Dimension Reduction

数 据 降 维

主成分分析与因子分析

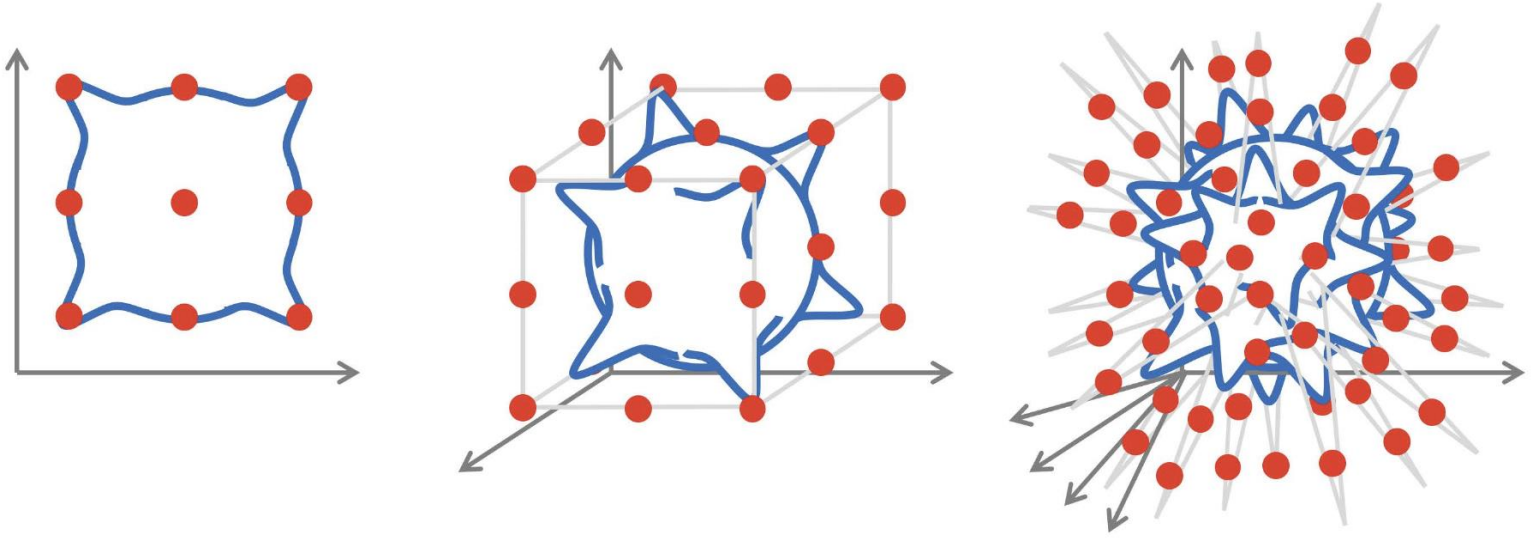
城市分析方法系列课程

苏州大学 王灿

大纲

- 数据降维概况
- 主成分分析
- 因子分析

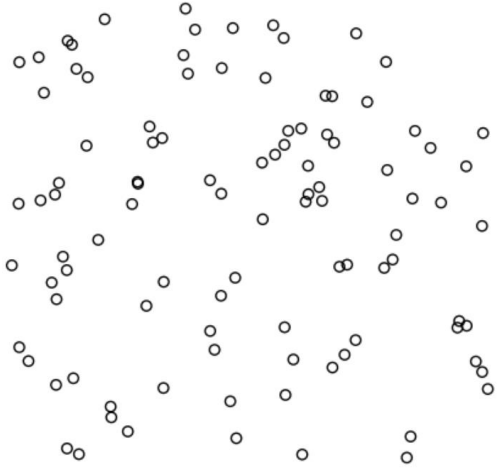
维度灾难（curse of dimensionality）



- 随着数据集的维度（即变量的数量）增加，分析难度也随之增加的现象。
- 3水平完全实验设计：2变量→9个，4变量→81个
- 多重共线性 ： 回归中的多个解释变量高度相关时，结果将存在严重问题。

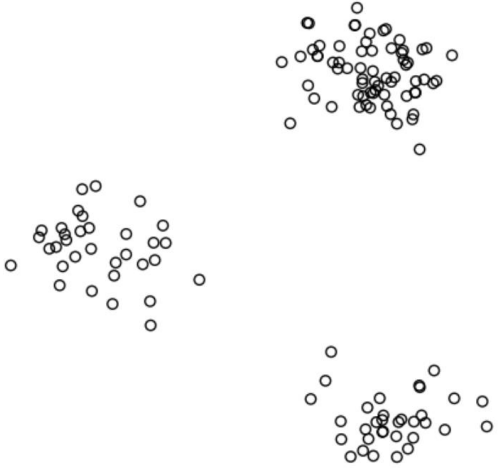
数据降维概况

真的需要多个维度吗?

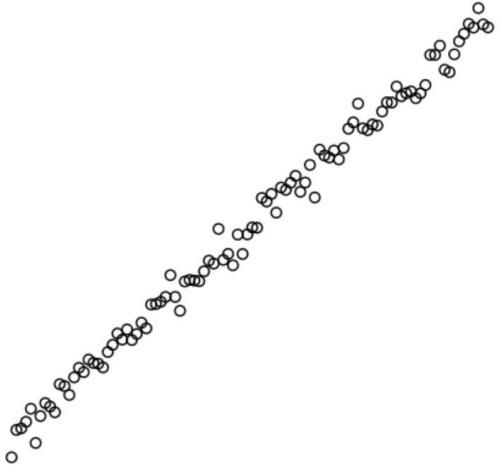


无规律，
需要两个维度

能把2D的面变成1D的线吗?



有类别特征，
需要两个维度



有相关特征，
可能只需要一个维度

数据降维概况

目的： 变量过多、信息重叠 ⟶ 变量缩减、相互独立

↓

- 计算量+、复杂性+
- 难以抓住研究对象的关键特征
- 变量高度相关可能带来统计污点

数据降维（dimension reduction）：对原始数据进行压缩，以相对**较少的新变量**保留原始变量的**大部分信息**，并且使新变量之间尽可能相互**独立**，减少信息重叠。

数据降维的方法

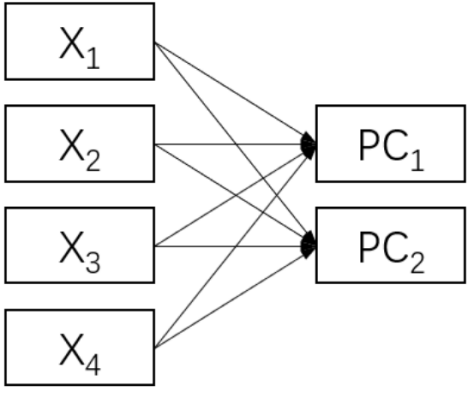
- 派代表
- 求平均

- **主成分分析**（principal component analysis, PCA）
- **因子分析**（factor analysis, FA）

基本原理

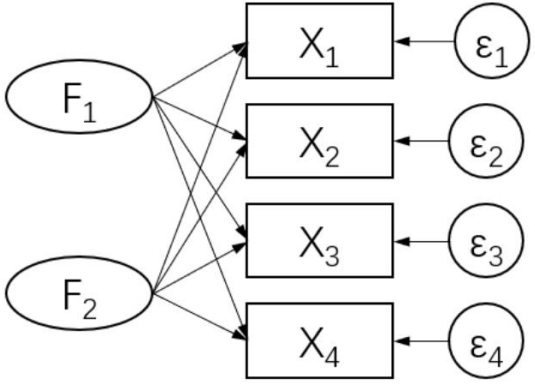
4 个原始变量（$X1$~$X4$）

2 个主成分（$PC1$, $PC2$）/ 2 个因子（$F1$, $F2$）



主成分分析：将原始变量转换为新变量——主成分。

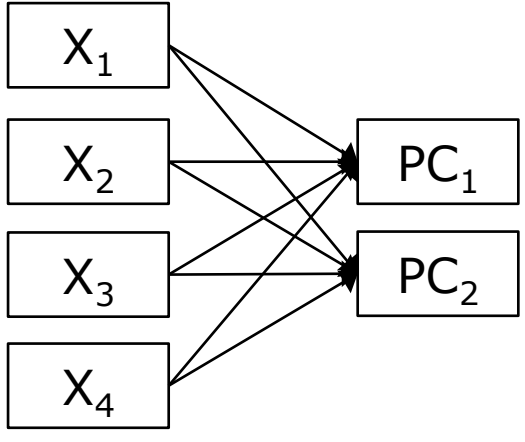
因子分析：认为原始变量是由背后的因子决定的，观察到的变量只不过是因子的一种“外显表现”，从而用因子解释原始变量，$\varepsilon 1$~$\varepsilon 4$代表了未能解释的部分。



转换矩阵（载荷）

- 对于有 n 个原始变量、k 个主成分/因子的问题，需要获得一个 n 行 k 列的转换矩阵，反映了原始变量与主成分/因子的关系，称为载荷（loading）。

		主成分/因子			
		PC_1/F_1	PC_2/F_2	…	PC_k/F_k
原始变量	X_1	b_{11}	b_{12}	…	b_{1k}
	X_2	b_{21}	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}

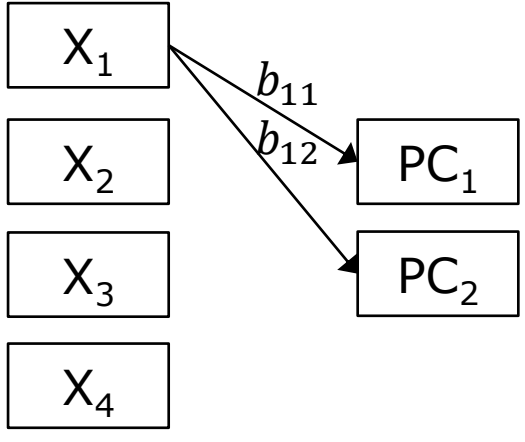


- $k<n$，从而实现了数据降维。

转换矩阵（载荷）

- 对于有 n 个原始变量、k 个主成分/因子的问题，需要获得一个 n 行 k 列的转换矩阵，反映了原始变量与主成分/因子的关系，称为载荷（loading）。

		主成分/因子			
		PC_1/F_1	PC_2/F_2	…	PC_k/F_k
原始变量	X_1	b_{11}	b_{12}	…	b_{1k}
	X_2	b_{21}	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}

