因子分析

01 多元数据

- 多元数据的数学表达
- 欧氏距离和统计距离

- 在多元数据分析中，所研究的数据是p个指标（变量），n次观测，常用p维随机向量$\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$表示对同一个体观测的$p$个变量。

表1 p维总体的样本资料数据

序号 \ 变量	X_1	X_2		X_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p)=\begin{bmatrix} \boldsymbol{X}'_{(1)} \\ \boldsymbol{X}'_{(2)} \\ \vdots \\ \boldsymbol{X}'_{(n)} \end{bmatrix}$$

- 一维随机变量 X 的数字特征
  - 期望: $\mu = E(X)$
  - 方差: $\sigma^2 = Var(X) = D(X) = V(X)$

设 $x_1, x_2, \cdots, x_n$ 为来自 p 维总体的样本，则有

- 样本均值： $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- 样本方差： $s^2 = \frac{l_{xx}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- 多维随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)'$ 的数字特征

  - 均值向量:

$$E(\boldsymbol{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)'$$

  - 协方差阵:

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{X}, \boldsymbol{X}) = E(\boldsymbol{X} - \boldsymbol{EX})(\boldsymbol{X} - \boldsymbol{EX})' = D(\boldsymbol{X})$$

$$= \begin{bmatrix} D(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_p) \\ \mathrm{cov}(X_2, X_1) & D(X_2) & \cdots & \mathrm{cov}(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(X_p, X_1) & \mathrm{cov}(X_p, X_2) & \cdots & D(X_p) \end{bmatrix} = (\sigma_{ij})$$

若无特别说明，所称向量均指列向量

- 协差阵$\boldsymbol{\Sigma}$既包含了$\boldsymbol{X}$各分量的方差，也包含了每两个分量之间的协方差。显然，$\boldsymbol{\Sigma}$是一个对称矩阵，也是一个非负定阵。

- 随机向量$\boldsymbol{X}$ 和$\boldsymbol{Y}$ 的协差阵
- 设$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$ 和$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_q)'$ 分别为 p维和q维随机向量，它们之间的协方差阵定义为一个 $p\times q$ 矩阵，其元素是 $\mathrm{cov}(X_i,Y_j)$，即

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=(\mathrm{cov}(X_i,Y_j))=(\sigma_{ij})_{p\times q}=\begin{pmatrix}\mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \cdots & \mathrm{Cov}(X_1,Y_q)\\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \cdots & \mathrm{Cov}(X_2,Y_q)\\ \vdots & \vdots & & \vdots\\ \mathrm{Cov}(X_p,Y_1) & \mathrm{Cov}(X_p,Y_2) & \cdots & \mathrm{Cov}(X_p,Y_q)\end{pmatrix}$$

$$=\begin{pmatrix}E\left[X_1-E(X_1)\right]\left[Y_1-E(Y_1)\right] & \cdots & E\left[X_1-E(X_1)\right]\left[Y_q-E(Y_q)\right]\\ \vdots & & \vdots\\ E\left[X_p-E(X_p)\right]\left[Y_1-E(Y_1)\right] & \cdots & E\left[X_p-EX(x_p)\right]\left[Y_q-E(Y_q)\right]\end{pmatrix}$$

$$=E\left(\begin{pmatrix}X_1-E(X_1)\\ \vdots\\ X_p-E(X_p)\end{pmatrix}\left(Y_1-E(Y_1),\cdots,Y_q-E(Y_q)\right)\right)=E\left[\boldsymbol{X}-E(\boldsymbol{X})\right]\left[\boldsymbol{Y}-E(\boldsymbol{Y})\right]'$$

- 随机向量$\boldsymbol{X}$和$\boldsymbol{Y}$的协差阵

  若$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=\boldsymbol{0}$，称$\boldsymbol{X}$和$\boldsymbol{Y}$是不相关的。

- 两个独立的随机向量必然不相关，但两个不相关的随机向量未必独立。

- 随机变量 X 和 Y 的相关系数定义为

$$\rho = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)D(Y)}}$$

- $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)'$ 和 $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_q)'$ 的相关阵定义为

$$\rho(\boldsymbol{X},\boldsymbol{Y}) = \begin{pmatrix} \rho(X_1,Y_1) & \rho(X_1,Y_2) & \cdots & \rho(X_1,Y_q) \\ \rho(X_2,Y_1) & \rho(X_2,Y_2) & \cdots & \rho(X_2,Y_q) \\ \vdots & \vdots & & \vdots \\ \rho(X_p,Y_1) & \rho(X_p,Y_2) & \cdots & \rho(X_p,Y_q) \end{pmatrix}$$

- 若$\rho(\boldsymbol{X}, \boldsymbol{Y}) = 0$，则表明 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 不相关。
- $\boldsymbol{X} = \boldsymbol{Y}$ 时的相关阵$\rho(\boldsymbol{X}, \boldsymbol{X})$称为 $\boldsymbol{X}$ 的相关阵，记作$\boldsymbol{R}=(\rho_{ij})$，这里$\rho_{ij}=\rho(X_i, X_j)$，$\rho_{ii}=1$。

- $\boldsymbol{X}$的相关阵$\boldsymbol{R}=(\rho_{ij})$，即

$$\boldsymbol{R}=\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

- $\boldsymbol{R}$和$\boldsymbol{\Sigma}$的相应元素之间的关系式为

$$\rho_{ij}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

$$\boldsymbol{D}=\operatorname{diag}\left(\sqrt{\sigma_{11}},\sqrt{\sigma_{22}},\cdots,\sqrt{\sigma_{pp}}\right)$$

$$=\begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix}\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix}$$

- $\boldsymbol{R}=(\rho_{ij})$和$\boldsymbol{\Sigma}=(\sigma_{ij})$之间有关系式：　$\boldsymbol{R}=\boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}$

- **标准化变换**
- 最常用的标准化变换是令 $X_i^* = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, \quad i = 1, 2, \cdots, p$

$$\boldsymbol{X}^* = \begin{pmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_p^* \end{pmatrix} = \begin{pmatrix} \dfrac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{pmatrix} = \boldsymbol{D}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

$$E(\boldsymbol{X}^*) = \boldsymbol{D}^{-1}\left[E(\boldsymbol{X}) - \boldsymbol{\mu}\right] = \boldsymbol{0}$$

$$V(\boldsymbol{X}^*) = V\left[\boldsymbol{D}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right] = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1} = \boldsymbol{R}$$

- 相关阵$\boldsymbol{R}$也是一个非负定阵。

- 当$\boldsymbol{A}$、$\boldsymbol{B}$为常数矩阵时，均值向量的性质

(1) $E(\boldsymbol{AX})=\boldsymbol{A}E(\boldsymbol{X})$

(2) $E(\boldsymbol{AXB}+\boldsymbol{C})=\boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}+\boldsymbol{C}$

(3) $E(\boldsymbol{AX}+\boldsymbol{BY})=\boldsymbol{A}E(\boldsymbol{X})+\boldsymbol{B}E(\boldsymbol{Y})$

- 当$\boldsymbol{A}$、$\boldsymbol{B}$为常数矩阵时，协差阵有如下性质：

(1) $D(\boldsymbol{AX}+\boldsymbol{b})=\boldsymbol{A}D(\boldsymbol{X})\boldsymbol{A}'=\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$

