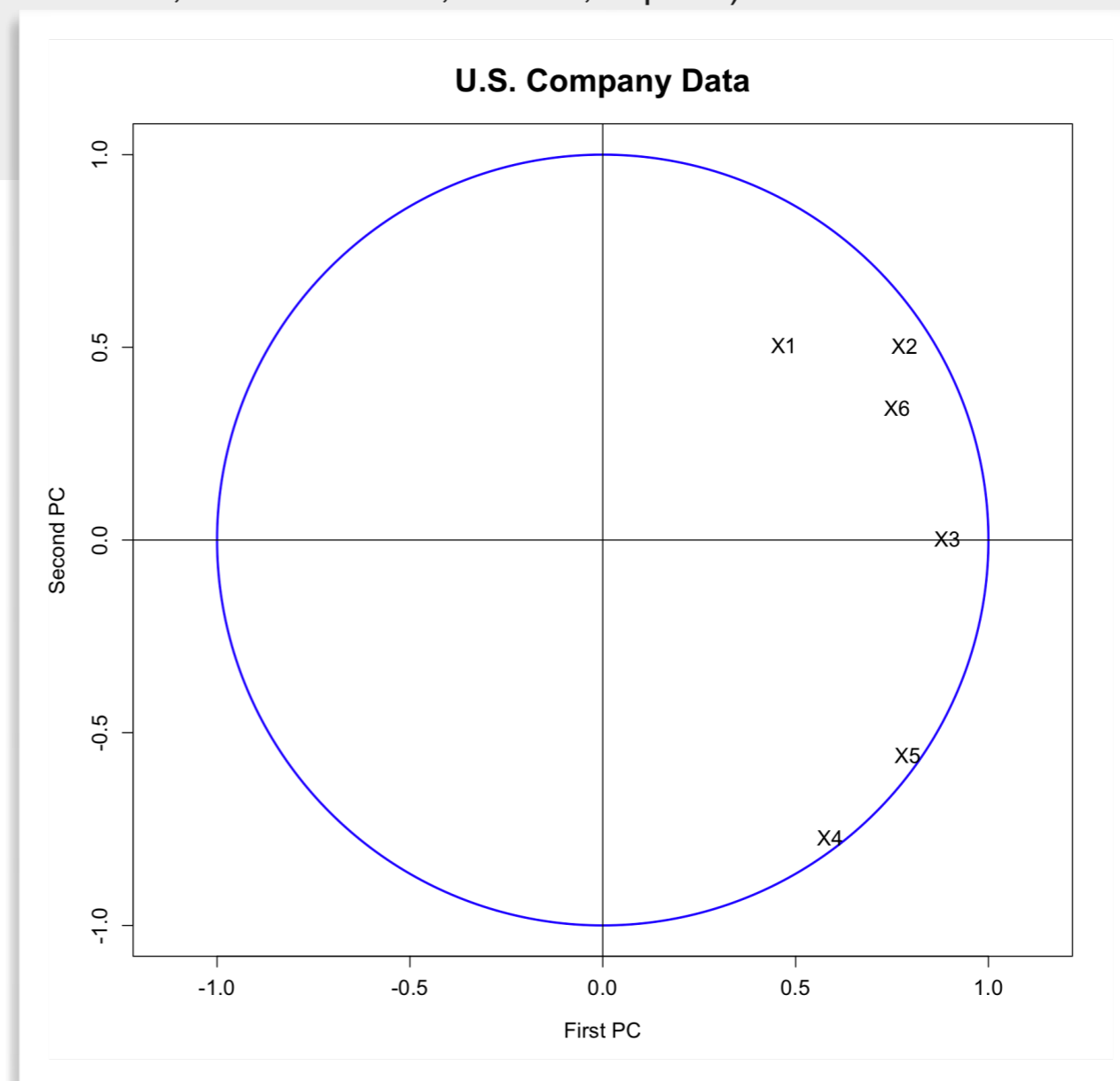
```
r = x %*% e$vector # principal components
r = cor(cbind(r, x)) # correlation between PCs and variables
r1 = r[7:12, 1:2] # correlation of the two most important PCs and variables
corpcx = data.frame(PC1 = r1[, 1], PC2 = r1[, 2], SquaredSum = r1[, 1]^2 + r1[, 2]^2)
rownames(corpcx) = c('X1', 'X2', 'X3', 'X4', 'X5', 'X6')
round(corpcx, digits = 2)
```

```
> round(corpcx, digits = 2)
      PC1  PC2 SquaredSum
X1 -0.47  0.50      0.48
X2 -0.78  0.50      0.87
X3 -0.89  0.00      0.80
X4 -0.59 -0.77      0.95
X5 -0.79 -0.56      0.94
X6 -0.76  0.34      0.70
```