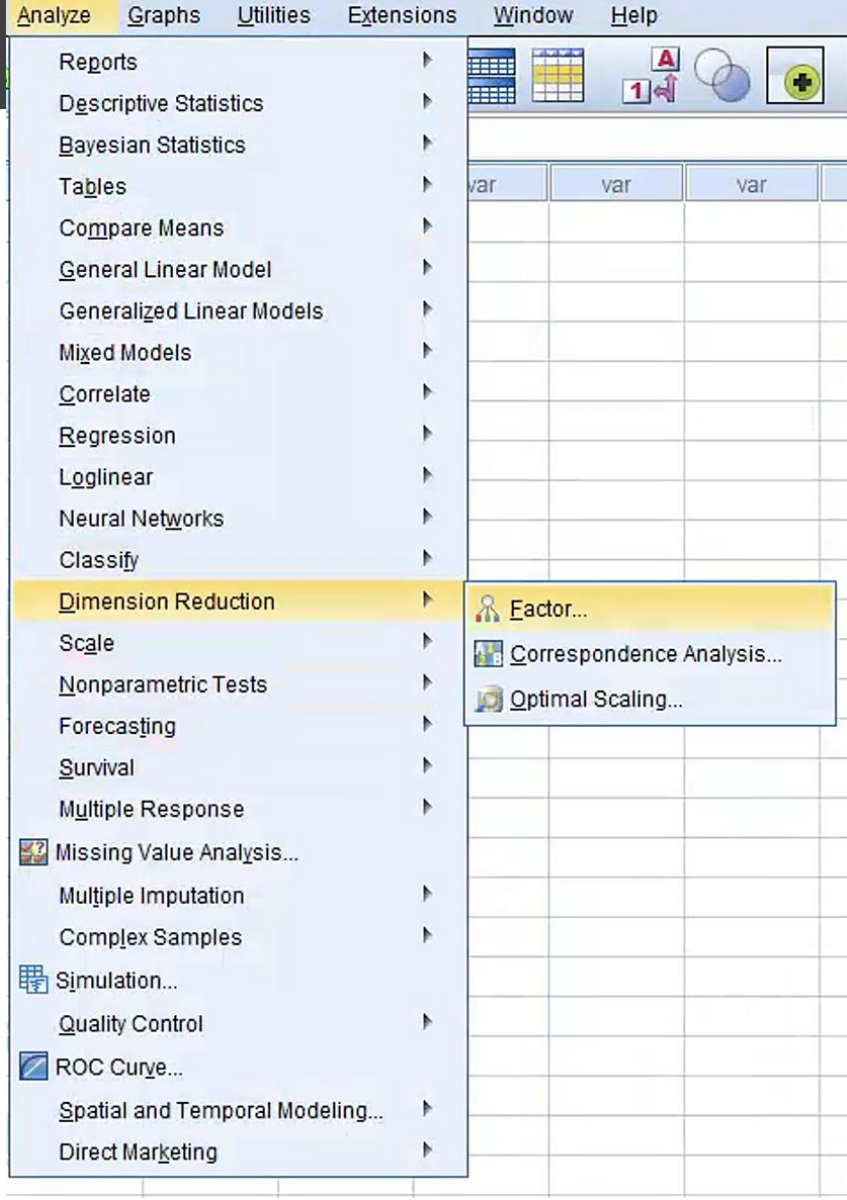


- $k<n$，从而实现了数据降维。

数据降维概况

软件

- 绝大多数的统计软件都可以运行主成分分析与因子分析，如 SPSS、Stata、R、 Python、Matlab等。
- SPSS中虽然没有专门的主成分分析功能，但可以借助因子分析实现。

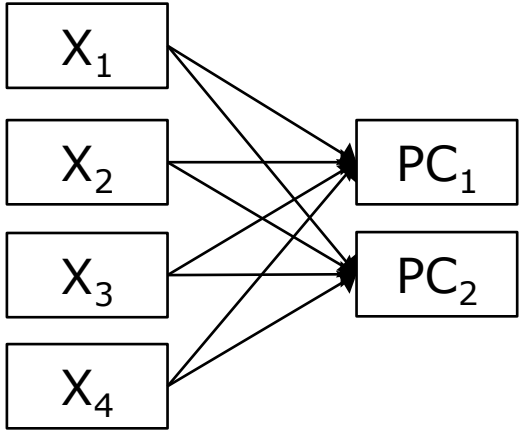


大纲

- 数据降维概况
- **主成分分析**
- 因子分析

原始变量与主成分的关系

		主成分			
		PC_1	PC_2	…	PC_k
原始变量	X_1	b_{11}	b_{12}	…	b_{1k}
	X_2	b_{21}	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}



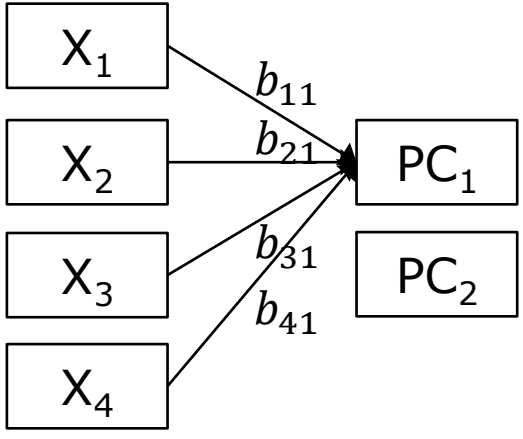
- k 个主成分（$PC_1 \sim PC_k$）与 n 个原始变量的关系为：

$$\begin{cases} PC_1 = b_{11}X_1 + b_{21}X_2 + \cdots + b_{n1}X_n \\ PC_2 = b_{12}X_1 + b_{22}X_2 + \cdots + b_{n2}X_n \\ \cdots\cdots \\ PC_k = b_{1k}X_1 + b_{2k}X_2 + \cdots + b_{nk}X_n \end{cases}$$

$X_1 \sim X_n$是标准化(减去均值，除以标准差)后的原始变量，各自的均值为0，标准差为1。

原始变量与主成分的关系

		主成分			
		PC_1	PC_2	…	PC_k
原始变量	X_1	b_{11}	b_{12}	…	b_{1k}
	X_2	b_{21}	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}



- k 个主成分（$PC_1 \sim PC_k$）与 n 个原始变量的关系为：

$$\begin{cases} PC_1 = b_{11}X_1 + b_{21}X_2 + \cdots + b_{n1}X_n \\ PC_2 = b_{12}X_1 + b_{22}X_2 + \cdots + b_{n2}X_n \\ \cdots\cdots \\ PC_k = b_{1k}X_1 + b_{2k}X_2 + \cdots + b_{nk}X_n \end{cases}$$

$X_1 \sim X_n$是标准化(减去均值，除以标准差)后的原始变量，各自的均值为0，标准差为1。

主成分分析

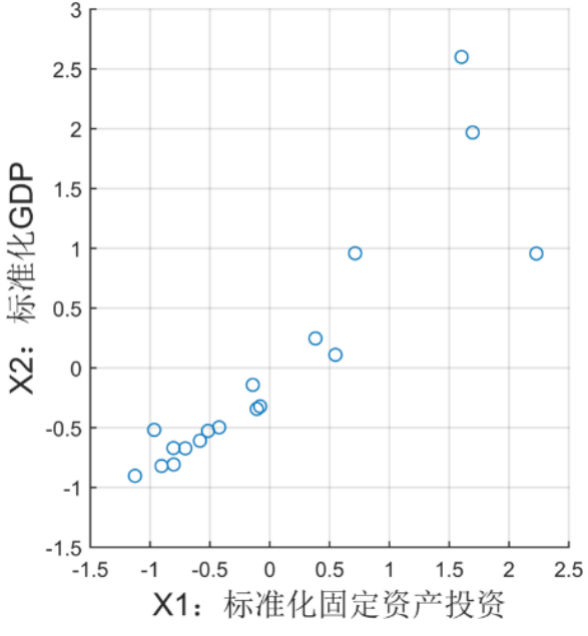
- 主成分分析实质上是对标准化原始变量的**加权求和**，将其变换为新变量。
- 为什么主成分分析中的权重比较科学呢？

- 从几何意义上，主成分分析是将原始变量定义的**坐标轴**进行**旋转**，得到更有利于表现数据特征的新坐标轴。
- 直观示例：**固定资产投资**与 **GDP** 均可作为表征经济发展水平的自变量，但二者之间存在很强的相关性，信息高度重叠，不宜同时纳入回归模型。如通过主成分分析，以一个主成分代替这两个变量。

主成分分析原理

城市	固定资产投资(亿元)	GDP(亿元)	标准化固定资产投资 (X_1)	标准化 GDP (X_2)
北京	5493.5	14113.6	1.695	1.969
上海	5317.7	17166.0	1.603	2.601
天津	6511.4	9224.5	2.228	0.957
济南	1987.4	3910.5	-0.141	-0.143
南京	3306.1	5130.7	0.549	0.109
泰安	1270.5	2051.7	-0.517	-0.528
苏州	3617.8	9228.9	0.713	0.958
……	……	……	……	……
平均值	2257.3	4602.7	0	0
标准差	1909.5	4831.0	1	1

减去均值、除以标准差



原坐标系

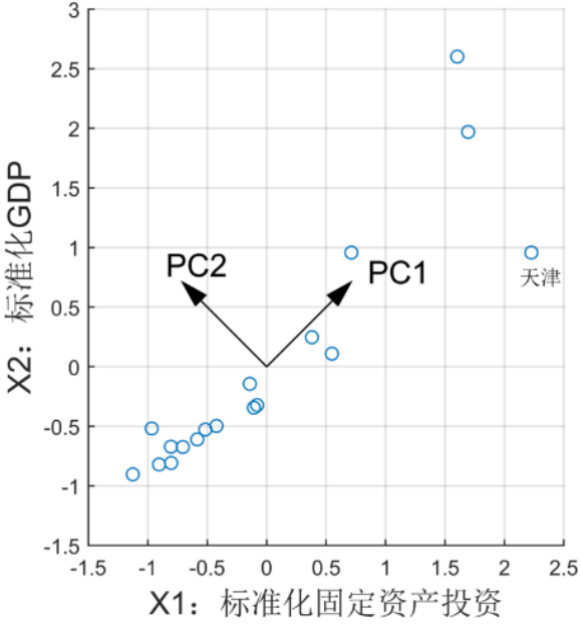
主成分分析原理：相关系数矩阵

	X_1	X_2
X_1	1	**0.901**
X_2	**0.901**	1

X_1与X_2的相关系数高达0.901，这种高度相关的数据有利于提取主成分。

主成分分析原理：特征分解

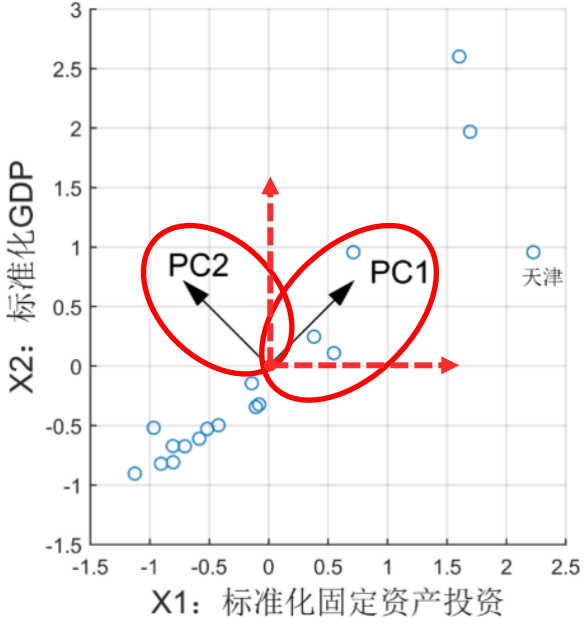
- 对相关系数矩阵进行**特征分解**：
  - 对于 n 个变量的相关系数矩阵，特征分解将得到 n 个特征值，每个特征值分别对应一个特征向量。每组“特征向量+特征值”即对应一个主成分。
  - **特征向量** (eigenvector)——**最重要的变化方向**。
  - **特征值** (eigenvalue)——**这个方向有多重要**。
  - 特征向量之间相互垂直，反映**新坐标轴的方向**。
  - 特征值是数据点在新坐标轴上投影的方差，反映了**该方向上信息量的大小**。