(2) $\mathrm{cov}(\boldsymbol{AX},\boldsymbol{BY})=\boldsymbol{A}\,\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}'$

(3) $\boldsymbol{\Sigma}\geq 0$，即随机向量 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 一定是非负定矩阵。

推论　若$|\boldsymbol{\Sigma}|\neq 0$，则$\boldsymbol{\Sigma}>0$。

- 研究多元数据中的变量及其相互关系时，通常以距离和数据多元正态分布假设为基础。距离的平方和多元正态密度可以用称为二次型的矩阵乘积来表示。
- 在多元分析中二次型起着重要作用。

- 设$\boldsymbol{A}$是p阶对称矩阵，$\boldsymbol{x}$是一p维向量，则$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$称为$\boldsymbol{A}$的二次型.
- 若对一切$\boldsymbol{x}\neq\boldsymbol{0}$，有$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}>0$，则称$\boldsymbol{A}$为正定矩阵，记作$\boldsymbol{A}>0$；
- 若对一切$\boldsymbol{x}$，有$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\geq 0$，则称$\boldsymbol{A}$为非负定矩阵，记作$\boldsymbol{A}\geq 0$。
- 对非负定矩阵$\boldsymbol{A}$和$\boldsymbol{B}$，$\boldsymbol{A}>\boldsymbol{B}$表示$\boldsymbol{A}-\boldsymbol{B}>0$；$\boldsymbol{A}\geq\boldsymbol{B}$表示$\boldsymbol{A}-\boldsymbol{B}\geq 0$。

- 多维样本随机向量 $\boldsymbol{X}=(X_1,\ X_2,\ \cdots,\ X_P)'$的数字特征
- 设 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$为来自 p 维总体的样本，其中 $\boldsymbol{x}_k=(x_{k1},x_{k2},\cdots,x_{kp})'$, $k=1,2,\cdots,n.$
- 样本均值向量:

$$\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k=\frac{1}{n}\left[\begin{bmatrix}x_{11}\\x_{12}\\\vdots\\x_{1p}\end{bmatrix}+\begin{bmatrix}x_{21}\\x_{22}\\\vdots\\x_{2p}\end{bmatrix}+\cdots+\begin{bmatrix}x_{n1}\\x_{n2}\\\vdots\\x_{np}\end{bmatrix}\right]=\frac{1}{n}\begin{bmatrix}x_{11}+x_{21}+\cdots+x_{n1}\\x_{12}+x_{22}+\cdots+x_{n2}\\\vdots\\x_{1p}+x_{2p}+\cdots+x_{np}\end{bmatrix}=\frac{1}{n}\begin{bmatrix}\sum_{i=1}^{n}x_{i1}\\\sum_{i=1}^{n}x_{i2}\\\vdots\\\sum_{i=1}^{n}x_{ip}\end{bmatrix}=\begin{bmatrix}\bar{x}_1\\\bar{x}_2\\\vdots\\\bar{x}_p\end{bmatrix}$$

$$\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k==\frac{1}{n}\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&&\vdots\\x_{n1}&x_{n2}&\cdots&x_{np}\end{bmatrix}'\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}=\frac{1}{n}\boldsymbol{X}'\boldsymbol{1}_n=\begin{pmatrix}\bar{x}_1\\\bar{x}_2\\\vdots\\\bar{x}_p\end{pmatrix}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_p)'$$

- 多维样本随机向量 $\boldsymbol{X}=(X_1,\ X_2,\ \cdots,\ X_P)'$的数字特征
- 设 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$为来自 p 维总体的样本，其中 $\boldsymbol{x}_k=(x_{k1},x_{k2},\cdots,x_{kp})'$, $k=1,2,\cdots,n.$
  - 样本协方差阵:

$$\hat{\boldsymbol{\Sigma}}_p=\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{L}=\frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{x}_k-\bar{\boldsymbol{x}})(\boldsymbol{x}_k-\bar{\boldsymbol{x}})'$$

  - 其中 $\boldsymbol{L}$ 是离差阵(样本叉积阵)，它是每一个样品（向量）与样本均值（向量）的离差积形成的n个 $p\times p$ 阶对称阵的和。

$$\boldsymbol{L}=\sum_{k=1}^{n}(\boldsymbol{x}_k-\bar{\boldsymbol{x}})(\boldsymbol{x}_k-\bar{\boldsymbol{x}})'=\boldsymbol{X}'\boldsymbol{X}-\frac{1}{n}\boldsymbol{X}'\mathbf{1}_n\mathbf{1}_n'\boldsymbol{X}=\boldsymbol{X}'(\boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{X}$$

- $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$ 和$\boldsymbol{y}=(y_1,y_2,\cdots,y_p)'$ 之间的欧氏距离为

$$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_p-y_p)^2}$$

- 平方欧氏距离为

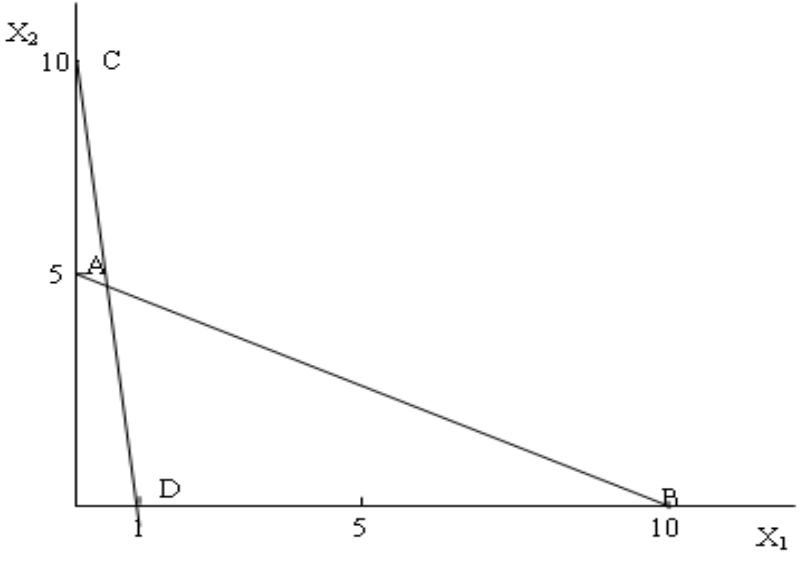
$$d^2(\boldsymbol{x},\boldsymbol{y})=(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_p-y_p)^2$$

$$=(x_1-y_1,\cdots,x_p-y_p)\begin{pmatrix}x_1-y_1\\ \vdots\\ x_p-y_p\end{pmatrix}=(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})$$

