多元统计

陈崇双

西南交通大学数学学院统计系
ccsmars@swjtu.edu.cn

2018-2019学年

1 主成分分析

- 问题背景
- 变量降维的可行性
- 基本原理
- 主成分的性质
- 由相关阵提取主成分
- 一些细节

第一节：问题背景

为了全面准确地反映事物的特征和变化规律，往往考虑与其有关的多个指标（或变量）。

问题背景

中国可持续发展指标体系，来源《中国经济网》，http://www.ce.cn/xwzx/gnsz/gdxw/201801/08/t20180108_27611518.shtml

问题背景

一级指标	评价维度	表征指标
创新驱动	科技投入	研究与实验发展经费投入强度（%）
	知识生产	国际科技论文被引次数
	科技价值	每万人口发明专利拥有量（件）
集约高效	价值创造	人均GDP（万元）
	产业质量	工业增加值率（%）
	空间集约	亩均增加值（万元）
平衡普惠	产业结构	知识密集型服务业增加值占GDP的比例（%）
	城乡均衡	城乡人均可支配收入比值
	分配公平	劳动收入占GDP比重（%）
绿色生态	能源消耗	单位GDP能耗（吨标准煤/万元）
	水消耗	单位GDP水耗（立方米/万元）
	碳排放	单位GDP碳排放（千克/万元）

高质量发展指标体系, 来源《凤凰网》,

http://wemedia.ifeng.com/88805720/wemedia.shtml

问题背景

- 优点：描述详尽，刻画细腻。
- 缺点：增加问题的复杂性，信息重叠，主次不清，难以获得直观清晰的把握。

问题背景

- 优点：描述详尽，刻画细腻。
- 缺点：增加问题的复杂性，信息重叠，主次不清，难以获得直观清晰的把握。

能否用较少的几项指标来代替原来的指标，并且能较多地反映原来指标的信息？

问题背景

Hotelling于1933年首先提出主成分分析(Principal Component Analysis, PCA)。

问题背景

Hotelling于1933年首先提出主成分分析(Principal Component Analysis, PCA)。

PCA并不是比较各指标的重要性，将不太重要的指标简单去掉，而是通过全面分析各项指标所携带的信息，从中提取出一些潜在的综合性指标(主成分)。用综合性指标替代原来较多的指标。

问题背景

Hotelling于1933年首先提出主成分分析(Principal Component Analysis, PCA)。

PCA并不是比较各指标的重要性，将不太重要的指标简单去掉，而是通过全面分析各项指标所携带的信息，从中提取出一些潜在的综合性指标(主成分)。用综合性指标替代原来较多的指标。

例如，上海证券综合指数，简称上证综指，是上海证券交易所编制，以上海证券交易所挂牌上市的全部股票为计算范围，以发行量为权数综合，反映了上海证券交易市场的总体走势。(来源搜狗百科)

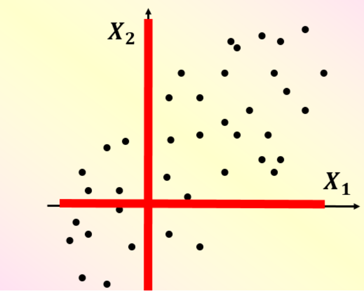
截止至2018.06


快速检索
标准检索
专业检索
作者发文检索
科研基金检索
句子检索
来源期刊检索
1.输入检索控制条件：
期刊年期：从 不限 年 到 不限 年
指定期： 请输入
更新时间： 不限
来源期刊： 输入期刊名称，ISSN，CN均可 模糊
来源类别： EI来源期刊
支持基金： 输入基金名称 模糊
作者 输入作者姓名 精确
作者单位： 输入作者单位，全称、简称、曾用名均可 模糊
2.输入内容检索条件：
篇名 主成分分析 词频 并且包含 输入检索词 词频 精确
或者 关键词 主成分分析 词频 并且包含 词频 精确
仅限优先出版论文
中英文扩展检索
在结果中检索
检索文献
定制本次检索式
3.您可以按如下文献分组排序方式选择文献：(分组只对前4万条记录分组,排序只在800万条记录以内有效)
文献分组浏览： 学科类别 期刊名称 研究资助基金 研究层次 文献作者 作者单位 中文关键词 发表年度 不分组
"紫色"刊名为"中国知网"独家出版刊物
文献排序浏览： 发表时间 相关度 被引频次 下载频次
列表显示
每页记录数： 10 20 50
找到 1,621 条结果 浏览1/82
1 2 3 4 5 6 7 8 9 后页
全选 清除 存盘 定制

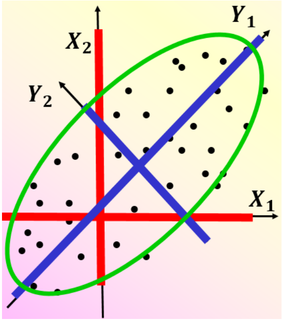
计算机软件及应用(309)	轻工业手工业(192)
自动化技术(178)	化学(167)
环境科学与资源利用(107)	电力工业(74)
工业通用技术及设备(60)	电信技术(57)
物理学(54)	地质学(48)
农作物(44)	植物保护(40)
农业基础科学(39)	宏观经济管理(38)
建筑科学与工程(32)	金属学及金属工艺(32)
矿业工程(29)	公路与水路运输(28)
生物学(28)	有机化工(25)
中药学(25)	机械工业(25)
互联网技术(25)	一般化学工业(24)
地球物理学(22)	汽车工业(21)
自然地理学和测绘学(19)	

例1

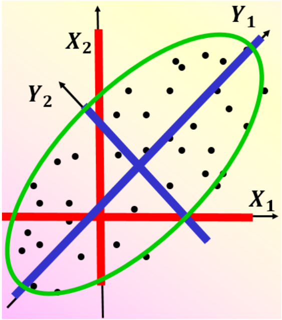
设有一批样品，每个样品涉及两个变量X_1和X_2。在二维平面中，样本点散布情况如图所示。



样本点沿X_1和X_2轴都有一定离散性。

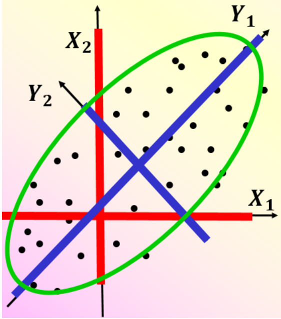


(1) 若以椭圆长短轴为坐标轴，相当于平移旋转变换。不论选用何种坐标系，同样的样品所反映的信息量肯定相同。



(1) 若以椭圆长短轴为坐标轴，相当于平移旋转变换。不论选用何种坐标系，同样的样品所反映的信息量肯定相同。

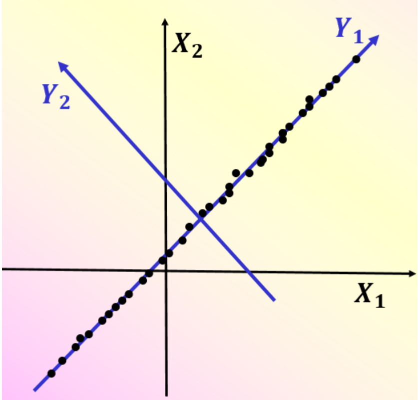
(2) 数据越离散，则分布越广泛，代表性越好，故可用方差表示信息量。一个变量Y_1集中反映X_1和X_2的大部分信息，忽略Y_2也无损大局。