原坐标系

主成分分析原理：特征分解

- 第一特征值 = 1.901，对应特征向量 (0.707, 0.707)，指向数据点变异程度最大、信息量最多的方向，即第一主成分（PC_1）的方向。
- 第二特征值 = 0.099，对应特征向量 (-0.707, 0.707)，与PC_1垂直，是第二主成分（PC_2）的方向。
- $PC_1 - PC_2$ 构成了新的坐标系，由原始的 $X_1 - X_2$ 坐标系旋转一定角度得到。



原坐标系

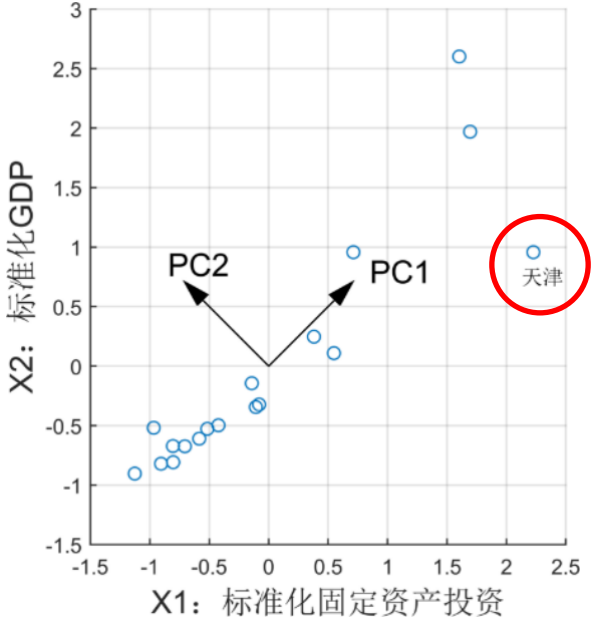
主成分分析原理：载荷矩阵

- 将两个特征向量组合，即得到**主成分荷载矩阵。**

	PC_1	PC_2
X_1	0.707	-0.707
X_1	0.707	0.707

- 主成分计算公式：$\begin{cases} PC_1 = 0.707X_1 + 0.707X_2 \\ PC_2 = -0.707X_1 + 0.707X_2 \end{cases}$

- 以天津为例：X_1=2.228，X_2=0.957，
  代入上式，得 PC_1=2.252，PC_2=-0.899



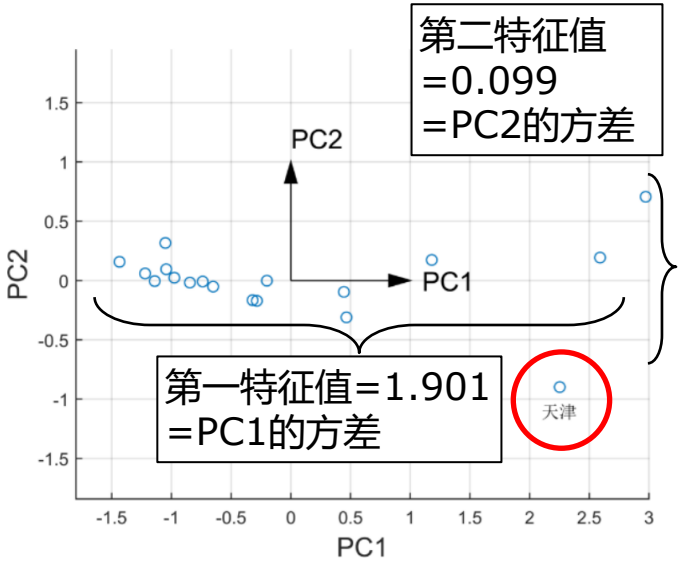
原坐标系

主成分分析原理

- **主成分的几何意义**：数据点在“PC1—PC2”这个旋转坐标系上的投影。

  以天津为例：PC_1=2.252，PC_2=-0.899

- 还记得**特征值的意义**吗：某一维度主成分的方差，反映了该维度信息量大小。



旋转坐标系：主成分散点图

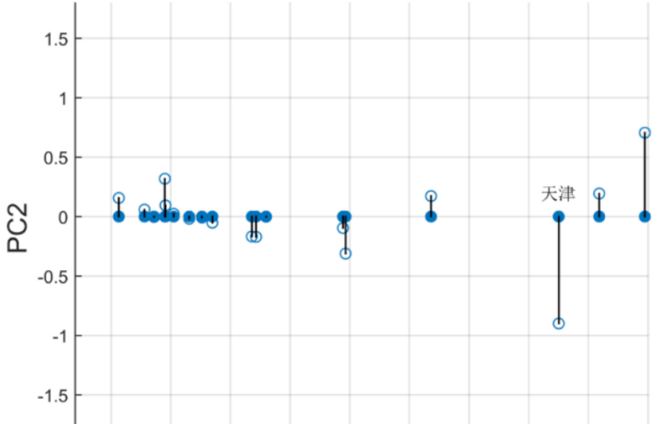
主成分分析原理

- **主成分的几何意义**：数据点在“PC1—PC2”这个旋转坐标系上的投影。

  以天津为例：PC_1=2.252，PC_2=-0.899

- 还记得**特征值的意义**吗：某一维度主成分的方差，反映了该维度信息量大小。

- **如何实现降维**（2维→1维）：把数据点投影到信息量最大的PC_1方向上，忽略PC_2方向上的变异，以PC_1（实心点）取代X_1与X_2（空心点），也只是损失很少的信息。



旋转坐标系：只保留第一主成分

主成分分析原理：特征值的意义

- 两个**特征值的和**为1.901+0.099=2。
  - 方差即是信息量：n 个变量的总信息量为 n（每个变量的方差为1）。
  - 分析得到的 n 个特征值之和也为必定为 n（总信息量恒定）。
  - 原先平均分布的信息量在新的坐标轴上重新分布，有大有小。

- 因此，可以用特征值与 n 的比值，计算每个主成分的方差贡献率，即所解释的信息量的比重。
  - PC_1的方差贡献率为$1.901/2 = 95.07\%$
  - PC_2的方差贡献率为$0.099/2 = 4.93\%$

主成分分析原理：特征值的意义

- 在确定主成分数量时，常常以“**特征值大于1**”作为标准。

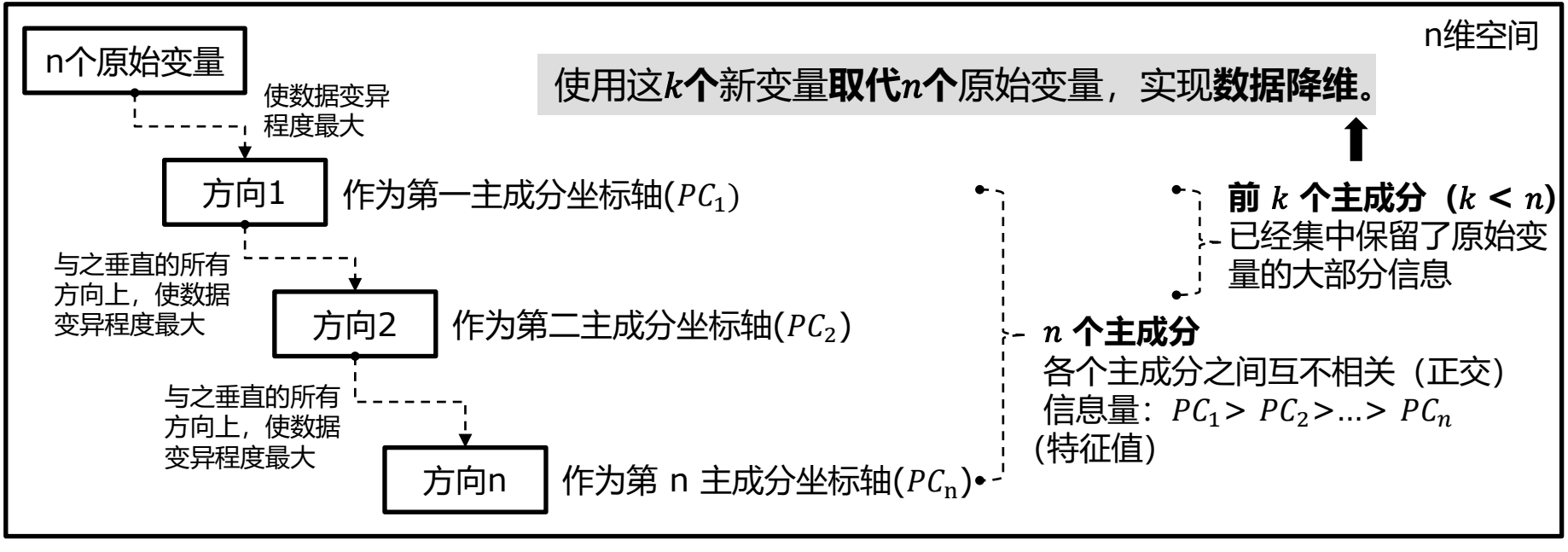
  这是由于特征值代表了主成分的方差，而1是每个原始变量标准化后的方差，当特征值小于1时，主成分所包含的信息量甚至还不如原始变量。

- 另外，也可以以“前k个主成分的**累积方差贡献率大于一定的阈值**（如85%）”为标准。

- 本例中，显然应当保留特征值为1.901、方差贡献率高达95.07%的 PC_1 ，舍弃PC_2。

主成分分析原理：一般化

- 主成分分析是对原始变量的加权求和，在几何意义上，是将原始变量定义的**坐标轴进行旋转**，得到更有利于表现数据特征的新坐标轴。



案例讲解

工业用地更新属性

主成分分析案例

数据：

上海市工业用地更新的11项属性数据。各属性之间有一定的信息重叠。

问题：

通过主成分分析实现数据降维，提取相互独立的主成分变量，用于进一步的回归分析中。

主成分分析案例

No	用地编号
LA	用地面积(ha)
FA	建筑面积(万㎡)
Dist	到市中心的直线距离(km)
Metro	到最近地铁站的距离(m)
BE	周边建成环境 BE=1差 BE=2中 BE=3好
BR	周边的写字楼租金(元/天*m²)
SR	周边的商铺租金(元/天*m²)
Loc	区位 Loc=1内环内 Loc=2内中环间 Loc=3中外环间
AQ	建筑质量 AQ=1差 AQ=2中 AQ=3好
AH	建筑高度(m²)
IC	用地不规则系数，=用地周长/用地面积的平方根值越低，用地越接近正圆形，则认为形状越规整