- 就大部分统计问题而言，欧氏距离是不能令人满意的。
- 欧氏距离有一个缺点：当各个分量为不同性质的量时，“距离”的大小与指标的单位有关。

不适合直接使用欧氏距离的例子

例如，横轴 x_1 代表重量（以kg为单位），纵轴 x_2 代表长度（以cm为单位）。有四个点A、B、C、D，它们的坐标如图所示：



这时

$$AB=\sqrt{5^2+10^2}=\sqrt{125}$$

$$CD=\sqrt{10^2+1^2}=\sqrt{101}$$

现在，如果 x_2 用*mm*作单位，x_1 单位保持不变，此时A坐标为（0，50），C坐标为（0，100），则

$$AB=\sqrt{50^2+10^2}=\sqrt{2600}$$

$$CD=\sqrt{100^2+1^2}=\sqrt{10001}$$

结果*CD*反而比*AB*长，是不够合理的。

不适合直接使用欧氏距离的例子

- 下面是各国家和地区男子径赛记录的数据：

国家和地区	100米（秒）	200米（秒）	400米（秒）	800米（分）	1500米（分）	5000米（分）	10000米（分）	马拉松（分）
阿根廷	10.39	20.81	46.84	1.81	3.7	14.04	29.36	137.72
澳大利亚	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.3
奥地利	10.44	20.81	46.82	1.79	3.6	13.26	27.72	135.9
比利时	10.34	20.68	45.04	1.73	3.6	13.22	27.45	129.95
百慕大	10.28	20.58	45.91	1.8	3.75	14.68	30.55	146.62
巴西	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
缅甸	10.64	21.52	48.3	1.8	3.85	14.45	30.28	139.95
加拿大	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
智利	10.34	20.8	46.2	1.79	3.71	13.61	29.3	134.03
中国	10.51	21.04	47.3	1.81	3.73	13.9	29.13	133.53
哥伦比亚	10.43	21.05	46.1	1.82	3.74	13.49	27.88	131.35
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- 在实际应用中，为了消除单位的影响和均等地对待每一分量，常须先对各分量作标准化变换，然后再计算欧氏距离。
- 令 $x_i^* = \dfrac{x_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, \cdots, p, \quad \boldsymbol{x}^* = \left(x_1^*, \cdots, x_p^*\right)'$，则

$$d^2\left(\boldsymbol{x}^*, \mathbf{0}\right) = \boldsymbol{x}^{*\prime}\boldsymbol{x}^* = x_1^{*2} + \cdots + x_p^{*2}$$

- 由于 $E\left(x_i^{*2}\right) = V\left(x_i^*\right) = 1, i = 1, 2, \cdots, p$，故平方和中各分量所起的平均作用都一样。
- 欧氏距离经变量的标准化之后能够消除各变量的单位或方差差异的影响，但不能消除变量之间相关性的影响。
- 有必要建立一种距离，要能够体现各个变量在变差大小上的不同，还要求距离与各变量所用的单位无关，以及有时存在着的相关性。选择的距离要依赖于样本方差和协方差，采用“统计距离”这个术语，以区别通常习惯用的欧氏距离。最常用的一种统计距离是印度统计学家马哈拉诺比斯（*Mahalanobis*）于1936年引入的距离，称为“马氏距离”。

- $\boldsymbol{x}=\left(x_1,x_2,\cdots,x_p\right)'$ 和$\boldsymbol{y}=\left(y_1,y_2,\cdots,y_p\right)'$之间的平方马氏距离定义为

$$d^2\left(\boldsymbol{x},\boldsymbol{y}\right)=\left(\boldsymbol{x}-\boldsymbol{y}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{y}\right)$$

- $\boldsymbol{x}=\left(x_1,x_2,\cdots,x_p\right)'$到总体$\pi$的平方马氏距离定义为

$$d^2\left(\boldsymbol{x},\pi\right)=\left(\boldsymbol{x}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)$$

- 特点(1) 马氏距离不受变量单位的影响，是一个无单位的数值。

带有常数项的单位变换

- 例　摄氏温度与华氏温度的换算公式：

$$F=(C\times 9/5)+32,\quad C=(F-32)\times 5/9$$

式中F——华氏温度，C——摄氏温度。

-

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}=\begin{pmatrix} c_1x_1+b_1 \\ c_2x_2+b_2 \\ \vdots \\ c_px_p+b_p \end{pmatrix}=\begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & c_p \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}+\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

$$\boldsymbol{y}=\boldsymbol{C}\boldsymbol{x}+\boldsymbol{b}$$

- 特点(2) 马氏距离是$\boldsymbol{x}$和$\boldsymbol{y}$经"标准化"之后的欧氏距离，即

$$d^2(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}^*-\boldsymbol{y}^*)'(\boldsymbol{x}^*-\boldsymbol{y}^*)$$

其中$\boldsymbol{x}^*=\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}),\boldsymbol{y}^*=\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu})$，它们的均值皆为$\boldsymbol{0}$，协差阵皆为单位阵$\boldsymbol{I}$。

- 特点(3) 若 $\boldsymbol{\Sigma}=\mathrm{diag}(\sigma_{11},\sigma_{22},\cdots,\sigma_{pp})$，则

$$d^2(\boldsymbol{x},\boldsymbol{y})=\frac{(x_1-y_1)^2}{\sigma_{11}}+\frac{(x_2-y_2)^2}{\sigma_{22}}+\cdots+\frac{(x_p-y_p)^2}{\sigma_{pp}}$$

即当各分量不相关时马氏距离即为各分量经标准化后的欧氏距离。

02

因子分析

- 正交因子模型
- 参数估计
- 因子旋转
- 因子得分

- 因子分析是一种数据简化的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。原始的变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因子。
- 因子分析的目的和用途与主成分分析类似，它也是一种降维方法。由于因子往往比主成分更易得到解释，故因子分析比主成分分析更容易成功，从而有更广泛的应用。
- 因子分析起源于20世纪初，K.皮尔逊(Pearson)和C.斯皮尔曼(Spearman)等学者为定义和测定智力所作的努力，主要是由对心理测量学有兴趣的科学家们培育和发展了因子分析。

- 例1　对运动员的奥林匹克十项全能比赛的得分作了因子分析研究。

x_1：100米跑	x_6：11米跨栏
x_2：跳远	x_7：铁饼
x_3：铅球	x_8：撑杆跳高
x_4：跳高	x_9：标枪
x_5：400米跑	x_{10}：1500米跑

- 经标准化后所作的因子分析表明，十项得分基本上可归结于他们的爆发性臂力强度、短跑速度、爆发性腿部强度和跑的耐力这四个方面，每一方面都称为一个因子。十项得分与这四个因子之间的关系可以描述为如下的因子模型：

$$x_i=\mu_i+a_{i1}f_1+a_{i2}f_2+a_{i3}f_3+a_{i4}f_4+\varepsilon_i,\ \ i=1,2,\cdots,10$$

其中f_1, f_2, f_3, f_4表示四个因子，称为公共因子(common factor), a_{ij}称为x_i在因子f_j上的载荷(loading)，μ_i是x_i的均值，ε_i是x_i不能被四个公共因子解释的部分，称之为特殊因子(specific factor)。

- 因子分析与主成分分析主要有如下一些区别：
  ① 主成分分析涉及的只是一般的变量变换，它不能作为一个模型来描述，本质上几乎不需要任何假定；而因子分析需要构造一个因子模型，并伴有几个关键性的假定。
  ② 主成分是原始变量的线性组合；而在因子分析中，原始变量是因子的线性组合，但因子却一般不能表示为原始变量的线性组合。

主成分分析：$x_1, x_2, \cdots, x_p \rightarrow y_1, y_2, \cdots, y_m$

$$y_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p = \boldsymbol{a}_1'\boldsymbol{x}$$
$$y_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p = \boldsymbol{a}_2'\boldsymbol{x}$$
$$\vdots$$
$$y_m = a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{pm}x_p = \boldsymbol{a}_m'\boldsymbol{x}$$

因子分析：$x_1, x_2, \cdots, x_p \rightarrow f_1, f_2, \cdots, f_m$

$$\begin{cases} x_1 = \mu_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = \mu_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \vdots \\ x_p = \mu_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases}$$