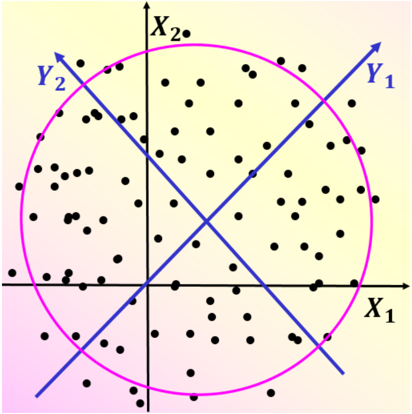
(1) 若以椭圆长短轴为坐标轴，相当于平移旋转变换。不论选用何种坐标系，同样的样品所反映的信息量肯定相同。

(2) 数据越离散，则分布越广泛，代表性越好，故可用方差表示信息量。一个变量Y_1集中反映X_1和X_2的大部分信息，忽略Y_2也无损大局。

(3) Y_1和Y_2正交，避免了信息重叠。



用Y_1来反映X_1和X_2共同的信息，几乎没有损失!



数据杂乱分散，一个变量难以反映X_1和X_2共同的信息!

变量降维的可行性

主成分与原始指标(变量)之间:

(1) 经坐标轴的平移和旋转变换后，样品的信息可在新坐标系中选用少数几个轴进行压缩。即，主成分的数目少于原始指标的数目。

变量降维的可行性

主成分与原始指标(变量)之间:

(1) 经坐标轴的平移和旋转变换后，样品的信息可在新坐标系中选用少数几个轴进行压缩。即，主成分的数目少于原始指标的数目。

(2) 坐标轴进行平移旋转变换，则样品在新坐标系(Y_1, Y_2)下的坐标，为原坐标系(X_1, X_2)下坐标的线性组合。即，每一个主成分都是各原始指标的线性组合。

变量降维的可行性

主成分与原始指标(变量)之间:

(1) 经坐标轴的平移和旋转变换后，样品的信息可在新坐标系中选用少数几个轴进行压缩。即，主成分的数目少于原始指标的数目。

(2) 坐标轴进行平移旋转变换，则样品在新坐标系(Y_1, Y_2)下的坐标，为原坐标系(X_1, X_2)下坐标的线性组合。即，每一个主成分都是各原始指标的线性组合。

(3) 少数几个主成分，一方面保留了原始指标的绝大多数信息，另一方面各主成分不重叠地反映某个方面的综合信息，即不相关。

概括起来：

主成分分析方法：全面分析各项指标携带的信息，从中提取较少几项综合性指标（主成分），且互不相关，最大限度地保留指标所反映的信息，进而用这较少的几项综合性指标来刻划个体。

第三节：基本原理

主要内容：数学模型，求解

基本原理

设有一个p维总体$(X_1, X_2, \cdots, X_p)$，从中抽取n个样本，观测值向量分别为$\boldsymbol{x}_{(i)}, i = 1, 2, \cdots, n$, 其中$\boldsymbol{x}_{(i)} = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}, i = 1, 2, \cdots, n$。希望通过这$p$项可观测指标$X_1, X_2, \cdots, X_p$提取出$m$(远小于$p$)项综合性指标$Y_1, Y_2, \cdots, Y_m$。

基本原理

设有一个p维总体$(X_1, X_2, \cdots, X_p)$，从中抽取n个样本，观测值向量分别为$\boldsymbol{x}_{(i)}, i = 1, 2, \cdots, n$, 其中$\boldsymbol{x}_{(i)} = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}$, $i = 1, 2, \cdots, n$。希望通过这p项可观测指标$X_1, X_2, \cdots, X_p$提取出m(远小于p)项综合性指标$Y_1, Y_2, \cdots, Y_m$。

PCA的处理办法：

1. 将每个综合性指标都看成是各可观测指标的线性组合;
2. 采用方差来度量一个随机变量所包含的信息量。
3. 有效性VS.充分性。

基本原理

建立如下数学模型：

$$
\begin{cases}
Y_1 = l_{11}X_1 + l_{12}X_2 + \cdots + l_{1p}X_p \triangleq \boldsymbol{l}_1^{\top}\boldsymbol{X} \\
Y_2 = l_{21}X_1 + l_{22}X_2 + \cdots + l_{2p}X_p \triangleq \boldsymbol{l}_2^{\top}\boldsymbol{X} \\
\cdots \\
Y_m = l_{m1}X_1 + l_{m2}X_2 + \cdots + l_{mp}X_p \triangleq \boldsymbol{l}_m^{\top}\boldsymbol{X}
\end{cases}
$$

建立如下数学模型：

$$\begin{cases} Y_1 = l_{11}X_1 + l_{12}X_2 + \cdots + l_{1p}X_p \triangleq \boldsymbol{l}_1^\mathsf{T}\boldsymbol{X} \\ Y_2 = l_{21}X_1 + l_{22}X_2 + \cdots + l_{2p}X_p \triangleq \boldsymbol{l}_2^\mathsf{T}\boldsymbol{X} \\ \cdots \\ Y_m = l_{m1}X_1 + l_{m2}X_2 + \cdots + l_{mp}X_p \triangleq \boldsymbol{l}_m^\mathsf{T}\boldsymbol{X} \end{cases}$$

其中$\boldsymbol{l}_i = (l_{i1}, l_{i2}, \cdots, l_{ip})^\mathsf{T}$, $i = 1, 2, \ldots m$是常向量;
$\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^\mathsf{T}$的均值向量为$\boldsymbol{\mu}$, 协方差阵为$\boldsymbol{\Sigma}$。

通过确定m个常数向量$\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_m$, 使得$Y_i = \boldsymbol{l}_i^\mathsf{T}\boldsymbol{X}$的方差尽可能大, 且各$Y_i$之间互不相关。

基本原理

注1：Y_i与Y_j不相关，则$Cov(Y_i, Y_j) = Cov(\boldsymbol{l}_i X, \boldsymbol{l}_j X) = \boldsymbol{l}_i^{\top} \boldsymbol{\Sigma} \boldsymbol{l}_j = 0$。

基本原理

注1：Y_i与Y_j不相关，则$Cov(Y_i, Y_j) = Cov(\boldsymbol{l}_i X, \boldsymbol{l}_j X) = \boldsymbol{l}_i^\top \boldsymbol{\Sigma} \boldsymbol{l}_j = 0$。

注2：由于$D(Y_i) = D(\boldsymbol{l}_i^\top \boldsymbol{X}) = \boldsymbol{l}_i^\top \boldsymbol{\Sigma} \boldsymbol{l}_i$, 则

$$D((c\boldsymbol{l}_i)^\top \boldsymbol{X}) = (c\boldsymbol{l}_i)^\top \boldsymbol{\Sigma}(c\boldsymbol{l}_i) = c^2 \boldsymbol{l}_i^\top \boldsymbol{\Sigma} \boldsymbol{l}_i = c^2 D(Y_i), \forall c \in R$$

注1：Y_i与Y_j不相关，则$Cov(Y_i,Y_j)=Cov(\boldsymbol{l}_iX,\boldsymbol{l}_jX)=\boldsymbol{l}_i^\top\boldsymbol{\Sigma}\boldsymbol{l}_j=0$。