No (1)	LA (2)	FA (3)	Dist (4)	Metro (5)	BE (6)	BR (7)	SR (8)	Loc (9)	AQ (10)	AH (11)	IC (12)
1	0.35	0.29	0.82	556.4	1	3.20	4.95	1	1	12	5.67
2	0.49	0.46	1.25	473.5	1	3.20	5.68	1	1	16	4.27
3	23.90	11.72	2.72	1290.3	1	2.20	2.73	3	2	15	4.22
4	0.23	0.14	2.05	1174.6	1	1.80	3.37	3	1	12	4.11
5	0.85	1.02	1.17	543.0	2	3.70	5.49	1	3	22	5.97
6	0.08	0.12	1.10	485.8	1	3.70	5.58	1	1	12	4.41
7	0.34	0.51	0.20	243.8	2	3.50	9.07	1	2	21	4.05
8	1.77	3.34	0.46	566.3	2	3.20	6.11	1	3	36	5.02
9	1.82	2.24	1.10	704.9	1	2.70	4.49	2	1	12	4.68
10	1.22	1.83	.64	540.9	2	2.10	5.29	3	3	32	4.09
……	……	……	……	……	……	……	……	……	……	……	……

主成分分析案例

① 变量标准化

② 计算相关系数矩阵

	LA	FA	Dist	Metro	BE	BR	SR	Loc	AQ	AH	IC
LA	1										
FA	0.917	1									
Dist	0.262	0.239	1								
Metro	0.227	0.236	0.753	1							
BE	-0.105	-0.044	-0.171	-0.102	1						
BR	-0.166	-0.189	-0.546	-0.419	0.065	1					
SR	-0.218	-0.217	-0.715	-0.681	0.070	0.361	1				
Loc	0.252	0.293	0.583	0.559	-0.021	-0.732	-0.608	1			
AQ	0.037	0.116	-0.193	-0.111	0.725	0.099	0.115	0.011	1		
AH	-0.040	0.109	-0.192	-0.127	0.760	-0.012	0.076	0.032	0.806	1	
IC	0.007	0.041	0.066	0.015	-0.132	0.047	0.003	-0.031	-0.027	-0.090	1

主成分分析案例

③ 提取主成分：对上述矩阵进行特征分解，根据特征值和方差贡献率提取主成分

主成分 i	特征值 λ_i	方差贡献率（%） λ_i/n	累积方差贡献率（%） $\sum_i^k \lambda_i/n$	抽取 $\lambda_i > 1$
1	3.771	34.283	34.283	√
2	2.529	22.992	57.274	√
3	1.613	14.666	71.940	√
4	1.004	9.127	81.067	√
5	0.793	7.205	88.272	
6	0.364s	3.305	91.577	
7	0.285	2.587	94.164	
8	0.243	2.213	96.377	
9	0.208	1.890	98.267	
10	0.132	1.198	99.465	
11	0.059	0.535	100	

特征值：是主成分变量的方差，反映了该主成分的信息量大小，在表中按从大到小的顺序排列。所有特征值之和等于变量总数 ($n=11$)。

主成分分析案例

③ 提取主成分：对上述矩阵进行特征分解，根据特征值和方差贡献率提取主成分

主成分 i	特征值 λ_i	方差贡献率（%） λ_i/n	累积方差贡献率（%） $\sum_i^k \lambda_i/n$	抽取 $\lambda_i > 1$
1	3.771	34.283	34.283	√
2	2.529	22.992	57.274	√
3	1.613	14.666	71.940	√
4	1.004	9.127	81.067	√
5	0.793	7.205	88.272	
6	0.364s	3.305	91.577	
7	0.285	2.587	94.164	
8	0.243	2.213	96.377	
9	0.208	1.890	98.267	
10	0.132	1.198	99.465	
11	0.059	0.535	100	

方差贡献率：即特征值与变量总数n的比值，反映了主成分的方差在全部方差中的比重。其值越高，说明主成分综合原始变量信息的能力越强。

主成分分析案例

③ 提取主成分：对上述矩阵进行特征分解，根据特征值和方差贡献率提取主成分

主成分 i	特征值 λ_i	方差贡献率（%） λ_i/n	累积方差贡献率（%） $\sum_i^k \lambda_i/n$	抽取 $\lambda_i > 1$
1	3.771	34.283	34.283	√
2	2.529	22.992	57.274	√
3	1.613	14.666	71.940	√
4	1.004	9.127	81.067	√
5	0.793	7.205	88.272	
6	0.364s	3.305	91.577	
7	0.285	2.587	94.164	
8	0.243	2.213	96.377	
9	0.208	1.890	98.267	
10	0.132	1.198	99.465	
11	0.059	0.535	100	

累积方差贡献率：即前k个主成分的方差贡献率之和，可以作为确定主成分数量的依据。

主成分分析案例

③ 提取主成分：对上述矩阵进行特征分解，根据特征值和方差贡献率提取主成分

主成分 i	特征值 λ_i	方差贡献率（%） λ_i/n	累积方差贡献率（%） $\sum_i^k \lambda_i/n$	抽取 $\lambda_i > 1$
1	3.771	34.283	34.283	√
2	2.529	22.992	57.274	√
3	1.613	14.666	71.940	√
4	1.004	9.127	81.067	√
5	0.793	7.205	88.272	
6	0.364s	3.305	91.577	
7	0.285	2.587	94.164	
8	0.243	2.213	96.377	
9	0.208	1.890	98.267	
10	0.132	1.198	99.465	
11	0.059	0.535	100	

依据“**特征值大于 1**”的标准，抽取前4个主成分。

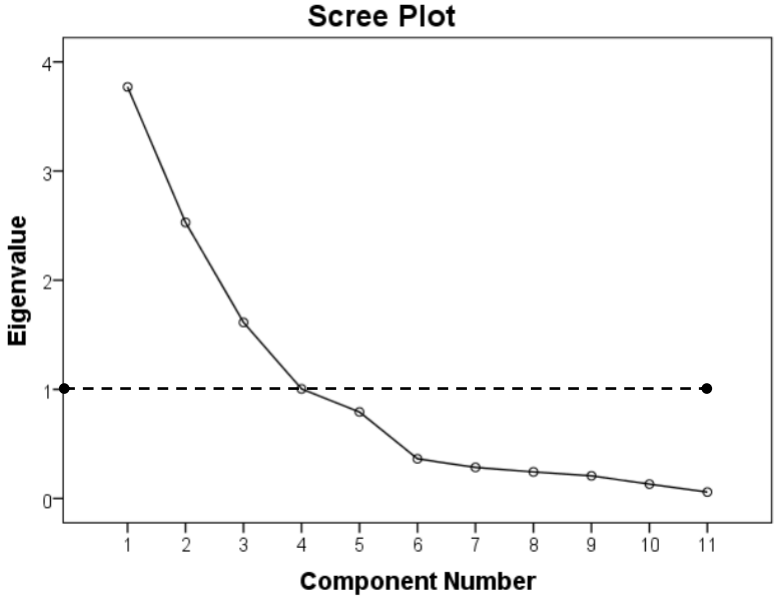
前4个主成分的累积方差贡献率超过 80%，已包含了原始变量大部分的信息。

主成分分析案例

碎石图

- 碎石图（scree plot）：又称陡坡图，反映了特征值由大到小的变化趋势。
- 左侧陡峭的部分对应特征值大、信息量高的主成分，右侧平缓的部分对应特征值小、信息量小的主成分。
- 辅助确定主成分数量：明显下跌的位置。

本例中，主成分数量在5、6之间明显下跌，可考虑保留5个主成分。但是由于第5个主成分的特征值小于1，最终还是只保留前4个主成分。



主成分分析案例

④ 得到主成分荷载矩阵，生成主成分变量。

	主成分			
	1	2	3	4
LA：用地面积	0.491	0.223	0.813	-0.085
FA：建筑面积	0.480	0.330	0.788	-0.037
Dist：中心距离	0.862	0.003	-0.204	0.112
Metro：地铁站距离	0.802	0.071	-0.216	0.096
BE：周边建成环境	-0.293	0.827	-0.218	0.026
BR：写字楼租金	-0.678	-0.152	0.252	0.147
SR：商铺租金	-0.782	-0.101	0.235	-0.079
Loc：区位	0.787	0.256	-0.219	-0.067
AQ：建筑质量	-0.270	0.875	-0.018	0.137
AH：建筑高度	-0.254	0.894	-0.104	0.053
IC：不规则系数	0.037	-0.139	0.128	0.959

4个主成分变量，可检验其两两之间的相关系数为0。

$$PC_1 = 0.491LA + 0.480FA + 0.862Dist + 0.802Metro - 0.293BE - 0.678BR - 0.782SR + 0.787Loc - 0.270AQ - 0.254AH + 0.037IC$$

$$PC_2 = 0.223LA + 0.330FA + 0.003Dist + 0.071Metro + 0.827BE - 0.152BR - 0.101SR + 0.256Loc + 0.875AQ + 0.894AH - 0.139IC$$

$$PC_3 = 0.813LA + 0.788FA - 0.204Dist - 0.216Metro - 0.218BE + 0.252BR + 0.235SR - 0.219Loc - 0.018AQ - 0.104AH + 0.128IC$$

$$PC_4 = -0.085LA - 0.037FA + 0.112Dist + 0.096Metro + 0.026BE + 0.147BR - 0.079SR - 0.067Loc + 0.137AQ + 0.053AH + 0.959IC$$

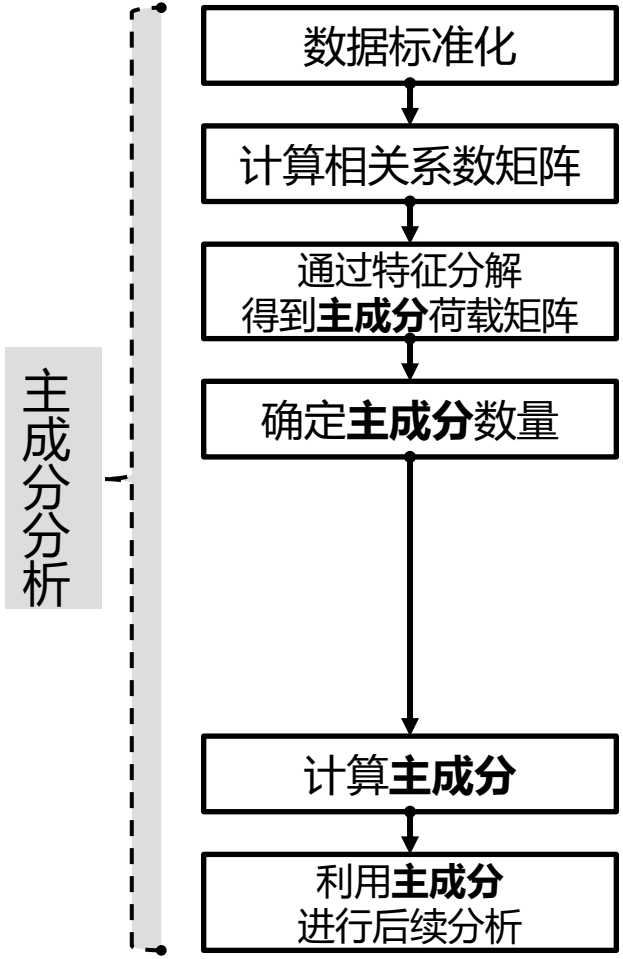
主成分分析案例

由此，将11个原始变量转换成为4个互不相关、无信息重叠的新变量，下一步可以利用它们进行回归、聚类等后续分析。

- 第一主成分（ PC_1 ）在 Dist、Metro、BR、SR、Loc 上的载荷较高，可以看作是反映中心距离、地铁站距离、写字楼租金、商铺租金、区位方面的综合指标。
- 第二主成分（ PC_2 ）在 BE、AQ、AH 上的荷载较高，可以看作是反映周边建成环境、建筑质量、建筑高度方面的综合指标。
- 第三主成分（ PC_3 ）在 LA、FA 上的荷载较高，可以看作是反映用地面积、建筑面积方面的综合指标。
- 第四主成分（ PC_4 ）在IC上的荷载较高，可以看作是反映用地不规则系数方面的综合指标。