- 数学模型
- 正交因子模型的性质
- 因子载荷矩阵的统计意义

数学模型

- 设有p维可观测的随机向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$，其均值为 $\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_p)'$，协差阵为$\boldsymbol{\Sigma}=(\sigma_{ij})$。
- 因子分析的一般模型为

$$\begin{cases} x_1=\mu_1+a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1 \\ x_2=\mu_2+a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2 \\ \vdots \\ x_p=\mu_p+a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p \end{cases}$$

其中$f_1,f_2,\cdots,f_m$为公共因子，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p$为特殊因子，它们都是不可观测的随机变量。公共因子出现在每一个原始变量的表达式中，可理解为原始变量共同具有的公共因素。

- 上式可用矩阵表示为：　　$\boldsymbol{x}=\boldsymbol{\mu}+\boldsymbol{A}\boldsymbol{f}+\boldsymbol{\varepsilon}$

$x=\mu+Af+\varepsilon$

- 式中 $\boldsymbol{f}=(f_1,f_2,\cdots,f_m)'$为公共因子向量，$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p)'$为特殊因子向量，$\boldsymbol{A}=(a_{ij}):p\times m$称为因子载荷矩阵。通常假定

$$\begin{cases} E(\boldsymbol{f})=\boldsymbol{0} \\ E(\boldsymbol{\varepsilon})=\boldsymbol{0} \\ V(\boldsymbol{f})=\boldsymbol{I} \\ V(\boldsymbol{\varepsilon})=\boldsymbol{D}=\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2) \\ \mathrm{Cov}(\boldsymbol{f},\boldsymbol{\varepsilon})=E(\boldsymbol{f}\boldsymbol{\varepsilon}')=\boldsymbol{0} \end{cases}$$

- 该假定和上述关系式构成了正交因子模型。
- 由上述假定可以看出，公共因子彼此不相关且具有单位方差，特殊因子也彼此不相关且和公共因子也不相关。

正交因子模型的性质

- $\boldsymbol{x}$的协差阵$\boldsymbol{\Sigma}$的分解
- 模型不受单位的影响
- 因子载荷是不惟一的

x的协差阵Σ的分解

$$\boldsymbol{\Sigma} = V(\boldsymbol{Af} + \boldsymbol{\varepsilon}) = V(\boldsymbol{Af}) + V(\boldsymbol{\varepsilon}) = \boldsymbol{A}V(\boldsymbol{f})\boldsymbol{A}' + V(\boldsymbol{\varepsilon}) = \boldsymbol{AA}' + \boldsymbol{D}$$

- 如果$\boldsymbol{A}$只有少数几列，则上述分解式揭示了$\boldsymbol{\Sigma}$的一个简单结构。
- 由于$\boldsymbol{D}$是对角矩阵，故$\boldsymbol{\Sigma}$的非对角线元素可由$\boldsymbol{A}$的元素确定，即因子载荷完全决定了原始变量之间的协方差。
- 如果$\boldsymbol{x}$为各分量已标准化了的随机向量，则$\boldsymbol{\Sigma}$就是相关阵$\boldsymbol{R}$，即有

$$\boldsymbol{R} = \boldsymbol{AA}' + \boldsymbol{D}$$

- 例2 设随机向量$\boldsymbol{x}=(x_1,x_2,x_3,x_4)'$的协方差矩阵为

$$\boldsymbol{\Sigma}=\begin{pmatrix} 9 & -11 & -5 & 20 \\ -11 & 27 & 17 & -42 \\ -5 & 17 & 52 & -5 \\ 20 & -42 & -5 & 86 \end{pmatrix}$$

则$\boldsymbol{\Sigma}$可分解为

$$\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{D}$$

其中

$$\boldsymbol{A}=\begin{pmatrix} 2 & -1 \\ -4 & 3 \\ 1 & 7 \\ 9 & -2 \end{pmatrix}, \quad \boldsymbol{D}=\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

模型不受单位的影响

- 将$\boldsymbol{x}$的单位作变化，通常是作一变换 $\boldsymbol{x}^* = \boldsymbol{Cx}$，这里

$$\boldsymbol{C} = \mathrm{diag}(c_1, c_2, \cdots, c_p),\ c_i > 0,\ i = 1, 2, \cdots, p,$$

于是 $\boldsymbol{x}^* = \boldsymbol{C\mu} + \boldsymbol{Caf} + \boldsymbol{C\varepsilon}$

令 $\boldsymbol{\mu}^* = \boldsymbol{C\mu}$，$\boldsymbol{A}^* = \boldsymbol{CA}$，$\boldsymbol{\varepsilon}^* = \boldsymbol{C\varepsilon}$，则有 $\boldsymbol{x}^* = \boldsymbol{\mu}^* + \boldsymbol{A}^*\boldsymbol{f} + \boldsymbol{\varepsilon}^*$

这个模型能满足类似于前述因子模型的假定，即

$$E(\boldsymbol{f}) = \boldsymbol{0},\ E(\boldsymbol{\varepsilon}^*) = \boldsymbol{0}$$

$$V(\boldsymbol{f}) = \boldsymbol{I},\ V(\boldsymbol{\varepsilon}^*) = \boldsymbol{D}^*$$

$$\mathrm{Cov}(\boldsymbol{f}, \boldsymbol{\varepsilon}^*) = \mathrm{Cov}(\boldsymbol{f}, \boldsymbol{\varepsilon})\boldsymbol{C}' = \boldsymbol{0}$$

$$\boldsymbol{D}^* = \mathrm{diag}(\sigma_1^{*2}, \sigma_2^{*2}, \cdots, \sigma_p^{*2}),\ \sigma_i^{*2} = c_i^2\sigma_i^2,\ i = 1, 2, \cdots, p$$

因子载荷是不惟一的

- 设$\boldsymbol{T}$为任一$m\times m$正交矩阵，令$\boldsymbol{A}^*=\boldsymbol{AT}$，$\boldsymbol{f}^*=\boldsymbol{T}'\boldsymbol{f}$，则模型能表示为

$$\boldsymbol{x}=\boldsymbol{\mu}+\boldsymbol{A}^*\boldsymbol{f}^*+\boldsymbol{\varepsilon}$$

因为

$$E(\boldsymbol{f}^*)=\boldsymbol{T}'E(\boldsymbol{f})=\boldsymbol{0}$$

$$V(\boldsymbol{f}^*)=\boldsymbol{T}'V(\boldsymbol{f})\boldsymbol{T}=\boldsymbol{T}'\boldsymbol{T}=\boldsymbol{I}$$

$$\mathrm{Cov}(\boldsymbol{f}^*,\boldsymbol{\varepsilon})=E(\boldsymbol{f}^*\boldsymbol{\varepsilon}')=\boldsymbol{T}'E(\boldsymbol{f}\boldsymbol{\varepsilon}')=\boldsymbol{0}$$

所以仍满足模型条件。$\boldsymbol{\Sigma}$也可分解为

$$\boldsymbol{\Sigma}=\boldsymbol{A}^*\boldsymbol{A}^{*\prime}+\boldsymbol{D}$$

- 因此，因子载荷矩阵$\boldsymbol{A}$不是惟一的，在实际应用中常常利用这一点，通过因子的旋转，使得新的因子有更好的实际意义。

因子载荷矩阵的统计意义

- $\boldsymbol{A}$的元素a_{ij}
- $\boldsymbol{A}$的行元素平方和 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$
- $\boldsymbol{A}$的列元素平方和 $g_j^2 = \sum_{i=1}^{p} a_{ij}^2$