注2：由于$D(Y_i)=D(\boldsymbol{l}_i^\top\boldsymbol{X})=\boldsymbol{l}_i^\top\boldsymbol{\Sigma}\boldsymbol{l}_i$, 则

$$D((c\boldsymbol{l}_i)^\top\boldsymbol{X})=(c\boldsymbol{l}_i)^\top\boldsymbol{\Sigma}(c\boldsymbol{l}_i)=c^2\boldsymbol{l}_i^\top\boldsymbol{\Sigma}\boldsymbol{l}_i=c^2D(Y_i),\forall c\in R$$

即有$D(Y_i)\to\infty\quad(\|\boldsymbol{l}_i\|=\sqrt{\boldsymbol{l}_i^\top\boldsymbol{l}_i}\to\infty)$, 从而无意义。

基本原理

因此，线性变换需满足如下约束:

1. $\boldsymbol{l}_i^{\top}\boldsymbol{l}_i = l_{i1}^2 + l_{i2}^2 + \cdots + l_{ip}^2 = 1$，$i = 1, 2, \cdots, p$;
2. $\boldsymbol{l}_i^{\top}\boldsymbol{\Sigma}\boldsymbol{l}_j = 0$，$i, j = 1, 2, \cdots, m$，且$i \neq j$。

基本原理

因此，线性变换需满足如下约束:

1. $\boldsymbol{l}_i^\top \boldsymbol{l}_i = l_{i1}^2 + l_{i2}^2 + \cdots + l_{ip}^2 = 1$，$i = 1, 2, \cdots, p$;
2. $\boldsymbol{l}_i^\top \boldsymbol{\Sigma} \boldsymbol{l}_j = 0$，$i, j = 1, 2, \cdots, m$，且$i \neq j$。

也即是说，在$X_1, X_2, \cdots, X_p$满足约束(1)的所有线性组合中，

1. Y_1是方差最大者;
2. Y_2是与Y_1不相关中的方差最大者;
3. $\ldots$
4. Y_m是与$Y_1, Y_2, \cdots, Y_{m-1}$都不相关中的方差最大者。

基本原理

因此，线性变换需满足如下约束:

1. $\boldsymbol{l}_i^\top \boldsymbol{l}_i = l_{i1}^2 + l_{i2}^2 + \cdots + l_{ip}^2 = 1$，$i = 1, 2, \cdots, p$;
2. $\boldsymbol{l}_i^\top \boldsymbol{\Sigma} \boldsymbol{l}_j = 0$，$i, j = 1, 2, \cdots, m$，且$i \neq j$。

也即是说，在$X_1, X_2, \cdots, X_p$满足约束(1)的所有线性组合中，

1. Y_1是方差最大者;
2. Y_2是与Y_1不相关中的方差最大者;
3. $\ldots$
4. Y_m是与$Y_1, Y_2, \cdots, Y_{m-1}$都不相关中的方差最大者。

综合变量$Y_1, Y_2, \cdots, Y_m$分别称为原始变量的第一、第二、……、第m个主成分。

基本原理

定理1

设p阶实对称矩阵$\boldsymbol{A}$，将其特征值按大小顺序排列为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$;
各特征值对应的标准正交特征向量分别为$\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p$。则$\boldsymbol{x} \in \mathbb{R}^p$有

$$\max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \lambda_1, \qquad \min_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \lambda_p$$

定理1

设p阶实对称矩阵$\boldsymbol{A}$，将其特征值按大小顺序排列为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$;
各特征值对应的标准正交特征向量分别为$\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p$。则$\boldsymbol{x} \in \mathbb{R}^p$有

$$\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \lambda_1, \qquad \min_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \lambda_p$$

且

$$argmax_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \boldsymbol{\xi}_1, \qquad argmin_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \boldsymbol{\xi}_p$$

Proof.

矩阵$\boldsymbol{A}$与单位阵$\boldsymbol{I}$有谱分解$\boldsymbol{A}=\sum_{i=1}^{p}\lambda_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\top}$，$\boldsymbol{I}=\sum_{i=1}^{p}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\top}$。

基本原理

Proof.

矩阵$\boldsymbol{A}$与单位阵$\boldsymbol{I}$有谱分解$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top$，$\boldsymbol{I} = \sum_{i=1}^{p} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top$。
对$\boldsymbol{x} \in \mathbb{R}^p$有$\boldsymbol{x} = \sum_{i=1}^{p} a_i \boldsymbol{\xi}_i$, 进而

Proof.

矩阵$\boldsymbol{A}$与单位阵$\boldsymbol{I}$有谱分解$\boldsymbol{A}=\sum_{i=1}^{p}\lambda_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathsf{T}}$，$\boldsymbol{I}=\sum_{i=1}^{p}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathsf{T}}$。

对$\boldsymbol{x}\in\mathbb{R}^p$有$\boldsymbol{x}=\sum_{i=1}^{p}a_i\boldsymbol{\xi}_i$, 进而

$$\max_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}}=\max_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\sum_{i=1}^{p}\lambda_i a_i^2}{\sum_{i=1}^{p}a_i^2}\le\frac{\lambda_1\sum_{i=1}^{p}a_i^2}{\sum_{i=1}^{p}a_i^2}=\lambda_1$$

$$\min_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}}=\min_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\sum_{i=1}^{p}\lambda_i a_i^2}{\sum_{i=1}^{p}a_i^2}\ge\frac{\lambda_p\sum_{i=1}^{p}a_i^2}{\sum_{i=1}^{p}a_i^2}=\lambda_p$$

□

类似地有如下结论:

命题1

设p阶实对称矩阵$\boldsymbol{A}$，将其特征值按大小顺序排列为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$; 各特征值对应的标准正交特征向量分别为$\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p$。则$\boldsymbol{x} \in \mathbb{R}^p$有

$$\max_{\boldsymbol{x}\neq\boldsymbol{0};\boldsymbol{x}^T\boldsymbol{\xi}_i=\boldsymbol{0},i=1,2,\cdots,k} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \lambda_{k+1}$$

$$\min_{\boldsymbol{x}\neq\boldsymbol{0};\boldsymbol{x}^T\boldsymbol{\xi}_i=\boldsymbol{0},i=1,2,\cdots,k} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \lambda_p$$

基本原理

命题2

设随机向量$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^T$的协方差矩阵为$\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}$的特征值从大到小排列为$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$, 各特征值对应的标准正交特征向量分别为$\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p$，则第$i$个主成分为

$$Y_i=\xi_{i1}X_1+\xi_{i2}X_2+\cdots+\xi_{ip}X_p=\boldsymbol{\xi}_i^T\boldsymbol{X}, i=1,2,\cdots,m$$

且

$$Cov(Y_i,Y_j)=\boldsymbol{\xi}_i^T\boldsymbol{\Sigma}\boldsymbol{\xi}_j=\begin{cases}\lambda_i, & i=j;\\ 0, & i\neq j.\end{cases}$$

基本原理

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$$

基本原理

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_1}{\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_1}$$

基本原理

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_1}{\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_1} = \lambda_1$$

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_1}{\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_1} = \lambda_1$$