主成分分析小结

分析步骤

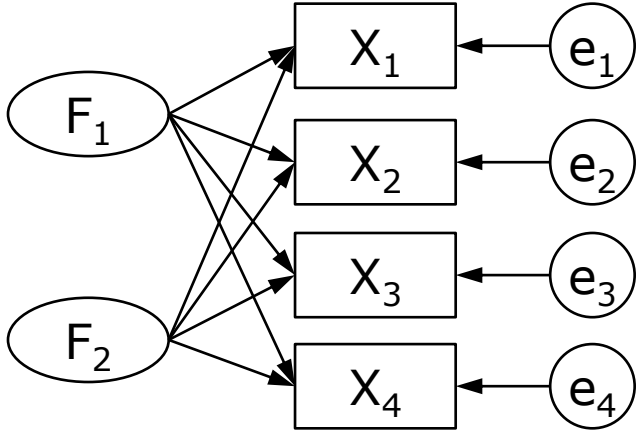


大纲

- 数据降维概况
- 主成分分析
- **因子分析**

原始变量与公因子的关系

		公因子			
		F_1	F_2	…	F_k
原始变量	X_1	b_{11}	b_{12}	…	b_{1k}
	X_2	b_{21}	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}



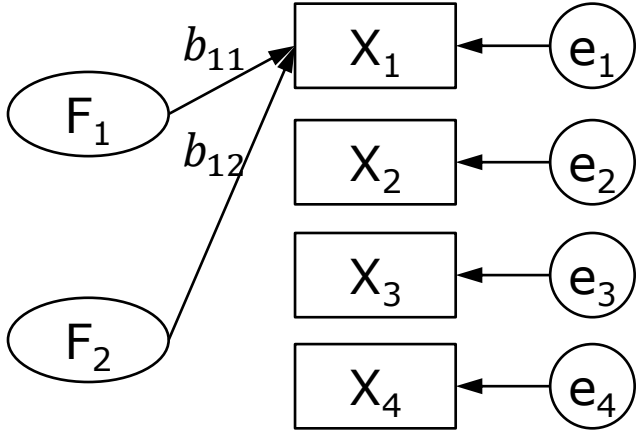
- k 个公因子（$F_1 \sim F_k$）与 n 个原始变量的关系为：

$$\begin{cases} X_1 = b_{11}F_1 + b_{12}F_2 + \cdots + b_{1k}F_k + e_1 \\ X_2 = b_{21}F_1 + b_{22}F_2 + \cdots + b_{2k}F_k + e_2 \\ \cdots\cdots \\ X_n = b_{n1}F_1 + b_{n2}F_2 + \cdots + b_{nk}F_k + e_n \end{cases}$$

$X_1 \sim X_n$是标准化(减去均值，除以标准差)后的原始变量，各自的均值为0，标准差为1。

原始变量与公因子的关系

		公因子			
		F_1	F_2	…	F_k
原始变量	X_1	b_{11}	b_{12}	…	b_{1k}
	X_2	b_{21}	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}



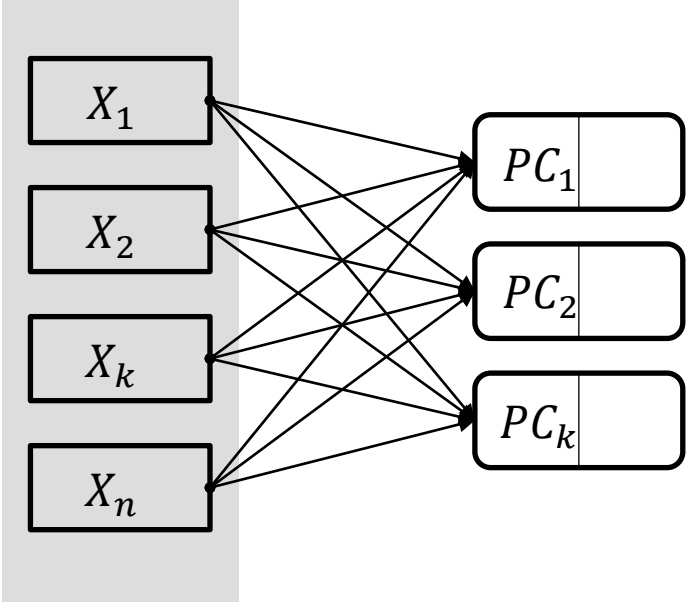
- k 个公因子（$F_1 \sim F_k$）与 n 个原始变量的关系为：

$$\begin{cases} X_1 = b_{11}F_1 + b_{12}F_2 + \cdots + b_{1k}F_k + e_1 \\ X_2 = b_{21}F_1 + b_{22}F_2 + \cdots + b_{2k}F_k + e_2 \\ \cdots\cdots \\ X_n = b_{n1}F_1 + b_{n2}F_2 + \cdots + b_{nk}F_k + e_n \end{cases}$$

$X_1 \sim X_n$是标准化(减去均值，除以标准差)后的原始变量，各自的均值为0，标准差为1。

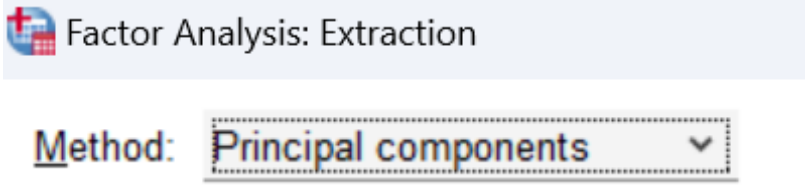
因子分析 vs. 主成分分析

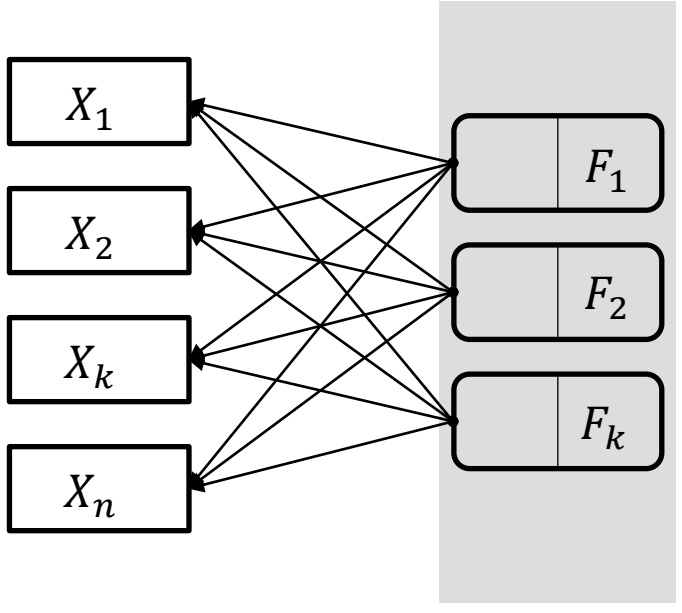
- 在一定条件下，主成分转换与因子转换是**互逆**的，二者可**共用**同一转换矩阵。



因子分析 vs. 主成分分析

- 在一定条件下，主成分转换与因子转换是**互逆**的，二者可**共用**同一转换矩阵。
- 主成分法是求解因子分析最常用的方法。

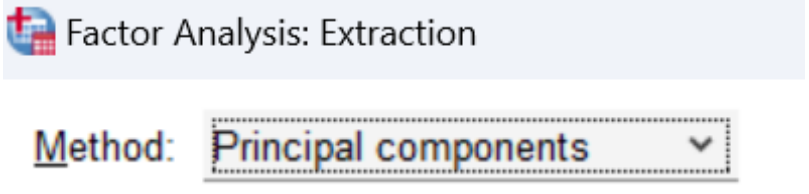




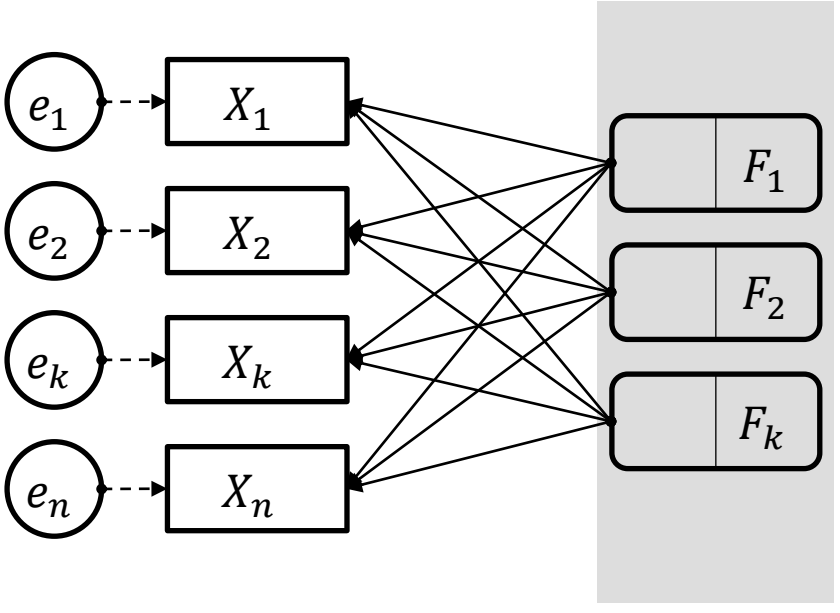
主成分分析 $(X \rightarrow PC)$ ⇌ 因子分析 $(F \rightarrow X)$