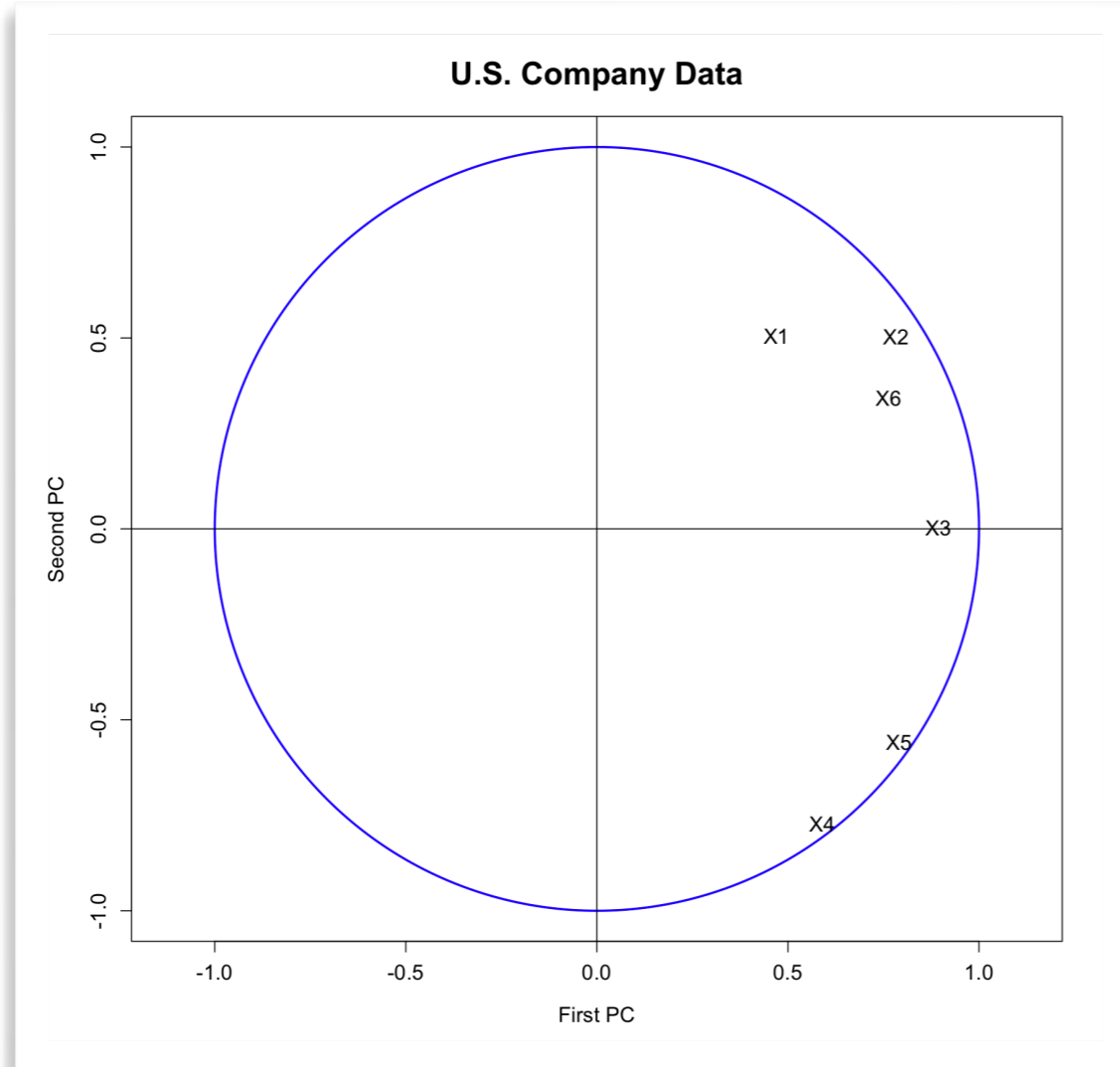
79 家美国公司

- 考虑 79 家美国公司的数据集。
 - ▶ 原始变量与前两个主成分的相关性.

```
ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))  
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab = "Second PC",  
     main = "U.S. Company Data", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.8, lwd = 2, asp = 1)  
abline(h = 0, v = 0)  
label = c("X1", "X2", "X3", "X4", "X5", "X6")  
text(-r1[, 1], r1[, 2], label, cex = 1.2)
```



79 家美国公司



```

> round(corpcx, digits = 2)
      PC1   PC2 SquaredSum
X1 -0.47  0.50      0.48
X2 -0.78  0.50      0.87
X3 -0.89  0.00      0.80
X4 -0.59 -0.77      0.95
X5 -0.79 -0.56      0.94
X6 -0.76  0.34      0.70
  
```

assets (X_1)

profits (X_4)

sales (X_2)