类似地可得:

$$D(Y_{k+1}) = \max_{\boldsymbol{x}\neq\boldsymbol{0};\boldsymbol{x}^\top\boldsymbol{\xi}_i=\boldsymbol{0},i=1,2,\cdots,k} \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$$

基本原理

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_1}{\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_1} = \lambda_1$$

类似地可得:

$$D(Y_{k+1}) = \max_{\boldsymbol{x} \neq \boldsymbol{0}; \boldsymbol{x}^\top \boldsymbol{\xi}_i = \boldsymbol{0}, i=1,2,\cdots,k} \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_{k+1}}{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\xi}_{k+1}}$$

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_1}{\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_1} = \lambda_1$$

类似地可得:

$$D(Y_{k+1}) = \max_{\boldsymbol{x} \neq \boldsymbol{0}; \boldsymbol{x}^\top \boldsymbol{\xi}_i = \boldsymbol{0}, i=1,2,\cdots,k} \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_{k+1}}{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\xi}_{k+1}} = \lambda_{k+1}$$

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_1}{\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_1} = \lambda_1$$

类似地可得:

$$D(Y_{k+1}) = \max_{\boldsymbol{x} \neq \boldsymbol{0}; \boldsymbol{x}^\top \boldsymbol{\xi}_i = \boldsymbol{0}, i=1,2,\cdots,k} \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_{k+1}}{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\xi}_{k+1}} = \lambda_{k+1}$$

$$Cov(Y_i, Y_j) = \boldsymbol{\xi}_i^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_j = \boldsymbol{\xi}_i^\top \left(\sum_{k=1}^{p} \lambda_k \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top\right) \boldsymbol{\xi}_j$$

Proof.

由定理1可知，对$\forall \boldsymbol{x} \in R^p$,有

$$D(Y_1) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_1^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_1}{\boldsymbol{\xi}_1^\top \boldsymbol{\xi}_1} = \lambda_1$$

类似地可得:

$$D(Y_{k+1}) = \max_{\boldsymbol{x} \neq \boldsymbol{0}; \boldsymbol{x}^\top \boldsymbol{\xi}_i = \boldsymbol{0}, i=1,2,\cdots,k} \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_{k+1}}{\boldsymbol{\xi}_{k+1}^\top \boldsymbol{\xi}_{k+1}} = \lambda_{k+1}$$

$$Cov(Y_i, Y_j) = \boldsymbol{\xi}_i^\top \boldsymbol{\Sigma} \boldsymbol{\xi}_j = \boldsymbol{\xi}_i^\top \left(\sum_{k=1}^{p} \lambda_k \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top\right) \boldsymbol{\xi}_j = \sum_{k=1}^{p} \lambda_k \boldsymbol{\xi}_i^\top \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\top \boldsymbol{\xi}_j$$

□

基本原理

总结：主成分的系数向量即为协差阵的标准正交特征向量，重根按重数计算。

基本原理

总结：主成分的系数向量即为协差阵的标准正交特征向量，重根按重数计算。

无论协差阵$\boldsymbol{\Sigma}$的各特征根是否相等，对应的标准化特征向量$\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p$总是存在，进一步施以正交化（如施密特正交化法）。这样，求主成分的问题就转化为求特征根与特征向量。

第四节：主成分的性质

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质1：主成分向量$\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1, \lambda_2, \cdots, \lambda_p) \triangleq \boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_{ii}$。

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$, 有$\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p)$, 则$\boldsymbol{\Sigma}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

性质1：主成分向量$\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1, \lambda_2, \cdots, \lambda_p) \triangleq \boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p)$，则$\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p} \sigma_{ii} = tr(\boldsymbol{\Sigma})$$

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$, 有$\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p)$, 则$\boldsymbol{\Sigma}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p}\sigma_{ii}=tr(\boldsymbol{\Sigma})=tr(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}})$$

第四节：主成分的性质

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p)$，则$\boldsymbol{\Sigma}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p}\sigma_{ii}=tr(\boldsymbol{\Sigma})=tr(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}})$$

第四节：主成分的性质

性质1：主成分向量$\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1, \lambda_2, \cdots, \lambda_p) \triangleq \boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p)$，则$\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p} \sigma_{ii} = tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}) = tr(\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}) = tr(\boldsymbol{\Lambda})$$

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p)$，则$\boldsymbol{\Sigma}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p}\sigma_{ii}=tr(\boldsymbol{\Sigma})=tr(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda})=\sum_{i=1}^{p}\lambda_i$$

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p)$，则$\boldsymbol{\Sigma}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p}\sigma_{ii}=tr(\boldsymbol{\Sigma})=tr(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda})=\sum_{i=1}^{p}\lambda_i$$

注2：意义说明。$\boldsymbol{\Sigma}$的主对角线元素$\sigma_{11},\sigma_{22},\cdots,\sigma_{pp}$分别是$\boldsymbol{X}$各分量$X_1,X_2,\cdots,X_p$的方差。

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p)$，则$\boldsymbol{\Sigma}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p}\sigma_{ii}=tr(\boldsymbol{\Sigma})=tr(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda})=\sum_{i=1}^{p}\lambda_i$$

注2：意义说明。$\boldsymbol{\Sigma}$的主对角线元素$\sigma_{11},\sigma_{22},\cdots,\sigma_{pp}$分别是$\boldsymbol{X}$各分量$X_1,X_2,\cdots,X_p$的方差。即$\sum_{i=1}^{p}\sigma_{ii}$刻画了原$p$个变量$X_1,X_2,\cdots,X_p$所携带的信息总量。

性质1：主成分向量$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathsf{T}}$的协差阵为对角阵$diag(\lambda_1,\lambda_2,\cdots,\lambda_p)\triangleq\boldsymbol{\Lambda}$。

性质2：记$\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$，有$\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

注1：关于证明。记$\boldsymbol{Q}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p)$，则$\boldsymbol{\Sigma}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$，且有

$$\sum_{i=1}^{p}\sigma_{ii}=tr(\boldsymbol{\Sigma})=tr(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}})=tr(\boldsymbol{\Lambda})=\sum_{i=1}^{p}\lambda_i$$

注2：意义说明。$\boldsymbol{\Sigma}$的主对角线元素$\sigma_{11},\sigma_{22},\cdots,\sigma_{pp}$分别是$\boldsymbol{X}$各分量$X_1,X_2,\cdots,X_p$的方差。即$\sum_{i=1}^{p}\sigma_{ii}$刻画了原$p$个变量$X_1,X_2,\cdots,X_p$所携带的信息总量。而$\sum_{i=1}^{p}\lambda_i$为$p$个主成分$Y_1,Y_2,\cdots,Y_p$所携带的信息量之和。即主成分数等于变量数时，无信息损失。

主成分的性质

定义1

第k个主成分Y_k的信息量在全部信息量中所占的比例，$\lambda_k / \sum_{i=1}^{p} \lambda_i$称为第$k$个主成分的贡献率。前$k$个主成分的贡献率之和$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$称为前$k$个主成分$Y_1, Y_2, \cdots, Y_k$的累积贡献率。

主成分的性质

定义1

第k个主成分Y_k的信息量在全部信息量中所占的比例，$\lambda_k / \sum_{i=1}^{p} \lambda_i$称为第$k$个主成分的贡献率。前$k$个主成分的贡献率之和$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$称为前$k$个主成分$Y_1, Y_2, \cdots, Y_k$的累积贡献率。