因子分析 vs. 主成分分析

- 在一定条件下，主成分转换与因子转换是**互逆**的，二者可**共用**同一转换矩阵。
- 主成分法是求解因子分析最常用的方法。



- 关键不同：因子分析具有明确的解释性。



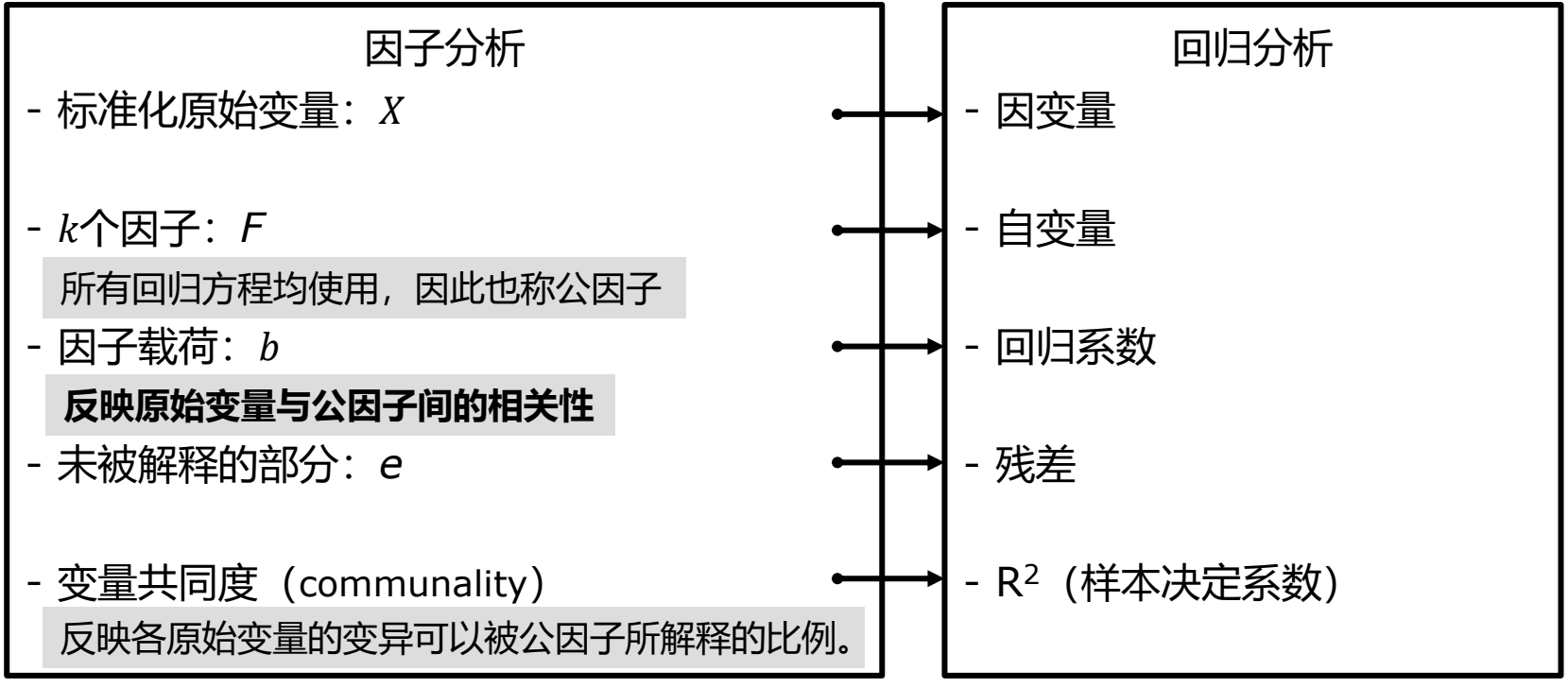
主成分分析 $(X \rightarrow PC)$ ⇌ 因子分析 $(F \rightarrow X)$

一种**压缩信息的数据变换**

一种**对关系结构的解释**

因子分析的深层意义

因子分析把原始变量X放到了”被解释”的位置，如同对各个X分别做线性回归。

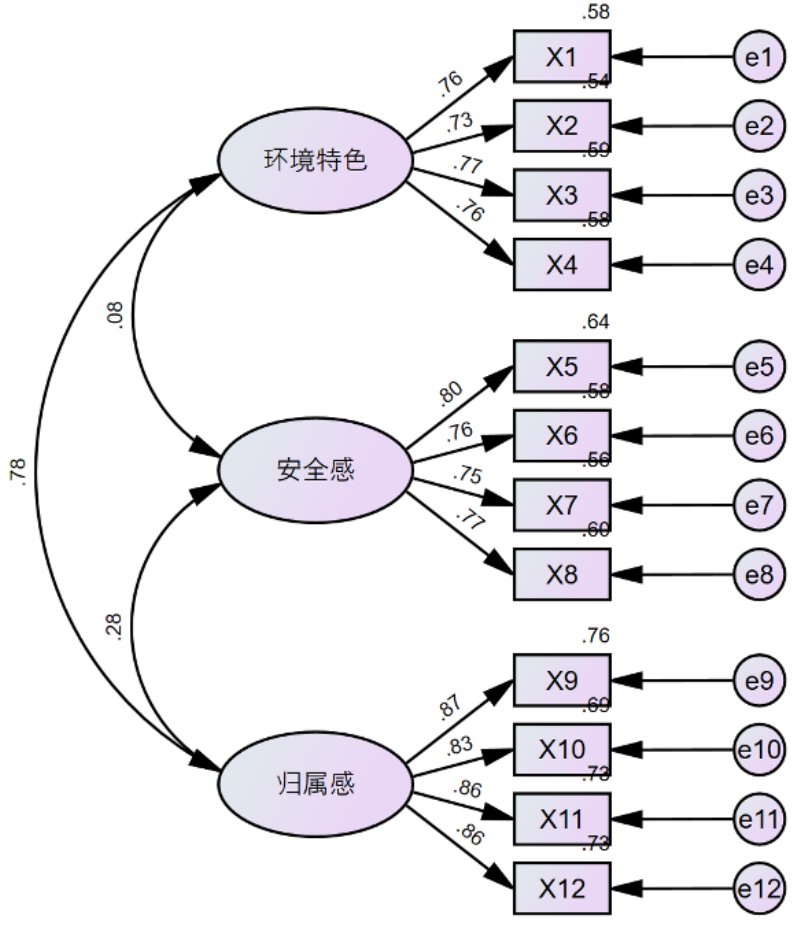


因子分析的深层意义

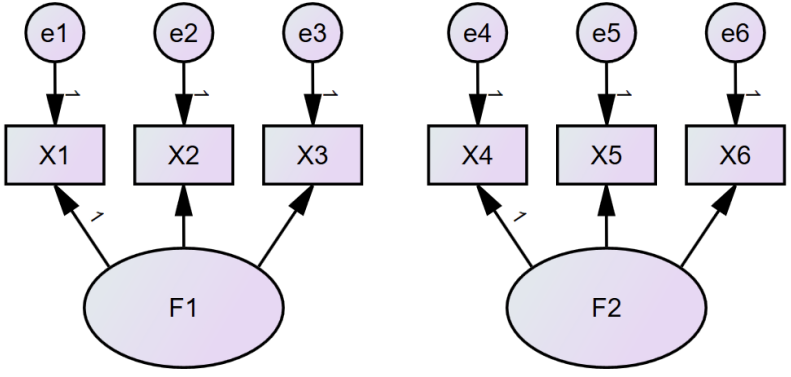
- 因子分析蕴含了研究者对数据结构的“因子化”假设。
  - **数据**：外显的可观察变量（observable variables）
  - **因子**：背后的潜变量（latent variables）/ 概念（construct）
- 因子分析比主成分分析更强调**可解释性**：每个因子作为一个概念，应当有明确的意义。

因子分析的深层意义

X1	我们小区是经过精心设计的。
X2	我们小区的住宅楼别具一格。
X3	我们小区的户外景化与众不同。
X4	大家从不会混淆我们小区和其他小区。
X5	我们小区很少发生犯罪事件。
X6	我不担心家中财物被盗。
X7	我们小区的安保服务令人放心。
X8	即使深夜回家，我也不觉得害怕。
X9	我对我们社区有很深的依恋。
X10	我很了解我们社区的情况。
X11	我希望付出时间，为社区做贡献。
X12	我和社区里的邻居相处得很愉快。

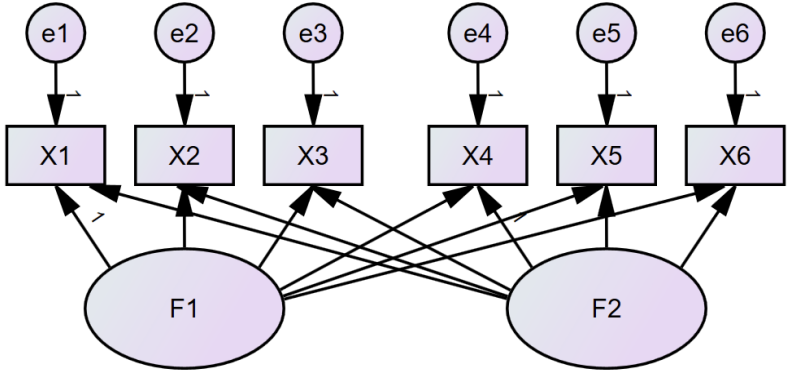


验证式 vs. 探索式



验证式因子分析
confirmatory factor analysis, **CFA**

- Restricted：每个变量只关联到一个因子
- Deductive：先理论，后数据

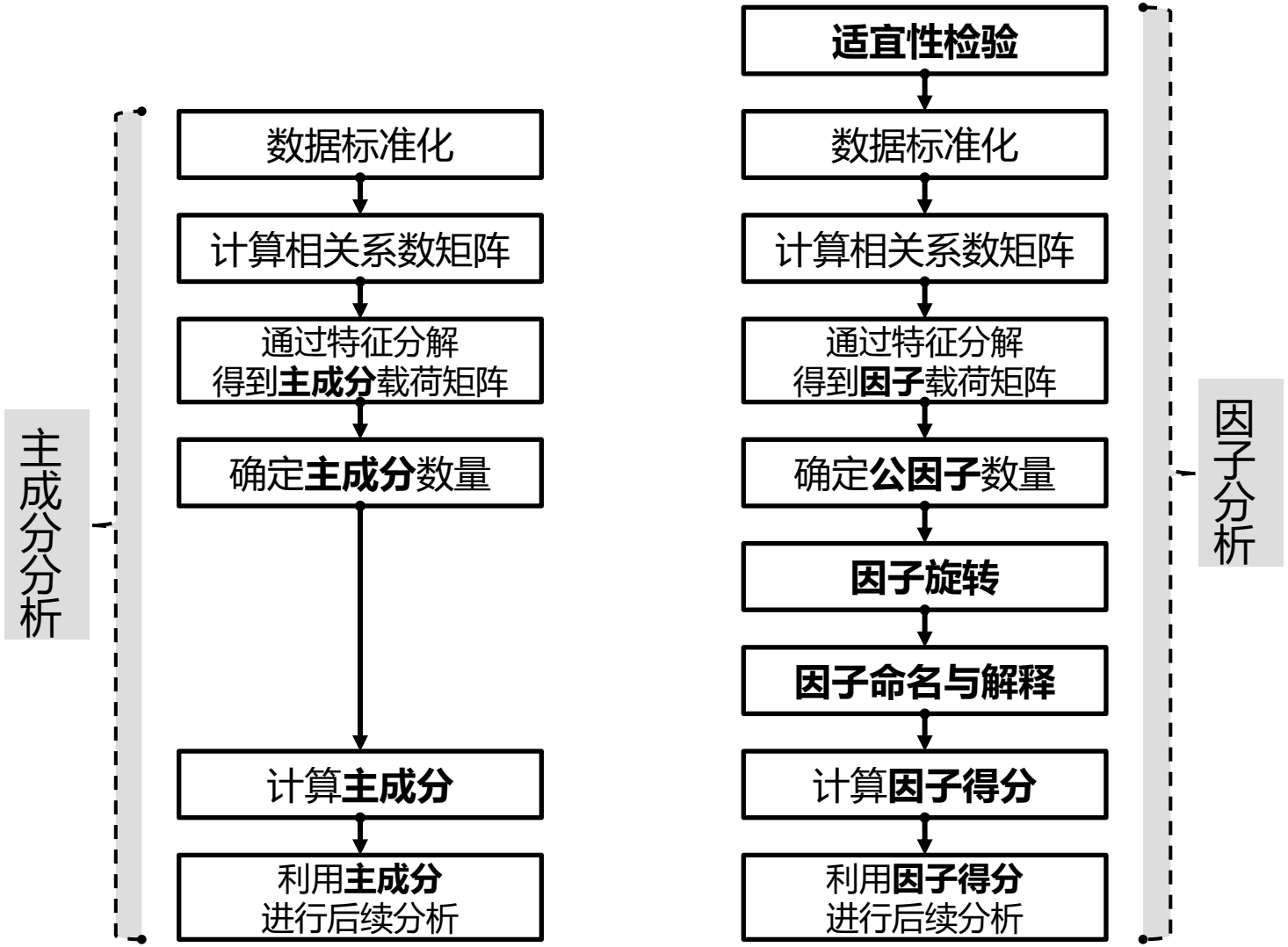


探索式因子分析
exploratory factor analysis, **EFA**

- Unrestricted：所有变量均关联到所有因子
- Inductive：先数据，后理论

因子分析

分析步骤


主成分分析
数据标准化
计算相关系数矩阵
通过特征分解
得到主成分载荷矩阵
确定主成分数量
计算主成分
利用主成分
进行后续分析
因子分析
适宜性检验
数据标准化
计算相关系数矩阵
通过特征分解
得到因子载荷矩阵
确定公因子数量
因子旋转
因子命名与解释
计算因子得分
利用因子得分
进行后续分析

