主成分分析

TUTU

主成分分析基本概念

♣ 基本思想：

主成分分析就是把原有的多个指标转化成少数几个代表性较好的综合指标，这少数几个指标能够反映原来指标大部分的信息 (85% 以上)，并且各个指标之间保持独立，避免出现重叠信息。主成分分析主要起着降维和简化数据结构的作用。

♣ 做法：

主成分分析通常的做法是，寻求原指标的线性组合 F_i：

$$F_i = \sum_{k=1}^{p} a_{ki} X_k$$

i 从 1 到 p，信息量逐渐减少

主成分的性质

♣ 主成分满足以下性质：

- $a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2 = 1$
- 若 $i \neq j$，则$\mathrm{Cov}(F_i, F_j) = 0$
- $\mathrm{Var}(F_1) \geq \mathrm{Var}(F_2) \geq \cdots \geq \mathrm{Var}(F_p)$

总体主成分

♣ 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的协方差矩阵

$$\boldsymbol{\Sigma}=E[(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))^{\mathrm{T}}]$$

♣ 第一主成分的定义为$F_1=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p=\alpha_1^{\mathrm{T}}X$ 是在 $\alpha_1^{\mathrm{T}}\alpha_1=1$ 的条件下，最大化$\mathrm{Var}(F_1)=\mathrm{Var}(\alpha_1^{\mathrm{T}}X)=\alpha_1^{\mathrm{T}}\boldsymbol{\Sigma}\alpha_1=\lambda_{\max}$

♣ 若 $F_1,F_2,\cdots,F_{i-1}$ 还不够，则需要第 i 个主成分，其定义为$F_i=\alpha_i^{\mathrm{T}}\boldsymbol{X}=a_{1i}X_1+a_{2i}X_2+\cdots+a_{pi}X_i$；是在 $\alpha_i^{\mathrm{T}}\alpha_i=1$ 和 $\mathrm{Cov}(F_i,F_k)=\alpha_i^{\mathrm{T}}\boldsymbol{\Sigma}\alpha_k=0$的条件下，最大化$\mathrm{Var}(F_i)=\alpha_i^{\mathrm{T}}\boldsymbol{\Sigma}\alpha_i=\lambda_{(i)}$

♣ 前 m 个主成分的累积贡献率 $(80\%\sim85\%)$：$\dfrac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\times100\%$

总体主成分

♣ 求主成分的基本步骤：

1. 对数据进行标准化变换 (R 型分析)
2. 求样本协方差矩阵 S
3. 求 S 的特征根 λ_k
4. 求特征根所对应的正交单位化特征向量 α_k
5. 写出主成分的表达式 F_k

♣ S 型分析和 R 型分析：

- S 型分析：协方差矩阵出发的主成分分析
- R 型分析：相关矩阵出发进行主成分分析，消除量纲影响，在计算之前先将原始数据标准化

总体主成分的性质

♣ 总体主成分的性质：

- $\mathrm{Var}(F_k) = \lambda_k$
- F_k 的方差贡献率：$\lambda_k / \sum \lambda_i$
- 因子载荷 $r(X_i, F_k) = \begin{cases} a_{ik}\frac{\sqrt{\lambda_k}}{\sqrt{s_{ii}}}, & \text{S 型分析} \\ a_{ik}\sqrt{\lambda_k}, & \text{R 型分析} \end{cases}$
- 主成分对每个原变量的方差贡献：$r(X_i, F_k)^2$
- $\sum_k r(X_i, F_k)^2 = 1$，$\sum_k r(X_i, F_k)^2 s_{ii} = \lambda_k$
- $\sum \lambda_i = \begin{cases} \sum s_{ii}, & \text{S 型分析} \\ p, & \text{R 型分析} \end{cases}$

系统评估

♣ 方法：

- 通过主成分分析得到综合指标 F_1
- 利用 F_1 作为评估指标，根据 F_1 得分对样本点进行排序比较
- 三个前提：
  - $r(X_i, F_k) = a_{ik}\sqrt{\lambda_k} > 0$
  - a_ik 在数值上分布较均匀
  - F_1 的方差贡献率较大

主成分分析 SAS 代码

♣ SAS 代码:

```
/*主成分分析*/
proc princomp data=yourdata out=out5_1;
run;

/*主成分分析，取前三个主成分，并绘制两个主成分的得分图*/
proc princomp data=yourdata n=3 out=out5_1 plot=score(ncomp=2);
id industry;
run;

/*根据第一主成分降序排序*/
proc sort data=out5_1;
by descending prin1;
run;

/*打印样品名与第一主成分得分*/
proc print;
var state prin1;
run;
```

主成分回归

♣ 主成分回归基本步骤：

1. 对自变量进行主成分分析，取反映 95% 的主成分
2. 以主成分为新的自变量，建立回归模型 $Y_i = \sum_k \hat{\gamma}_k F_{ki}$
3. 把主成分的表达式代入，得到最终的回归模型 $Y_i = \sum_k \hat{\beta}_k x_{ik}$

♣ 无量纲化方法：

- 标准化
- 均值化：$x'_{ji} = \frac{x_{ji}}{\bar{x}_i}$，协方差矩阵 S 中的元素 $u_{ji} = \frac{s_{jk}}{\bar{x}_j \bar{x}_k}$

  均值化处理后的协方差矩阵不仅消除了指标量纲与数量级的影响，还能包含原始数据的全部信息
- 最大最小化

主成分回归 SAS 代码

♣ SAS 代码:

```
/*主成分回归*/

proc reg data=yourdata outest=out;

model y=x1-x3/pcomit=1,2;

run;

quit;


/*打印结果*/

proc print data=out;

run;
```

逐步回归与岭回归 SAS 代码

♣ SAS 代码:

```
/*岭回归*/
proc reg data=yourdata outest=out;
model y=x1-x3/ridge=0 to 2 by 0.1;
run;

/*向前逐步回归*/
proc reg data=yourdata outest=out;
model y=x1-x3/selection=stepwise slentry=0.1 slstay=0.15;
run;

/*向后逐步回归*/
proc reg data=yourdata outest=out1;
model y=x1-x3/selection=backward slstay=0.15;
run;

quit;
proc print data=out;
run;
```