A的元素a_{ij}

- $$x_i = \mu_i + a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{im}f_m + \varepsilon_i$$

$$\mathrm{Cov}(x_i, f_j) = \sum_{\alpha=1}^{m} a_{i\alpha}\,\mathrm{Cov}(f_\alpha, f_j) + \mathrm{Cov}(\varepsilon_i, f_j) = a_{ij}$$

即a_{ij}是x_i与f_j之间的协方差。

- 若$\boldsymbol{x}$为各分量已标准化了的随机向量，则x_i与f_j的相关系数

$$\rho(x_i, f_j) = \frac{\mathrm{Cov}(x_i, f_j)}{\sqrt{V(x_i)V(f_j)}} = \mathrm{Cov}(x_i, f_j) = a_{ij}$$

此时a_{ij}表示x_i与f_j之间的相关系数。

A的行元素平方和

- $$x_i = \mu_i + a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{im}f_m + \varepsilon_i$$

$$V(x_i) = a_{i1}^2 V(f_1) + a_{i2}^2 V(f_2) + \cdots + a_{im}^2 V(f_m) + V(\varepsilon_i)$$

$$= a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \sigma_i^2, \quad i = 1, 2, \cdots, p$$

令 $$h_i^2 = \sum_{j=1}^{m} a_{ij}^2, \quad i = 1, 2, \cdots, p$$

于是 $$\sigma_{ii} = h_i^2 + \sigma_i^2, \quad i = 1, 2, \cdots, p$$

- h_i^2反映了公共因子对x_i的影响，可以看成是公共因子$f_1, f_2, \cdots, f_m$对x_i的方差贡献，称为共性方差(communality)；而σ_i^2是特殊因子ε_i对x_i的方差贡献，称为特殊方差(specific variance)。
- 当$\boldsymbol{x}$为各分量已标准化了的随机向量时，$\sigma_{ii}=1$，此时有

$$h_i^2 + \sigma_i^2 = 1, \quad i = 1, 2, \cdots, p$$

A的列元素平方和 $g_j^2=\sum_{i=1}^{p}a_{ij}^2$

- $$\sum_{i=1}^{p}V(x_i)=\sum_{i=1}^{p}a_{i1}^2V(f_1)+\cdots+\sum_{i=1}^{p}a_{im}^2V(f_m)+\sum_{i=1}^{p}V(\varepsilon_i)$$

$$=g_1^2+\cdots+g_m^2+\sum_{i=1}^{p}\sigma_i^2$$

其中

$$g_j^2=\sum_{i=1}^{p}a_{ij}^2,\quad j=1,2,\cdots,m$$

g_j^2反映了公共因子f_j对$x_1,x_2,\cdots,x_p$的影响，是衡量公共因子f_j重要性的一个尺度，可视为公共因子f_j对$x_1,x_2,\cdots,x_p$的总方差贡献。

- $$\sum_{i=1}^{p}h_i^2=\sum_{i=1}^{p}\sum_{j=1}^{m}a_{ij}^2=\sum_{j=1}^{m}\sum_{i=1}^{p}a_{ij}^2=\sum_{j=1}^{m}g_j^2$$

- 因子分析的参数估计方法有很多，常见的有：主成分法、主因子法、极大似然法等。

主成分法

- 设样本协方差矩阵$\boldsymbol{S}$的特征值依次为$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$，相应的正交单位特征向量为 $\hat{\boldsymbol{t}}_1, \hat{\boldsymbol{t}}_2, \cdots, \hat{\boldsymbol{t}}_p$ 。选取相对较小的因子数m，并使得累计贡献率 $\sum_{i=1}^{m} \hat{\lambda}_i \Big/ \sum_{i=1}^{p} \hat{\lambda}_i$

  达到一个较高的百分比，则$\boldsymbol{S}$可近似分解如下：

$$\boldsymbol{S} = \hat{\lambda}_1 \hat{\boldsymbol{t}}_1 \hat{\boldsymbol{t}}_1' + \cdots + \hat{\lambda}_m \hat{\boldsymbol{t}}_m \hat{\boldsymbol{t}}_m' + \hat{\lambda}_{m+1} \hat{\boldsymbol{t}}_{m+1} \hat{\boldsymbol{t}}_{m+1}' + \cdots + \hat{\lambda}_p \hat{\boldsymbol{t}}_p \hat{\boldsymbol{t}}_p'$$

$$\approx \hat{\lambda}_1 \hat{\boldsymbol{t}}_1 \hat{\boldsymbol{t}}_1' + \cdots + \hat{\lambda}_m \hat{\boldsymbol{t}}_m \hat{\boldsymbol{t}}_m' + \hat{\boldsymbol{D}} = \hat{\boldsymbol{A}} \hat{\boldsymbol{A}}' + \hat{\boldsymbol{D}}$$

  其中

$$\hat{\boldsymbol{A}} = \left(\sqrt{\hat{\lambda}_1} \hat{\boldsymbol{t}}_1, \cdots, \sqrt{\hat{\lambda}_m} \hat{\boldsymbol{t}}_m \right) = \left(\hat{a}_{ij} \right) \text{为} p \times m \text{矩阵，} \quad \hat{\boldsymbol{D}} = \operatorname{diag}\left(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_p^2 \right),$$

$\hat{\sigma}_i^2 = s_{ii} - \sum_{j=1}^{m} \hat{a}_{ij}^2$ ，$i=1,2,\cdots,p$。这里的 $\hat{\boldsymbol{A}}$和 $\hat{\boldsymbol{D}}$就是因子模型的一个主成分分解.

- 对主成分解，当因子数增加时，原来因子的估计载荷并不变，第j个因子f_j对$\boldsymbol{x}$的总方差贡献仍为 $\hat{\lambda}_i$。
- 主成分法与主成分分析有着很相似的名称，两者很容易混淆。虽然第j个因子与第j个主成分的解释完全相同，但主成分法与主成分分析本质上却是两个不同的概念。主成分法是因子分析中的一种参数估计方法，它并不计算任何主成分，且旋转后的因子解释一般就与主成分明显不同了。
- 称 $\boldsymbol{S}-\left(\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'+\hat{\boldsymbol{D}}\right)$ 为残差矩阵，对于主成分解，有

$$\boldsymbol{S}-\left(\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'+\hat{\boldsymbol{D}}\right)\text{的元素平方和}\leq\hat{\lambda}_{m+1}^2+\cdots+\hat{\lambda}_p^2$$

- 当p个原始变量的单位不同，或虽单位相同，但各变量的数值变异性相差较大时，应首先对原始变量作标准化变换。

- 例3 在例1中，分别取m=1和m=2，用主成分法估计的因子载荷和共性方差列于表1。

表1 当m=1和m=2时的主成分解

变 量	m=1		m=2		
	因子载荷	共性方差	因子载荷		共性方差
	f_1	$\hat{h}_i^2$	f_1	f_2	$\hat{h}_i^2$
x_1^*：100米	0.817	0.668	0.817	0.531	0.950
x_2^*：200米	0.867	0.752	0.867	0.432	0.939
x_3^*：400米	0.915	0.838	0.915	0.233	0.892
x_4^*：800米	0.949	0.900	0.949	0.012	0.900
x_5^*：1500米	0.959	0.920	0.959	-0.131	0.938
x_6^*：5000米	0.938	0.879	0.938	-0.292	0.965
x_7^*：10000米	0.944	0.891	0.944	-0.287	0.973
x_8^*：马拉松	0.880	0.774	0.880	-0.411	0.943
所解释的总方差的累计比例	0.828		0.828	0.938	