适宜性检验

- **KMO 统计量**：指示了变量的方差能够被公因子所解释的比例。

KMO 统计量	结论
0.00～0.49	不可接受
0.50～0.59	勉强
0.60～0.69	中度适合
0.70～0.79	适合
0.80～0.89	很适合
0.90～1.00	非常适合

- **Barlett球度检验**：原假设是“变量之间的相关系数矩阵与单位矩阵没有显著差异”，即“变量之间互不相关”。

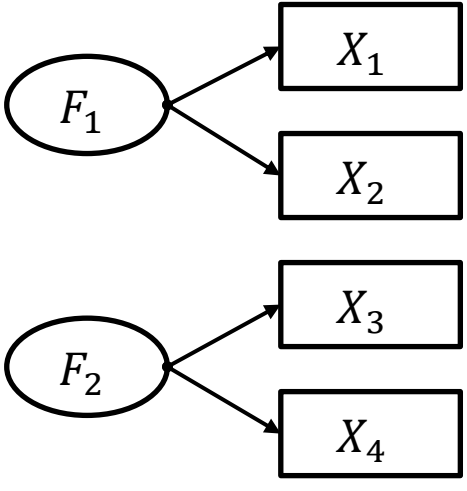
只有当原假设被拒绝（$p<0.05$），才适宜因子分析。

因子命名

- **我代表了谁？——“命名依据变量”**：以矩阵中荷载 b 的绝对值大小，判断特定的因子与哪些变量的相关性较高。

		公因子			
		F_1	F_2	…	F_k
原始变量	X_1	$\boldsymbol{b_{11}}$	b_{12}	…	b_{1k}
	X_2	$\boldsymbol{b_{21}}$	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}

- **理想的因子结构**：因子荷载矩阵的每一行只有一个非零值，即每个原始变量只与一个因子相关，这种结构最易于解释。

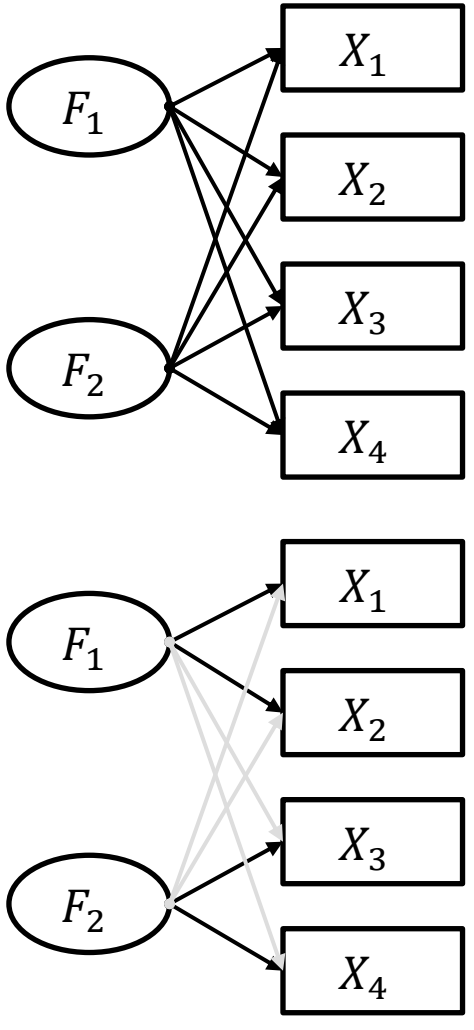


		公因子	
		F_1	F_2
原始变量	X_1	$\boldsymbol{b_{11}}$	0
	X_2	$\boldsymbol{b_{21}}$	0
	X_3	0	$\boldsymbol{b_{32}}$
	X_4	0	$\boldsymbol{b_{42}}$

因子命名

- 初始因子结构：一个原始变量在两个及以上的因子上均具有较高的荷载，难以判断该变量被哪个因子代表，不利于解释。

- **因子旋转**：为提高可解释性，可再次旋转坐标轴，尽可能接近简单结构。最常用的方法是最大方差法（varimax）：
  - 保持因子之间互不相关；
  - 使得**载荷两极分化**：要么接近于±1，要么接近于0。



因子得分

- 因子得分与主成分的作用相似。

载荷矩阵 $F \to X$		公因子			
		F_1	F_2	…	F_k
原始变量	X_1	b_{11}	b_{12}	…	b_{1k}
	X_2	b_{21}	b_{22}	…	b_{2k}
	…	…	…	…	…
	X_n	b_{n1}	b_{n2}	…	b_{nk}

得分矩阵 $X \to F$		公因子			
		F_1	F_2	…	F_k
原始变量	X_1	β_{11}	β_{12}	…	β_{1k}
	X_2	β_{21}	β_{22}	…	β_{2k}
	…	…	…	…	…
	X_n	β_{n1}	β_{n2}	…	β_{nk}

F 的值怎么求?

$$\begin{cases} X_1 = b_{11}F_1 + b_{12}F_2 + \cdots + b_{1k}F_k + e_1 \\ X_2 = b_{21}F_1 + b_{22}F_2 + \cdots + b_{2k}F_k + e_2 \\ \cdots\cdots \\ X_n = b_{n1}F_1 + b_{n2}F_2 + \cdots + b_{nk}F_k + e_n \end{cases}$$

$$\begin{cases} F_1 = \beta_{11}X_1 + \beta_{21}X_2 + \cdots + \beta_{n1}X_n \\ F_2 = \beta_{12}X_1 + \beta_{22}X_2 + \cdots + \beta_{n2}X_n \\ \cdots\cdots \\ F_k = \beta_{1k}X_1 + \beta_{2k}X_2 + \cdots + \beta_{nk}X_n \end{cases}$$

案例讲解

工业用地更新属性

因子分析案例

① 因子分析适宜性检验

KMO (Kaiser-Meyer-Olkin Measure of Sampling Adequacy)		0.661
Bartlett球度检验 (Bartlett's Test of Sphericity)	卡方值（近似）	848.081
	自由度	55
	p值	<0.001

- KMO统计量为0.661， 在0.6~0.69区间内，中度适合；
- Bartlett球度检验的p值小于 0.001，拒绝各变量独立的零假设。
- 因此，本数据适宜开展因子分析。

因子分析案例

② **提取公因子**：将上一节中的主成分荷载矩阵直接作为因子荷载矩阵。

	公因子			
	1	2	3	4
LA：用地面积	0.491	0.223	0.813	-0.085
FA：建筑面积	0.480	0.330	0.788	-0.037
Dist：中心距离	0.862	0.003	-0.204	0.112
Metro：地铁站距离	0.802	0.071	-0.216	0.096
BE：周边建成环境	-0.293	0.827	-0.218	0.026
BR：写字楼租金	-0.678	-0.152	0.252	0.147
SR：商铺租金	-0.782	-0.101	0.235	-0.079
Loc：区位	0.787	0.256	-0.219	-0.067
AQ：建筑质量	-0.270	0.875	-0.018	0.137
AH：建筑高度	-0.254	0.894	-0.104	0.053
IC：不规则系数	0.037	-0.139	0.128	0.959

$$LA = 0.491F_1 + 0.223F_2 + 0.813F_3 - 0.085F_4 + \varepsilon_{LA}$$

$$FA = 0.480F_1 + 0.330F_2 + 0.788F_3 - 0.037F_4 + \varepsilon_{FA}$$

$$Dist = 0.862F_1 + 0.003F_2 - 0.204F_3 + 0.112F_4 + \varepsilon_{Dist}$$

……

$$AH = -0.254F_1 + 0.894F_2 - 0.104F_3 + 0.053F_4 + \varepsilon_{AH}$$

$$IC = 0.037F_1 - 0.139F_2 + 0.128F_3 + 0.959F_4 + \varepsilon_{IC}$$

因子分析案例

② 提取公因子：变量共同度

各原始变量的信息能够被公因子解释的比例，相当于回归中的R^2，其值等于因子荷载矩阵中每一行的平方和。

本例中，大部分变量共同度在80%以上。但是，写字楼租金、商铺租金的变量共同度不足70%，表明公因子对这两个变量的解释能力相对较弱。

LA：用地面积	0.959
FA：建筑面积	0.962
Dist：中心距离	0.797
Metro：地铁站距离	0.704
BE：周边建成环境	0.818
BR：写字楼租金	0.567
SR：商铺租金	0.683
Loc：区位	0.737
AQ：建筑质量	0.858
AH：建筑高度	0.876
IC：不规则系数	0.956

因子分析案例

③ **因子旋转与命名**：通过最大方差法得到更加清晰的旋转因子荷载矩阵。

初始因子载荷矩阵

	公因子			
	1	2	3	4
LA：用地面积	0.491	0.223	0.813	-0.085
FA：建筑面积	0.480	0.330	0.788	-0.037
Dist：中心距离	0.862	0.003	-0.204	0.112
Metro：地铁站距离	0.802	0.071	-0.216	0.096
BE：周边建成环境	-0.293	0.827	-0.218	0.026
BR：写字楼租金	-0.678	-0.152	0.252	0.147
SR：商铺租金	-0.782	-0.101	0.235	-0.079
Loc：区位	0.787	0.256	-0.219	-0.067
AQ：建筑质量	-0.270	0.875	-0.018	0.137
AH：建筑高度	-0.254	0.894	-0.104	0.053
IC：不规则系数	0.037	-0.139	0.128	0.959