注1：主成分的贡献率越大，可认为其综合能力越强。

主成分的性质

定义1

第k个主成分Y_k的信息量在全部信息量中所占的比例，$\lambda_k / \sum_{i=1}^{p} \lambda_i$称为第$k$个主成分的**贡献率**。前$k$个主成分的贡献率之和$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$称为前$k$个主成分$Y_1, Y_2, \cdots, Y_k$的**累积贡献率**。

注1：主成分的贡献率越大，可认为其综合能力越强。

注2：若前k个主成分的累积贡献率大于85%，可认为前k个主成分已综合了原始变量的大部分信息，可不再提取新的主成分了。

主成分的性质

定义2

主成分Y_k与原始变量X_i的相关系数$\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} l_{ki}, k, i = 1, 2, \cdots, p$, 称之为$Y_k$关于$X_i$的因子负荷量。

主成分的性质

定义2

主成分Y_k与原始变量X_i的相关系数$\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} l_{ki}, k, i = 1, 2, \cdots, p,$ 称之为Y_k关于X_i的因子负荷量。

注1：关于证明。记$\boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^\mathsf{T}$, 则

定义2

主成分Y_k与原始变量X_i的相关系数$\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} l_{ki}, k, i = 1, 2, \cdots, p,$ 称之为Y_k关于X_i的因子负荷量。

注1：关于证明。记$\boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^\top$, 则

$$\begin{aligned}\rho(Y_k, X_i) =& \frac{Cov(Y_k, X_i)}{\sqrt{Var(Y_k)}\sqrt{Var(X_i)}} = \frac{Cov(\boldsymbol{l}_k^\top \boldsymbol{X}, \boldsymbol{e}_i^\top \boldsymbol{X})}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}} \\ =& \frac{\boldsymbol{l}_k^\top \boldsymbol{\Sigma} \boldsymbol{e}_i}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}} = \frac{\lambda_k \boldsymbol{l}_k^\top \boldsymbol{e}_i}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}} = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} l_{ki}\end{aligned}$$

注2：意义说明。$\rho(Y_k, X_i)$的绝对值越大，表示Y_k所反映的信息与X_i越密切。如果$\rho(Y_k, X_i)$符号为正，说明主成分Y_k与X_i正相关，二者变化趋势相同；否则为负相关，变化趋势相反。

主成分的性质

注2：意义说明。$\rho(Y_k, X_i)$的绝对值越大，表示Y_k所反映的信息与X_i越密切。如果$\rho(Y_k, X_i)$符号为正，说明主成分Y_k与X_i正相关，二者变化趋势相同；否则为负相关，变化趋势相反。

注3：主成分Y_k与原变量X_i的线性组合系数l_{ki}, 表达了原变量X_i每增减一个单位，主成分Y_k相应的增减量。因此可以通过观察各组合系数$l_{k1}, l_{k2}, \cdots, l_{kp}$的符号、大小，并结合原变量$X_i$的实际含义，对主成分$Y_k$的综合含义做出解释并命名。

主成分的性质

性质3：$\sum_{k=1}^{p} \rho^2(Y_k, X_i) = 1$。

主成分的性质

性质3：$\sum_{k=1}^{p}\rho^2(Y_k, X_i) = 1$。

注1：关于证明。由前面分析可知，第i个主成分的系数向量$\boldsymbol{l}_i$，为第i大特征根所对应的单位正交特征向量ξ_i，$i = 1, 2, \cdots, p$。一方面，

$$\boldsymbol{l}_i^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{l}_i = (l_{i1}, l_{i2}, \cdots, l_{ip})diag(\lambda_1, \lambda_2, \cdots, \lambda_p)(l_{i1}, l_{i2}, \cdots, l_{ip})^{\mathsf{T}} = \sum_{k=1}^{p}\lambda_k l_{ki}^2$$

主成分的性质

性质3：$\sum_{k=1}^{p}\rho^2(Y_k, X_i) = 1$。

注1：关于证明。由前面分析可知，第i个主成分的系数向量$\boldsymbol{l}_i$，为第i大特征根所对应的单位正交特征向量ξ_i，$i = 1, 2, \cdots, p$。一方面，

$$\boldsymbol{l}_i^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{l}_i = (l_{i1}, l_{i2}, \cdots, l_{ip})diag(\lambda_1, \lambda_2, \cdots, \lambda_p)(l_{i1}, l_{i2}, \cdots, l_{ip})^{\mathsf{T}} = \sum_{k=1}^{p}\lambda_k l_{ki}^2$$

另一方面，若记$\boldsymbol{Q} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p), \boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^{\mathsf{T}}$,
则$\boldsymbol{l}_i = \boldsymbol{Q}\boldsymbol{e}_i$并且

$$\boldsymbol{l}_i^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{l}_i = \boldsymbol{e}_i^{\mathsf{T}}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{e}_i = \boldsymbol{e}_i^{\mathsf{T}}\Sigma\boldsymbol{e}_i = \sigma_{ii}$$

性质3：$\sum_{k=1}^{p} \rho^2(Y_k, X_i) = 1$。

注1：关于证明。由前面分析可知，第i个主成分的系数向量$\boldsymbol{l}_i$，为第i大特征根所对应的单位正交特征向量ξ_i，$i = 1, 2, \cdots, p$。一方面，

$$\boldsymbol{l}_i^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{l}_i = (l_{i1}, l_{i2}, \cdots, l_{ip}) diag(\lambda_1, \lambda_2, \cdots, \lambda_p)(l_{i1}, l_{i2}, \cdots, l_{ip})^{\mathsf{T}} = \sum_{k=1}^{p} \lambda_k l_{ki}^2$$

另一方面，若记$\boldsymbol{Q} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p)$, $\boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^{\mathsf{T}}$，
则$\boldsymbol{l}_i = \boldsymbol{Q}\boldsymbol{e}_i$并且

$$\boldsymbol{l}_i^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{l}_i = \boldsymbol{e}_i^{\mathsf{T}} \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{Q} \boldsymbol{e}_i = \boldsymbol{e}_i^{\mathsf{T}} \Sigma \boldsymbol{e}_i = \sigma_{ii}$$

综上有，$\sum_{k=1}^{p} \rho^2(Y_k, X_i) = \frac{1}{\sigma_{ii}} \sum_{k=1}^{p} \lambda_k l_{ki}^2 = 1$。得证。

性质3：$\sum_{k=1}^{p} \rho^2(Y_k, X_i) = 1$。

注1：关于证明。由前面分析可知，第i个主成分的系数向量$\boldsymbol{l}_i$，为第i大特征根所对应的单位正交特征向量ξ_i，$i = 1, 2, \cdots, p$。一方面，

$$\boldsymbol{l}_i^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{l}_i = (l_{i1}, l_{i2}, \cdots, l_{ip}) diag(\lambda_1, \lambda_2, \cdots, \lambda_p)(l_{i1}, l_{i2}, \cdots, l_{ip})^{\mathsf{T}} = \sum_{k=1}^{p} \lambda_k l_{ki}^2$$

