主成分分析

主成分分析方法

- ➢假设某一事件发生可能受 $x_1, x_2, \cdots, x_N$ 因素影响
  - ➢ 识别出到底哪些因素起主要作用，哪些可以忽略
- ➢而实测数据共有 M 组
- ➢这样可以假设这些数据由一个 $N \times M$ 矩阵 X 表示
- ➢记该矩阵的每一列的均值为 $\bar{x}_i, i = 1, 2, \cdots, N$

$$\texttt{mean}(\boldsymbol{X})$$

主成分分析方法的一般步骤

- 由 X 建立起协方差矩阵 R

$$r_{ij} = \frac{\sqrt{\sum_{k=1}^{M}(x_{ki}-\bar{x}_i)(k_{kj}-\bar{x}_j)}}{\sqrt{\sum_{k=1}^{M}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{M}(x_{kj}-\bar{x}_j)^2}}$$

$$\boldsymbol{R}=\mathtt{corr}(\boldsymbol{X}-\mathtt{mean}(\boldsymbol{X}))$$

处理R矩阵的特征值与特征向量

- 由 R 矩阵可以分别得出
  - 特征向量 $\boldsymbol{e}_i$
  - 对应的排序特征值λ_i, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N \geqslant 0$
- 特征向量矩阵的每一列进行了相应的归一化

$$||\boldsymbol{e}_i|| = 1 \quad 或 \quad \sum_{j=1}^{N} e_{ij}^2 = 1$$

```
[e,d]=eig(R); d=diag(d)
d=d(end:-1:1); e=fliplr(e);
```

找出哪些因素贡献大

- 计算主成分贡献率和累计贡献率
  - 主成分贡献率：

$$\gamma_i = \frac{\lambda_i}{\sum_{k=1}^{N} \lambda_k}$$

  - 累计贡献率：

$$\delta_i = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{N} \lambda_k}$$

- 若前 n 个特征值的累计贡献率大于85% ~ 95%
  - 可以认为这 n 个因素是原问题的主成分

$$\boldsymbol{d}/\mathrm{sum}(\boldsymbol{d})$$

构造变换矩阵

➢建立新变量坐标 $Z = XL$，即

$$\begin{cases} z_1 = l_{11}x_1 + l_{21}x_2 + \cdots + l_{N1}x_N \\ z_2 = l_{12}x_1 + l_{22}x_2 + \cdots + l_{N2}x_N \\ \quad \vdots \\ z_N = l_{1N}x_1 + l_{2N}x_2 + \cdots + l_{NN}x_N \end{cases}$$

其中变换矩阵第 i 列的系数 l_{ji} 可以计算 $l_{ji} = \sqrt{\lambda_i} e_{ji}$

$\boldsymbol{D}$=[$\boldsymbol{d}$'; $\boldsymbol{d}$'; ⋯; $\boldsymbol{d}$'];

$\boldsymbol{L}$=real(sqrt($\boldsymbol{D}$)).*$\boldsymbol{e}$, $\boldsymbol{Z}$=$\boldsymbol{X}$*$\boldsymbol{L}$

降维矩阵的构造

- 若前 n 个成分作主成分，则矩阵 L 的m 列以后各值应该趋于0，上式化为

$$\begin{cases} z_1 = l_{11}x_1 + l_{21}x_2 + \cdots + l_{N1}x_N \\ \quad\vdots \\ z_n = l_{1n}x_1 + l_{2n}x_2 + \cdots + l_{Nn}x_N \end{cases}$$

- 即，在适当的线性变换下，原来的 N 维问题就可以简化成 n 维问题

例9-44 主成分分析的降维处理

- 假设某三维曲线上的样本点由下列函数直接生成

$$x = t\cos 2t, y = t\sin 2t, z = 0.2x + 0.6y$$

- 试用主成分分析的方法对其降维处理
- MATLAB生成数据：

```
>> t=[0:0.1:3*pi]'; x=t.*cos(2*t);
   y=t.*sin(2*t); z=0.2*x+0.6*y;
   X=[x y z]; R=corr(X);
   [e,d]=eig(R), d=diag(d), plot3(x,y,z)
```

降维处理与效果

➢降维处理

```
>> d=d(end:-1:1); e=fliplr(e);
   D=[d'; d'; d']; L=real(sqrt(D)).*e, Z=X*L;
```

➢降维的效果

```
>> plot(Z(:,1),Z(:,2))
```

➢降维矩阵与坐标变换（ $Z = XL$ ）

$$\boldsymbol{L} = \begin{bmatrix} 0.184 & -0.9829 & 0 \\ 0.9653 & 0.261 & 0 \\ 0.9975 & -0.0713 & 0 \end{bmatrix} \Rightarrow \begin{cases} z_1 = 0.1840x + 0.9653y + 0.9975z \\ z_2 = -0.9829x + 0.2610y - 0.0713z \end{cases}$$