主成分解的近似关系式

$$
\begin{cases}
x_1^*(100米) \approx 0.817 f_1 + 0.531 f_2 + \varepsilon_1 \\
x_2^*(200米) \approx 0.867 f_1 + 0.432 f_2 + \varepsilon_2 \\
x_3^*(400米) \approx 0.915 f_1 + 0.233 f_2 + \varepsilon_3 \\
x_4^*(800米) \approx 0.949 f_1 + 0.012 f_2 + \varepsilon_4 \\
x_5^*(1500米) \approx 0.959 f_1 - 0.131 f_2 + \varepsilon_5 \\
x_6^*(5000米) \approx 0.938 f_1 - 0.292 f_2 + \varepsilon_6 \\
x_7^*(10000米) \approx 0.944 f_1 - 0.287 f_2 + \varepsilon_7 \\
x_8^*(马拉松) \approx 0.880 f_1 - 0.411 f_2 + \varepsilon_8
\end{cases}
$$

- 因子解释。因子f_1代表在径赛项目上的总体实力，可称为强弱因子；因子f_2反映了速度与耐力的对比。

主因子法

- 假定原始向量$\boldsymbol{x}$的各分量已作了标准化变换。如果随机向量$\boldsymbol{x}$满足正交因子模型，则有

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}' + \boldsymbol{D}$$

其中$\boldsymbol{R}$为$\boldsymbol{x}$的相关矩阵，令

$$\boldsymbol{R}^* = \boldsymbol{R} - \boldsymbol{D} = \boldsymbol{A}\boldsymbol{A}'$$

则称$\boldsymbol{R}^*$为$\boldsymbol{x}$的约相关矩阵(reduced correlation matrix)。

- $\boldsymbol{R}^*$中的对角线元素是h_i^2，而不是1，非对角线元素和$\boldsymbol{R}$中是完全一样的，并且$\boldsymbol{R}^*$也是一个非负定矩阵。

- 设 $\hat{\sigma}_i^2$是特殊方差 σ_i^2的一个合适的初始估计，则约相关矩阵可估计为

$$\hat{\boldsymbol{R}}^* = \hat{\boldsymbol{R}} - \hat{\boldsymbol{D}} = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{pmatrix}$$

其中 $\hat{\boldsymbol{R}} = (r_{ij}), \hat{\boldsymbol{D}} = \mathrm{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_p^2)$，$\hat{h}_i^2 = 1 - \hat{\sigma}_i^2$ 是 h_i^2 的初始估计。

又设 $\hat{\boldsymbol{R}}^*$的前m个特征值依次为 $\hat{\lambda}_1^* \geq \hat{\lambda}_2^* \geq \cdots \geq \hat{\lambda}_m^* > 0$，相应的正交单位特征向量为 $\hat{\boldsymbol{t}}_1^*, \hat{\boldsymbol{t}}_2^*, \cdots, \hat{\boldsymbol{t}}_m^*$，则$\boldsymbol{A}$的主因子解为

$$\hat{\boldsymbol{A}} = \left(\sqrt{\hat{\lambda}_1^*}\hat{\boldsymbol{t}}_1^*, \sqrt{\hat{\lambda}_2^*}\hat{\boldsymbol{t}}_2^*, \cdots, \sqrt{\hat{\lambda}_m^*}\hat{\boldsymbol{t}}_m^* \right)$$

由此可以重新估计特殊方差，σ_i^2的最终估计为

$$\hat{\sigma}_i^2 = 1 - \hat{h}_i^2 = 1 - \sum_{j=1}^{m} \hat{a}_{ij}^2, \quad i = 1, 2, \cdots, p$$

- 如果我们希望求得拟合程度更好的解，则可以采用迭代的方法，即利用上式中的$\hat{\sigma}_i^2$再作为特殊方差的初始估计，重复上述步骤，直至解稳定为止。该估计方法称为迭代主因子法。

- 例4　在例1中，取$m=2$，为求得主因子解，选用x_i与其他七个变量的复相关系数平方作为 h_i^2 的初始估计值。计算得

$$\hat{h}_1^2=0.877,\quad \hat{h}_2^2=0.888,\quad \hat{h}_3^2=0.845,\quad \hat{h}_4^2=0.884$$
$$\hat{h}_5^2=0.927,\quad \hat{h}_6^2=0.955,\quad \hat{h}_7^2=0.967,\quad \hat{h}_8^2=0.905$$

于是约相关矩阵为

$$\hat{\boldsymbol{R}}^*=\begin{pmatrix}
0.877 & & & & & & & \\
0.923 & 0.888 & & & & & & \\
0.841 & 0.851 & 0.845 & & & & & \\
0.756 & 0.807 & 0.870 & 0.884 & & & & \\
0.700 & 0.775 & 0.835 & 0.918 & 0.927 & & & \\
0.619 & 0.695 & 0.779 & 0.864 & 0.928 & 0.955 & & \\
0.633 & 0.697 & 0.787 & 0.869 & 0.935 & 0.975 & 0.967 & \\
0.520 & 0.596 & 0.705 & 0.806 & 0.866 & 0.932 & 0.943 & 0.905
\end{pmatrix}$$

$\hat{\boldsymbol{R}}^*$ 的特征值为

$$\hat{\lambda}_1^* = 6.530, \quad \hat{\lambda}_2^* = 0.779, \quad \hat{\lambda}_3^* = 0.051, \quad \hat{\lambda}_4^* = 0.006$$

$$\hat{\lambda}_5^* = -0.014, \quad \hat{\lambda}_6^* = -0.015, \quad \hat{\lambda}_7^* = -0.036, \quad \hat{\lambda}_8^* = -0.053$$

从 $\hat{\lambda}_3^*$起特征值已接近于0，故取m=2，相应的计算结果列于表2。

表2 当m=2时的主因子解

变 量	因子载荷		共性方差
	f_1	f_2	$\hat{h}_i^2$
x_1^*：100米	0.807	0.496	0.897
x_2^*：200米	0.858	0.412	0.906
x_3^*：400米	0.890	0.216	0.856
x_4^*：800米	0.939	0.024	0.881
x_5^*：1500米	0.956	−0.114	0.926
x_6^*：5000米	0.938	−0.282	0.960
x_7^*：10000米	0.946	−0.281	0.974
x_8^*：马拉松	0.874	−0.378	0.907
所解释的总方差的累计比例	0.816	0.914	

因子旋转

- 因子的解释带有一定的主观性，常常通过旋转因子的方法来减少这种主观性且使之更易解释。
- 因子是否易于解释，很大程度上取决于因子载荷矩阵$\boldsymbol{A}$的元素结构。假设$\boldsymbol{A}$是从$\boldsymbol{R}$出发求得的，则有$|a_{ij}|\leq 1$。
- 如果$\boldsymbol{A}$的所有元素都接近0或± 1，则模型的因子就易于解释。这时可将$x_1,x_2,\cdots,x_p$分成m个部分，分别对应$f_1,\cdots,f_m$，这是一种使因子解释大为简化的理想情形，称之为简单结构。
- 因子旋转方法有正交旋转和斜交旋转两类，在这些方法中使用最普遍的是最大方差旋转法(varimax) 。