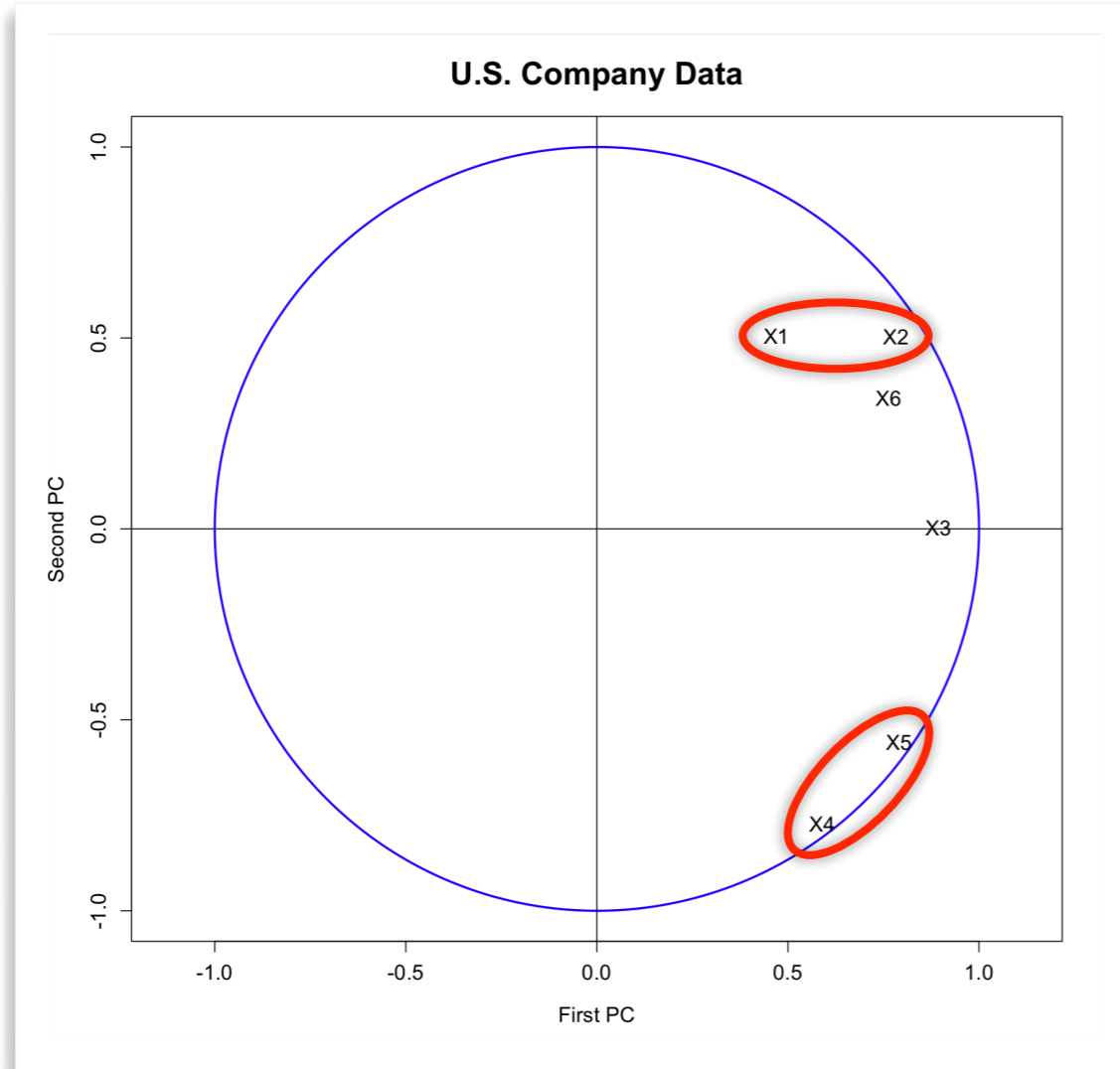
cash flow (X_5)

market value (X_3)

employees (X_6)

- ▶ 看起来第一主成分是一种“规模效应”，它与所有描述公司经营活动规模的变量都呈正相关。
- ▶ 这也是衡量企业经济实力的一种度量方式。

79 家美国公司



```
> round(corpcx, digits = 2)
```

	PC1	PC2	SquaredSum
X1	-0.47	0.50	0.48
X2	-0.78	0.50	0.87
X3	-0.89	0.00	0.80
X4	-0.59	-0.77	0.95
X5	-0.79	-0.56	0.94
X6	-0.76	0.34	0.70

assets (X_1)

profits (X_4)

sales (X_2)

cash flow (X_5)

market value (X_3)

employees (X_6)

- ▶ 第二主成分描述了公司的“形态”（“利润—现金流”与“资产—销售额”因素），从经济角度来看，这更难解释。