另一方面，若记$\boldsymbol{Q} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p)$, $\boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^{\mathsf{T}}$，则$\boldsymbol{l}_i = \boldsymbol{Q}\boldsymbol{e}_i$并且

$$\boldsymbol{l}_i^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{l}_i = \boldsymbol{e}_i^{\mathsf{T}} \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{Q} \boldsymbol{e}_i = \boldsymbol{e}_i^{\mathsf{T}} \Sigma \boldsymbol{e}_i = \sigma_{ii}$$

综上有，$\sum_{k=1}^{p} \rho^2(Y_k, X_i) = \frac{1}{\sigma_{ii}} \sum_{k=1}^{p} \lambda_k l_{ki}^2 = 1$。得证。

主成分的性质

注2：意义说明。既然主成分$\boldsymbol{Y}$是原始变量$\boldsymbol{X}$的线性组合，因此X_i也可以表示成$Y_1, Y_2, \cdots, Y_p$的线性组合。

主成分的性质

注2：意义说明。既然主成分$\boldsymbol{Y}$是原始变量$\boldsymbol{X}$的线性组合，因此X_i也可以表示成$Y_1, Y_2, \cdots, Y_p$的线性组合。从线性回归的角度来看，X_i完全由$Y_1, Y_2, \cdots, Y_p$解释而没有随机扰动，从而拟合优度$R^2 = 1$。也即有，X_i和$Y_1, Y_2, \cdots, Y_p$的全相关系数的平方等于1。

主成分的性质

注2：意义说明。既然主成分$\boldsymbol{Y}$是原始变量$\boldsymbol{X}$的线性组合，因此X_i也可以表示成$Y_1, Y_2, \cdots, Y_p$的线性组合。从线性回归的角度来看，X_i完全由$Y_1, Y_2, \cdots, Y_p$解释而没有随机扰动，从而拟合优度$R^2 = 1$。也即有，X_i和$Y_1, Y_2, \cdots, Y_p$的全相关系数的平方等于1。同时$Y_1, Y_2, \cdots, Y_p$之间互不相关，所以X_i与$Y_1, Y_2, \cdots, Y_p$的全相关系数的平方，也就是X_i与每个解释变量的相关系数的平方和，从而$\sum_{i=1}^{p} \rho^2(Y_k, X_i) = 1$。

注2：意义说明。既然主成分$\boldsymbol{Y}$是原始变量$\boldsymbol{X}$的线性组合，因此X_i也可以表示成$Y_1, Y_2, \cdots, Y_p$的线性组合。从线性回归的角度来看，X_i完全由$Y_1, Y_2, \cdots, Y_p$解释而没有随机扰动，从而拟合优度$R^2 = 1$。也即有，X_i和$Y_1, Y_2, \cdots, Y_p$的全相关系数的平方等于1。同时$Y_1, Y_2, \cdots, Y_p$之间互不相关，所以X_i与$Y_1, Y_2, \cdots, Y_p$的全相关系数的平方，也就是X_i与每个解释变量的相关系数的平方和，从而$\sum_{i=1}^{p} \rho^2(Y_k, X_i) = 1$。

定义3

X_i与前m个主成分$Y_1, Y_2, \cdots, Y_m$的全相关系数的平方和，称为$Y_1, Y_2, \cdots, Y_m$对原始变量X_i的方差贡献率，即

$$v_i = \frac{1}{\sigma_{ii}} \sum_{k=1}^{m} \lambda_k l_{ki}^2, i = 1, 2, \cdots, p$$

定义3

X_i与前m个主成分$Y_1, Y_2, \cdots, Y_m$的全相关系数的平方和，称为$Y_1, Y_2, \cdots, Y_m$对原始变量X_i的方差贡献率，即

$$v_i = \frac{1}{\sigma_{ii}} \sum_{k=1}^{m} \lambda_k l_{ki}^2, i = 1, 2, \cdots, p$$

注1：$v_i \leq 1$，当$m = p$时，$v_i = 1$。

定义3

X_i与前m个主成分$Y_1, Y_2, \cdots, Y_m$的全相关系数的平方和，称为$Y_1, Y_2, \cdots, Y_m$对原始变量X_i的方差贡献率，即

$$v_i = \frac{1}{\sigma_{ii}} \sum_{k=1}^{m} \lambda_k l_{ki}^2, i = 1, 2, \cdots, p$$

注1：$v_i \leq 1$，当$m = p$时，$v_i = 1$。

注2：方差贡献率刻画了，提取的主成分（前m个）表达了原始变量的信息，也即是解释原始变量的能力。

实际问题中，原p个变量$X_1, X_2, \cdots, X_p$可能有不同的量纲，因而可能由于单位选择不合理导致方差相差很悬殊。

第五节：由相关阵提取主成分

实际问题中，原p个变量$X_1, X_2, \cdots, X_p$可能有不同的量纲，因而可能由于单位选择不合理导致方差相差很悬殊。此时通过协差阵提取主成分，会过分关照方差大的变量，而漠视方差小的变量，从而给出不合理的主成分分析结果。

第五节：由相关阵提取主成分

实际问题中，原p个变量$X_1, X_2, \cdots, X_p$可能有不同的量纲，因而可能由于单位选择不合理导致方差相差很悬殊。此时通过协差阵提取主成分，会过分关照方差大的变量，而漠视方差小的变量，从而给出不合理的主成分分析结果。

为了避免这种不合理结果，往往先将各变量标准化，而标准化变量的协差阵正好是原变量$X_1, X_2, \cdots, X_p$的相关阵，由此提取主成分。

由相关阵提取主成分

令$Z_i = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$,$i = 1, 2, \cdots, p$，其中$\mu_i$与$\sigma_{ii}$分别表示变量$X_i$的期望与方差，则$E(Z_i) = 0$, $D(Z_i) = 1$。

由相关阵提取主成分

令$Z_i = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, 2, \cdots, p$，其中$\mu_i$与$\sigma_{ii}$分别表示变量$X_i$的期望与方差，则$E(Z_i) = 0, D(Z_i) = 1$。并令$\boldsymbol{\Sigma}^* = diag(\sigma_{11}, \sigma_{22}, \cdots, \sigma_{pp})$，则$\boldsymbol{Z}$
$= (Z_1, Z_2, \cdots, Z_p)^\mathsf{T} = (\boldsymbol{\Sigma}^*)^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$，

由相关阵提取主成分

令$Z_i = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, 2, \cdots, p$，其中$\mu_i$与$\sigma_{ii}$分别表示变量$X_i$的期望与方差，则$E(Z_i) = 0, D(Z_i) = 1$。并令$\mathbf{\Sigma}^* = diag(\sigma_{11}, \sigma_{22}, \cdots, \sigma_{pp})$，则$\mathbf{Z} = (Z_1, Z_2, \cdots, Z_p)^\mathsf{T} = (\mathbf{\Sigma}^*)^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$，且均值向量$E(\mathbf{Z}) = \mathbf{0}$，协差阵

$$Cov(\mathbf{Z}) = (\mathbf{\Sigma}^*)^{-1/2}\mathbf{\Sigma}(\mathbf{\Sigma}^*)^{-1/2}$$