表3 旋转后的因子载荷估计

变 量	主成分		主因子		极大似然	
	f_1^*	f_2^*	f_1^*	f_2^*	f_1^*	f_2^*
x_1^*：100米	0.274	0.935	0.287	0.903	0.288	0.914
x_2^*：200米	0.376	0.893	0.381	0.872	0.379	0.883
x_3^*：400米	0.543	0.773	0.541	0.751	0.541	0.746
x_4^*：800米	0.712	0.627	0.695	0.631	0.689	0.624
x_5^*：1500米	0.813	0.525	0.799	0.537	0.797	0.532
x_6^*：5000米	0.902	0.389	0.895	0.399	0.899	0.397
x_7^*：10000米	0.903	0.397	0.900	0.405	0.906	0.402
x_8^*：马拉松	0.936	0.261	0.909	0.284	0.914	0.281
所解释的总方差的累计比例	0.523	0.938	0.510	0.914	0.512	0.917

- 三种方法的因子载荷估计经因子旋转之后给出了大致相同的结果，在因子f_1^*上的载荷依次增大，在因子f_2^*上的载荷依次减小，可称f_1^*为耐力因子，称f_2^*为（短跑）速度因子。

- 横轴f_1表示原始变量在因子f_1上的载荷，将主成分解的因子载荷配对$(\hat{a}_{i1},\hat{a}_{i2})$在图中用点表示,在点上标出相应变量的序号。使用最大方差旋转法后，因子按顺时针方向旋转θ, 点i在新坐标系下的坐标为旋转后的因子载荷配对$(\hat{a}_{i1},\hat{a}_{i2})$. 从图中容易直接看出旋转后因子的实际意义.

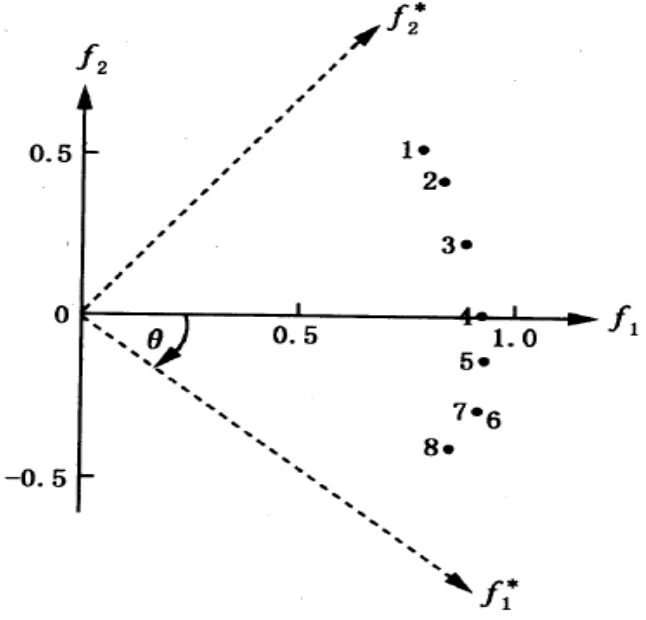


图　主成分解的因子旋转

- 例5 沪市604家上市公司某年财务报表中有这样十个主要财务指标：

x_1：主营业务收入(元)　　x_6：每股净资产(元)

x_2：主营业务利润(元)　　x_7：净资产收益率(%)

x_3：利润总额(元)　　x_8：总资产收益率(%)

x_4：净利润(元)　　x_9：资产总计(元)

x_5：每股收益(元)　　x_{10}：股本

上述十个指标的样本相关矩阵列于表。

表　　十个财务指标的样本相关矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.000									
x_2	0.723	1.000								
x_3	0.427	0.743	1.000							
x_4	0.407	0.697	0.982	1.000						
x_5	0.171	0.325	0.539	0.559	1.000					
x_6	0.149	0.228	0.284	0.274	0.585	1.000				
x_7	0.096	0.177	0.362	0.402	0.776	0.218	1.000			
x_8	0.066	0.204	0.455	0.500	0.849	0.290	0.833	1.000		
x_9	0.748	0.768	0.574	0.567	0.125	0.138	0.067	0.058	1.000	
x_{10}	0.622	0.619	0.485	0.500	0.002	-0.066	0.033	0.051	0.861	1.000

- 从相关矩阵出发，选择主成分法，相关阵的前三个特征值为

$$\hat{\lambda}_1 = 4.879, \quad \hat{\lambda}_2 = 2.574, \quad \hat{\lambda}_3 = 0.929$$

累计贡献率为83.82%，取因子数m=3，相应结果列于表。

表 m=3时的主成分解

变量	因子载荷			共性方差
	f_1	f_2	f_3	$\hat{h}_i^2$
x_1^*：主营业务收入	0.659	−0.472	0.121	0.672
x_2^*：主营业务利润	0.835	−0.346	0.097	0.826
x_3^*：利润总额	0.886	0.003	−0.037	0.786
x_4^*：净利润	0.888	0.037	−0.082	0.796
x_5^*：每股收益	0.666	0.692	0.109	0.934
x_6^*：每股净资产	0.391	0.367	0.814	0.951
x_7^*：净资产收益率	0.527	0.670	−0.325	0.832
x_8^*：总资产收益率	0.581	0.703	−0.260	0.899
x_9^*：资产总计	0.747	−0.564	0.019	0.877
x_{10}^*：股本	0.636	−0.596	−0.219	0.808
所解释的总方差的累计比例	0.488	0.745	0.838	

表 旋转后的因子载荷估计

变量	因子载荷 f_1^*	f_2^*	f_3^*	共性方差 $\hat{h}_i^2$
x_1^*：主营业务收入	0.809	-0.029	0.129	0.672
x_2^*：主营业务利润	0.874	0.171	0.182	0.826
x_3^*：利润总额	0.706	0.509	0.167	0.786
x_4^*：净利润	0.688	0.552	0.135	0.796
x_5^*：每股收益	0.115	0.849	0.447	0.934
x_6^*：每股净资产	0.082	0.199	0.951	0.951
x_7^*：净资产收益率	0.022	0.912	0.004	0.832
x_8^*：总资产收益率	0.045	0.943	0.087	0.899
x_9^*：资产总计	0.936	-0.012	0.028	0.877
x_{10}^*：股本	0.869	-0.013	-0.228	0.808
所解释的总方差的累计比例	0.404	0.712	0.838	

加权最小二乘法

- 采用类似于回归分析中加权最小二乘估计的想法将 $\boldsymbol{f}=(f_1, f_2, \cdots, f_m)'$ 估计为

$$\hat{\boldsymbol{f}}=\left(\boldsymbol{A}'\boldsymbol{D}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\boldsymbol{D}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

称为巴特莱特（Bartlett,1937）因子得分。

- 在实际应用中，用估计值 $\bar{\boldsymbol{x}}, \hat{\boldsymbol{A}}$和$\hat{\boldsymbol{D}}$分别代替上述公式中的$\boldsymbol{\mu}, \boldsymbol{A}$和$\boldsymbol{D}$，并将样品$\boldsymbol{x}_j$的数据代入，便可得到相应的因子得分

$$\hat{\boldsymbol{f}}_j=\left(\hat{\boldsymbol{A}}'\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{A}}\right)^{-1}\hat{\boldsymbol{A}}'\hat{\boldsymbol{D}}^{-1}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})$$