由相关阵提取主成分

令$Z_i = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, 2, \cdots, p$，其中$\mu_i$与$\sigma_{ii}$分别表示变量$X_i$的期望与方差，则$E(Z_i) = 0, D(Z_i) = 1$。并令$\boldsymbol{\Sigma}^* = diag(\sigma_{11}, \sigma_{22}, \cdots, \sigma_{pp})$，则$\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_p)^\mathsf{T} = (\boldsymbol{\Sigma}^*)^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$，且均值向量$E(\boldsymbol{Z}) = \boldsymbol{0}$，协差阵

$$Cov(\boldsymbol{Z}) = (\boldsymbol{\Sigma}^*)^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^*)^{-1/2} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{pmatrix} \triangleq \boldsymbol{R}$$

由相关阵提取主成分

由相关阵求主成分的过程与主成分个数的确定准则，与由协差阵情形相同。仍用λ_i, $\boldsymbol{\xi}_i$分别表示相关阵$\boldsymbol{R}$的特征值与对应的标准正交特征向量。

由相关阵提取主成分

由相关阵求主成分的过程与主成分个数的确定准则，与由协差阵情形相同。仍用λ_i, $\boldsymbol{\xi}_i$分别表示相关阵$\boldsymbol{R}$的特征值与对应的标准正交特征向量。则主成分与原始变量的关系式为：

$$Y_i = \boldsymbol{\xi}_i^{\mathsf{T}} \boldsymbol{Z} = \boldsymbol{\xi}_i^{\mathsf{T}} (\boldsymbol{\Sigma}^*)^{-1/2} (\boldsymbol{X} - \boldsymbol{\mu}), i = 1, 2, \cdots, p$$

由相关阵提取主成分

性质1：$\sum_{i=1}^{p} D(Y_i) = \sum_{i=1}^{p} D(Z_i) = p$。

由相关阵提取主成分

性质1：$\sum_{i=1}^{p} D(Y_i) = \sum_{i=1}^{p} D(Z_i) = p$。

注1：关于证明：由于$Y_i = \xi_i^{\mathsf{T}} \mathbf{Z}$，则$D(Y_i) = \xi_i^{\mathsf{T}} D(Z_i) \xi_i = \xi_i^{\mathsf{T}} \xi_i = 1$。

由相关阵提取主成分

性质1：$\sum_{i=1}^{p} D(Y_i) = \sum_{i=1}^{p} D(Z_i) = p$。

注1：关于证明：由于$Y_i = \xi_i^{\mathsf{T}} \mathbf{Z}$，则$D(Y_i) = \xi_i^{\mathsf{T}} D(Z_i) \xi_i = \xi_i^{\mathsf{T}} \xi_i = 1$。

注2：关于直观意思。经过标准化后，只要主成分数等于变量数时，信息就没有损失。

由相关阵提取主成分

第k个主成分的方差占总方差的比例，即第k个主成分的方差贡献率为λ_k/p，前k个主成分的累积方差贡献率为$\sum_{i=1}^{k}\lambda_i/p$。公共因子$Y_k$关于$Z_i$的因子负荷量为$\rho(Y_k,Z_i)=\sqrt{\lambda_k}l_{ki}, k,i=1,2,\cdots,p$

第六节：一些细节

主要内容：协差阵或相关阵选取，协差阵或相关阵估计，主成分的解释与命名，主成分分析的应用

协差阵或相关阵选取

- 基于协差阵或相关阵都可求得主成分，二者一般来说有差别，甚至差别很大。

协差阵或相关阵选取

- 基于协差阵或相关阵都可求得主成分，二者一般来说有差别，甚至差别很大。
- 度量单位不同或指标取值范围差异非常大时，由相关阵求主成分。

协差阵或相关阵选取

- 基于协差阵或相关阵都可求得主成分，二者一般来说有差别，甚至差别很大。
- 度量单位不同或指标取值范围差异非常大时，由相关阵求主成分。
- 指标为同度量或取值范围为同量级时，根据协差阵求解主成分。

协差阵或相关阵选取

- 基于协差阵或相关阵都可求得主成分，二者一般来说有差别，甚至差别很大。
- 度量单位不同或指标取值范围差异非常大时，由相关阵求主成分。
- 指标为同度量或取值范围为同量级时，根据协差阵求解主成分。
- 标准化抹杀了原始数据的一部分重要信息，使得标准化后各变量对主成分的作用趋于相等。

协差阵或相关阵估计

以上分析的前提是，随机向量$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathsf{T}}$的协差阵$\boldsymbol{\Sigma}$或相关阵$\boldsymbol{R}$已知。如果二者未知，那么需要先对其进行估计。

协差阵或相关阵估计

以上分析的前提是，随机向量$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathsf{T}}$的协差阵$\boldsymbol{\Sigma}$或相关阵$\boldsymbol{R}$已知。如果二者未知，那么需要先对其进行估计。

设$\boldsymbol{x}_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathsf{T}}$是来自总体$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathsf{T}}$的$n$个个体，$i=1,2,\cdots,n$。则均值向量的估计为

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{(i)}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_p)^{\mathsf{T}}$$

协差阵或相关阵估计

以上分析的前提是，随机向量$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathsf{T}}$的协差阵$\boldsymbol{\Sigma}$或相关阵$\boldsymbol{R}$已知。如果二者未知，那么需要先对其进行估计。

设$\boldsymbol{x}_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathsf{T}}$是来自总体$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathsf{T}}$的$n$个个体，$i=1,2,\cdots,n$。则均值向量的估计为

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{(i)}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_p)^{\mathsf{T}}$$

协差阵估计为

$$\boldsymbol{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathsf{T}}$$

协差阵或相关阵估计

以上分析的前提是，随机向量$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathsf{T}}$的协差阵$\boldsymbol{\Sigma}$或相关阵$\boldsymbol{R}$已知。如果二者未知，那么需要先对其进行估计。

设$\boldsymbol{x}_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathsf{T}}$是来自总体$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathsf{T}}$的$n$个个体，$i=1,2,\cdots,n$。则均值向量的估计为

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{(i)}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_p)^{\mathsf{T}}$$

协差阵估计为

$$\boldsymbol{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathsf{T}}$$

相关阵估计为$\boldsymbol{R}=(r_{ij})_{p\times p}$, 其中$r_{ij}=\dfrac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$。

主成分的解释与命名

提取出的每个主成分Y_k都是原p个变量$X_1,X_2,\cdots,X_p$的线性组合

$$Y_k = \boldsymbol{l}_k^\top \boldsymbol{X} = l_{k1}X_1 + l_{k2}X_2 + \cdots + l_{kp}X_p$$

主成分Y_k可看成是对原变量$X_1,X_2,\cdots,X_p$中某一类信息的综合。

主成分的解释与命名

提取出的每个主成分Y_k都是原p个变量$X_1, X_2, \cdots, X_p$的线性组合

$$Y_k = \boldsymbol{l}_k^\top \boldsymbol{X} = l_{k1}X_1 + l_{k2}X_2 + \cdots + l_{kp}X_p$$

主成分Y_k可看成是对原变量$X_1, X_2, \cdots, X_p$中某一类信息的综合。原变量$X_1, X_2, \cdots, X_p$都有明确的实际含义，那么线性组合后得到的新变量Y_k的含义又是什么？这就是主成分的解释与命名。

主成分的解释与命名

提取出的每个主成分Y_k都是原p个变量$X_1,X_2,\cdots,X_p$的线性组合

$$Y_k = \boldsymbol{l}_k^{\top}\boldsymbol{X} = l_{k1}X_1 + l_{k2}X_2 + \cdots + l_{kp}X_p$$

主成分Y_k可看成是对原变量$X_1,X_2,\cdots,X_p$中某一类信息的综合。原变量$X_1,X_2,\cdots,X_p$都有明确的实际含义，那么线性组合后得到的新变量Y_k的含义又是什么？这就是主成分的解释与命名。

主成分的解释与命名，除了需结合与问题有关的专业知识之外，数学手段是分析主成分与原变量的相关性方向和程度。

主成分的解释与命名

例2

考察某地区上市股票石化板块的五种股票。记X_1, X_2, X_3分别表示三家化工企业的股票回升率，X_4, X_5分别表示两家石油公司的股票回升率。这五项指标虽较详尽地刻画了石化板块在一周内的股票涨跌，但进行长年累月的趋势分析就会感到指标多了。为此进行主成分分析。

例2

考察某地区上市股票石化板块的五种股票。记X_1, X_2, X_3分别表示三家化工企业的股票回升率，X_4, X_5分别表示两家石油公司的股票回升率。这五项指标虽较详尽地刻画了石化板块在一周内的股票涨跌，但进行长年累月的趋势分析就会感到指标多了。为此进行主成分分析。

根据过去一年累计50个交易周这五种股票的周回升率数据，估计出协差阵$\boldsymbol{S}$，然后求其特征根与特征向量，仅列出前两个

$$\lambda_1 = 2.854, \quad \boldsymbol{l}_1^{\mathrm{T}} = (0.464, 0.457, 0.470, 0.421, 0.421)$$
$$\lambda_2 = 0.809, \quad \boldsymbol{l}_2^{\mathrm{T}} = (0.240, 0.509, 0.260, -0.526, -0.582)$$

且$(\lambda_1 + \lambda_2)/\sum_{i=1}^{5} \lambda_i > 0.85$。

主成分的解释与命名

由此得到两个主成分

$$
\begin{aligned}
Y_1 =& 0.464X_1 + 0.457X_2 + 0.470X_3 + 0.421X_4 + 0.421X_5 \\
Y_2 =& 0.240X_1 + 0.509X_2 + 0.260X_3 - 0.526X_4 - 0.582X_5
\end{aligned}
$$

主成分的解释与命名

由此得到两个主成分

$$
\begin{aligned}
Y_1 &= 0.464X_1 + 0.457X_2 + 0.470X_3 + 0.421X_4 + 0.421X_5 \\
Y_2 &= 0.240X_1 + 0.509X_2 + 0.260X_3 - 0.526X_4 - 0.582X_5
\end{aligned}
$$

原来通过五项指标(X_1,X_2,X_3,X_4,X_5)刻画石化板块的股市行情, 现就只需要两项综合性指标(Y_1,Y_2)了。

主成分的解释与命名

由此得到两个主成分

$$
\begin{aligned}
Y_1 =& 0.464X_1 + 0.457X_2 + 0.470X_3 + 0.421X_4 + 0.421X_5 \\
Y_2 =& 0.240X_1 + 0.509X_2 + 0.260X_3 - 0.526X_4 - 0.582X_5
\end{aligned}
$$

原来通过五项指标(X_1,X_2,X_3,X_4,X_5)刻画石化板块的股市行情, 现就只需要两项综合性指标(Y_1,Y_2)了。如已知五支股票的回升率为(3%,2.5%,4%,−1%,1.5%), 则对应的综合性指标为(4.687%,2.686%)。

主成分的解释与命名

由此得到两个主成分

$$
\begin{aligned}
Y_1 &= 0.464X_1 + 0.457X_2 + 0.470X_3 + 0.421X_4 + 0.421X_5 \\
Y_2 &= 0.240X_1 + 0.509X_2 + 0.260X_3 - 0.526X_4 - 0.582X_5
\end{aligned}
$$

原来通过五项指标(X_1,X_2,X_3,X_4,X_5)刻画石化板块的股市行情，现就只需要两项综合性指标(Y_1,Y_2)了。如已知五支股票的回升率为$(3\%,2.5\%,4\%,-1\%,1.5\%)$，则对应的综合性指标为$(4.687\%,2.686\%)$。

Y_1的系数皆为正且大小差异不大。可认为Y_1是石化股票的一种综合指数。该指数越大，表示石化板块的股票涨势越强。

主成分的解释与命名

由此得到两个主成分

$$Y_1 = 0.464X_1 + 0.457X_2 + 0.470X_3 + 0.421X_4 + 0.421X_5$$
$$Y_2 = 0.240X_1 + 0.509X_2 + 0.260X_3 - 0.526X_4 - 0.582X_5$$

原来通过五项指标(X_1,X_2,X_3,X_4,X_5)刻画石化板块的股市行情, 现就只需要两项综合性指标(Y_1,Y_2)了。如已知五支股票的回升率为(3%,2.5%,4%,−1%,1.5%), 则对应的综合性指标为(4.687%,2.686%)。

Y_1的系数皆为正且大小差异不大。可认为Y_1是石化股票的一种综合指数。该指数越大，表示石化板块的股票涨势越强。

Y_2的系数向量的前三个分量（对应化工企业）为正，后两个分量（对应石油公司）为负。可认为Y_2是石化板块中的行业对比指数，该指数越大说明加工相对于生产更受投资者追捧，

由此得到两个主成分

$$
\begin{aligned}
Y_1 &= 0.464X_1 + 0.457X_2 + 0.470X_3 + 0.421X_4 + 0.421X_5 \\
Y_2 &= 0.240X_1 + 0.509X_2 + 0.260X_3 - 0.526X_4 - 0.582X_5
\end{aligned}
$$

原来通过五项指标(X_1,X_2,X_3,X_4,X_5)刻画石化板块的股市行情, 现就只需要两项综合性指标(Y_1,Y_2)了。如已知五支股票的回升率为$(3\%,2.5\%,4\%,-1\%,1.5\%)$, 则对应的综合性指标为$(4.687\%,2.686\%)$。

Y_1的系数皆为正且大小差异不大。可认为Y_1是石化股票的一种综合指数。该指数越大，表示石化板块的股票涨势越强。

Y_2的系数向量的前三个分量（对应化工企业）为正，后两个分量（对应石油公司）为负。可认为Y_2是石化板块中的行业对比指数，该指数越大说明加工相对于生产更受投资者追捧，且X_2的系数明显大于X_1,X_3的系数，故X_2所对应的企业在加工行业中起着龙头作用。

主成分分析的应用

①降维：用少数的主成分替换较多的原始变量；

②主元回归：用主成分作为自变量进行回归。

本章作业

自选一个关注的问题，作为因变量；收集可能的影响因素（多个）作为自变量。对于这些影响因素合理地提取主成分，并将因变量关于这些主成分进行回归，并对比将原始变量直接进行线性回归的效果，尝试解释其中的原因。

要求：限5人组队；需提供数据出处，原始数据；若采用spss等不编程，需注明操作步骤并解释计算结果，严禁直接粘贴软件的图表；若采用matlab, R等编程，需提供代码；提交纸质版，封面为组内成员的姓名和学号。