回归法

- 在正交因子模型中，假设 $\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{\varepsilon} \end{pmatrix}$ 服从$(m+p)$元正态分布，

用回归预测方法可将 $\boldsymbol{f}=(f_1, f_2, \cdots, f_m)'$ 估计为

$$\hat{\boldsymbol{f}}=\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

称为汤姆森（Thompson,1951）因子得分。

- 在实际应用中，可用 $\bar{\boldsymbol{x}}, \hat{\boldsymbol{A}}$和$\hat{\boldsymbol{\Sigma}}$ 分别代替上式中的$\boldsymbol{\mu}, \boldsymbol{A}$和$\boldsymbol{\Sigma}$ 来得到因子得分。样品$\boldsymbol{x}_j$的因子得分

$$\hat{\boldsymbol{f}}_j=\hat{\boldsymbol{A}}'\boldsymbol{S}^{-1}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})$$

- 例6 在例5中，用回归法得到的因子得分为

$$\hat{\boldsymbol{f}}^{*} = \hat{\boldsymbol{A}}^{*\prime}\hat{\boldsymbol{R}}^{-1}\boldsymbol{x}^{*}$$

其中 $\boldsymbol{x}^{*} = \left(x_1^{*}, x_2^{*}, \cdots, x_p^{*}\right)'$，$x_i^{*}$ 为x_i的标准化值，i=1,2,⋯,p，经计算：

$$\begin{aligned}
\hat{f}_1^{*} &= 0.217x_1^{*} + 0.216x_2^{*} + 0.145x_3^{*} + 0.138x_4^{*} - 0.054x_5^{*} \\
&\quad -0.032x_6^{*} - 0.066x_7^{*} - 0.066x_8^{*} + 0.254x_9^{*} + 0.246x_{10}^{*} \\
\hat{f}_2^{*} &= -0.109x_1^{*} - 0.043x_2^{*} + 0.116x_3^{*} + 0.144x_4^{*} + 0.235x_5^{*} \\
&\quad -0.165x_6^{*} + 0.381x_7^{*} + 0.371x_8^{*} - 0.086x_9^{*} - 0.016x_{10}^{*} \\
\hat{f}_3^{*} &= 0.100x_1^{*} + 0.098x_2^{*} + 0.004x_3^{*} - 0.037x_4^{*} + 0.216x_5^{*} \\
&\quad +0.876x_6^{*} - 0.229x_7^{*} - 0.157x_8^{*} - 0.008x_9^{*} - 0.255x_{10}^{*}
\end{aligned}$$

- 分别按因子得分的数值大小由高到低排序列于表4，表5和表6，每张表只列出了排在前十位和后十位的股票。

表4 按规模因子得分 $\hat{f}_1^*$ 的排序

序号	股票名称	$\hat{f}_1^*$	$\hat{f}_2^*$	$\hat{f}_3^*$	序号	股票名称	$\hat{f}_1^*$	$\hat{f}_2^*$	$\hat{f}_3^*$
1	上海石化	5.580	-2.704	-2.168	⋮	⋮	⋮	⋮	⋮
2	东方航空	5.446	-2.089	-1.861	595	康美药业	-0.701	0.231	1.624
3	兖州煤碳	5.924	1.513	-0.044	596	潜江制药	-0.706	-0.430	2.085
4	马钢股份	5.175	-1.251	-2.804	597	浏阳花炮	-0.709	0.146	0.655
5	宁沪高速	5.341	0.835	-2.220	598	浪潮软件	-0.713	1.625	-1.313
6	广州控股	4.101	2.596	0.640	599	兆维科技	-0.728	2.511	-1.366
7	青岛海尔	4.022	0.954	3.160	600	PT农商社	-0.751	0.516	0.510
8	四川长虹	3.996	-2.027	1.907	601	三佳模具	-0.776	0.527	0.385
9	仪征化工	3.873	-0.964	-1.598	602	雄震集团	-0.817	1.175	-1.407
10	上海汽车	3.834	1.293	-0.666	603	中软股份	-1.023	2.715	-1.685
⋮	⋮	⋮	⋮	⋮	604	天地科技	-1.023	2.355	-0.946

表5 按盈利因子得分 $\hat{f}_2^*$的排序

序号	股票名称	$\hat{f}_1^*$	$\hat{f}_2^*$	$\hat{f}_3^*$	序号	股票名称	$\hat{f}_1^*$	$\hat{f}_2^*$	$\hat{f}_3^*$
1	中软股份	-1.023	2.715	-1.685	⋮	⋮	⋮	⋮	⋮
2	广州控股	4.101	2.596	0.640	595	东方电机	-0.246	-3.212	-0.385
3	广汇股份	0.517	2.534	-1.608	596	ST嘉陵	-0.144	-3.570	-0.284
4	兆维科技	-0.728	2.511	-1.366	597	ST海药	-0.089	-3.709	0.225
5	长江通讯	-0.657	2.369	1.899	598	鼎天科技	0.034	-4.230	-0.209
6	天地科技	-1.023	2.355	-0.946	599	大元股份	0.111	-4.559	0.284
7	申能股份	3.248	2.158	-0.498	600	新城B股	-0.080	-4.687	-0.086
8	上港集箱	2.992	2.112	1.624	601	银鸽投资	-0.063	-4.869	-0.086
9	中远航运	-0.588	1.957	-1.449	602	济南百货	0.083	-4.968	0.012
10	创业环保	0.797	1.755	-2.099	603	ST东锅	0.263	-5.979	0.272
⋮	⋮	⋮	⋮	⋮	604	国嘉实业	0.491	-5.730	1.055

表6 按每股价值因子得分 $\hat{f}_3^*$ 的排序

序号	股票名称	$\hat{f}_1^*$	$\hat{f}_2^*$	$\hat{f}_3^*$	序号	股票名称	$\hat{f}_1^*$	$\hat{f}_2^*$	$\hat{f}_3^*$
1	贵州茅台	0.877	1.366	5.750	⋮	⋮	⋮	⋮	⋮
2	用友软件	-0.581	-0.061	5.165	595	PT宝信	-0.571	1.145	-1.760
3	亿阳信通	-0.523	0.124	4.059	596	东方航空	5.446	-2.089	-1.861
4	华泰股份	-0.224	0.061	3.420	597	ST成量	-0.525	0.042	-1.873
5	太太药业	0.047	0.747	3.234	598	ST自仪	-0.185	-0.012	-1.905
6	赣粤高速	0.206	0.100	3.178	599	创业环保	0.797	1.755	-2.099
7	青岛海尔	4.022	0.954	3.160	600	上海石化	5.580	-2.704	-2.168
8	美克股份	-0.699	0.088	2.752	601	山东基建	2.275	0.797	-2.180
9	宇通客车	-0.264	0.604	2.619	602	ST中纺机	-0.390	0.278	-2.182
10	东方通讯	2.401	-0.750	2.593	603	宁沪高速	5.341	0.835	-2.220
⋮	⋮	⋮	⋮	⋮	604	马钢股份	5.175	-1.251	-2.804