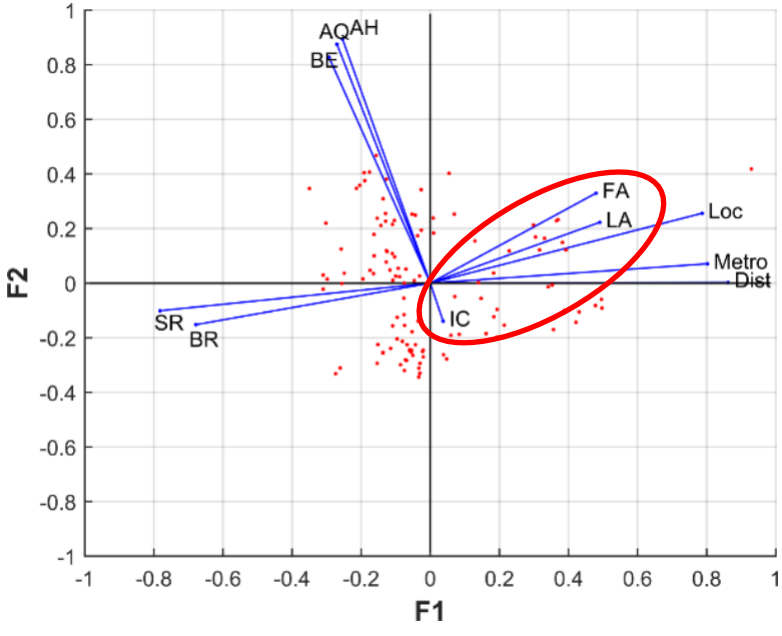
旋转因子载荷矩阵

	公因子			
	1	2	3	4
LA：用地面积	0.156	-0.044	0.966	-0.007
FA：建筑面积	0.175	0.070	0.962	0.028
Dist：中心距离	0.863	-0.167	0.111	0.110
Metro：地铁站距离	0.825	-0.087	0.097	0.085
BE：周边建成环境	-0.039	0.894	-0.098	-0.081
BR：写字楼租金	-0.731	0.001	-0.061	0.170
SR：商铺租金	-0.819	0.052	-0.082	-0.063
Loc：区位	0.837	0.075	0.148	-0.094
AQ：建筑质量	-0.086	0.916	0.093	0.047
AH：建筑高度	-0.037	0.934	0.033	-0.048
IC：不规则系数	-0.006	-0.054	0.016	0.976

因子分析案例

③ **因子旋转与命名**：通过最大方差法得到更加清晰的旋转因子荷载矩阵。

每条蓝线代表一个原始变量，其端点是该变量在横轴F_1、纵轴F_2这两个公因子上的荷载。

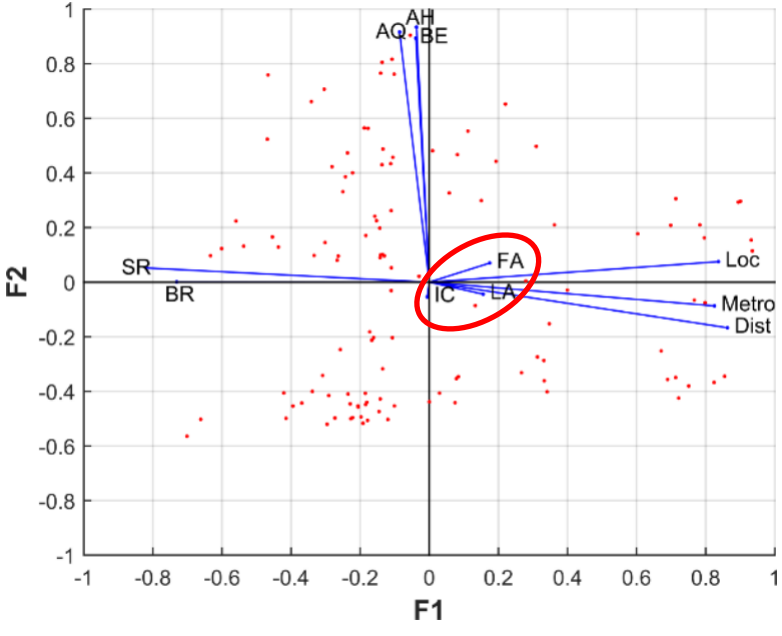


基于初始因子荷载

因子分析案例

③ **因子旋转与命名**：通过最大方差法得到更加清晰的旋转因子荷载矩阵。

每条蓝线代表一个原始变量，其端点是该变量在横轴F_1、纵轴F_2这两个公因子上的荷载。



基于**旋转**因子荷载

因子分析案例

③ **因子旋转与命名**：找到与公因子相关性最强的变量，用它们为公因子命名。

旋转因子载荷矩阵

	公因子			
	1	2	3	4
LA：用地面积	0.156	-0.044	0.966	-0.007
FA：建筑面积	0.175	0.070	0.962	0.028
Dist：中心距离	0.863	-0.167	0.111	0.110
Metro：地铁站距离	0.825	-0.087	0.097	0.085
BE：周边建成环境	-0.039	0.894	-0.098	-0.081
BR：写字楼租金	-0.731	0.001	-0.061	0.170
SR：商铺租金	-0.819	0.052	-0.082	-0.063
Loc：区位	0.837	0.075	0.148	-0.094
AQ：建筑质量	-0.086	0.916	0.093	0.047
AH：建筑高度	-0.037	0.934	0.033	-0.048
IC：不规则系数	-0.006	-0.054	0.016	0.976

F_1： 区位因子

- 与中心距离、地铁站距离、区位高度正相关，与写字楼租金、商铺租金高度负相关。
- F_1得分越高，区位条件越差。

因子分析案例

③ **因子旋转与命名**：找到与公因子相关性最强的变量，用它们为公因子命名。

旋转因子载荷矩阵

	公因子			
	1	2	3	4
LA：用地面积	0.156	-0.044	0.966	-0.007
FA：建筑面积	0.175	0.070	0.962	0.028
Dist：中心距离	0.863	-0.167	0.111	0.110
Metro：地铁站距离	0.825	-0.087	0.097	0.085
BE：周边建成环境	-0.039	0.894	-0.098	-0.081
BR：写字楼租金	-0.731	0.001	-0.061	0.170
SR：商铺租金	-0.819	0.052	-0.082	-0.063
Loc：区位	0.837	0.075	0.148	-0.094
AQ：建筑质量	-0.086	0.916	0.093	0.047
AH：建筑高度	-0.037	0.934	0.033	-0.048
IC：不规则系数	-0.006	-0.054	0.016	0.976

F_2：建设品质因子

- 与周边建成环境、建筑质量、建筑高度正相关。
- F_2得分越高，建设品质越高。

因子分析案例

③ **因子旋转与命名**：找到与公因子相关性最强的变量，用它们为公因子命名。

旋转因子载荷矩阵

	公因子			
	1	2	3	4
LA：用地面积	0.156	-0.044	0.966	-0.007
FA：建筑面积	0.175	0.070	0.962	0.028
Dist：中心距离	0.863	-0.167	0.111	0.110
Metro：地铁站距离	0.825	-0.087	0.097	0.085
BE：周边建成环境	-0.039	0.894	-0.098	-0.081
BR：写字楼租金	-0.731	0.001	-0.061	0.170
SR：商铺租金	-0.819	0.052	-0.082	-0.063
Loc：区位	0.837	0.075	0.148	-0.094
AQ：建筑质量	-0.086	0.916	0.093	0.047
AH：建筑高度	-0.037	0.934	0.033	-0.048
IC：不规则系数	-0.006	-0.054	0.016	0.976

F_3：建设规模因子

- 与用地面积、建筑面积正相关。
- F_3得分越高，建设规模越高。

因子分析案例

③ **因子旋转与命名**：找到与公因子相关性最强的变量，用它们为公因子命名。

旋转因子载荷矩阵

	公因子			
	1	2	3	4
LA：用地面积	0.156	-0.044	0.966	-0.007
FA：建筑面积	0.175	0.070	0.962	0.028
Dist：中心距离	0.863	-0.167	0.111	0.110
Metro：地铁站距离	0.825	-0.087	0.097	0.085
BE：周边建成环境	-0.039	0.894	-0.098	-0.081
BR：写字楼租金	-0.731	0.001	-0.061	0.170
SR：商铺租金	-0.819	0.052	-0.082	-0.063
Loc：区位	0.837	0.075	0.148	-0.094
AQ：建筑质量	-0.086	0.916	0.093	0.047
AH：建筑高度	-0.037	0.934	0.033	-0.048
IC：不规则系数	-0.006	-0.054	0.016	0.976

F_4：形态因子

- 仅与不规则系数高度正相关。
- F_4得分越高，用地形态越规整。

因子分析案例

③ 因子旋转与命名：找到与公因子相关性最强的变量，用它们为公因子命名。

旋转因子载荷矩阵

	公因子			
	区位	品质	规模	形态
LA：用地面积	0.156	-0.044	0.966	-0.007
FA：建筑面积	0.175	0.070	0.962	0.028
Dist：中心距离	0.863	-0.167	0.111	0.110
Metro：地铁站距离	0.825	-0.087	0.097	0.085
BE：周边建成环境	-0.039	0.894	-0.098	-0.081
BR：写字楼租金	-0.731	0.001	-0.061	0.170
SR：商铺租金	-0.819	0.052	-0.082	-0.063
Loc：区位	0.837	0.075	0.148	-0.094
AQ：建筑质量	-0.086	0.916	0.093	0.047
AH：建筑高度	-0.037	0.934	0.033	-0.048
IC：不规则系数	-0.006	-0.054	0.016	0.976

X1	我们小区是经过精心设计的。
X2	我们小区的住宅楼别具一格。
X3	我们小区的户外景化与众不同。
X4	大家从不会混淆我们小区和其他小区。
X5	我们小区很少发生犯罪事件。
X6	我不担心家中财物被盗。
X7	我们小区的安保服务令人放心。
X8	即使深夜回家，我也不觉得害怕。
X9	我对我们社区有很深的依恋。
X10	我很了解我们社区的情况。
X11	我希望付出时间，为社区做贡献。
X12	我和社区里的邻居相处得很愉快。

因子分析案例

③ 因子旋转与命名

在已确定公因子数量为k的前提下，因子旋转不会改变k个公因子解释原始变量的总比例，但是会改变每个因子的方差贡献率。

公因子	初始			旋转后		
	解释的方差	方差贡献率（%）	累积方差贡献率（%）	解释的方差	方差贡献率（%）	累积方差贡献率（%）
1	3.771	34.283	34.283	3.396	30.870	30.870
2	2.529	22.992	57.274	2.565	23.314	54.184
3	1.613	14.666	71.940	1.932	17.561	71.745
4	1.004	9.127	81.067	1.025	9.321	81.067

本例中，4个公因子在旋转前后均解释了 81.067%的方差。

因子分析案例

④ **因子得分**：根据因子得分矩阵，推算因子得分。

因子得分矩阵

	公因子			
	1	2	3	4
LA：用地面积	-0.070	-0.035	0.528	-0.036
FA：建筑面积	-0.058	0.014	0.519	0.006
Dist：中心距离	0.260	-0.024	-0.042	0.102
Metro：地铁站距离	0.253	0.005	-0.047	0.083
BE：周边建成环境	0.039	0.353	-0.071	-0.021
BR：写字楼租金	-0.229	-0.020	0.051	0.163
SR：商铺租金	-0.254	-0.018	0.055	-0.064
Loc：区位	0.257	0.058	-0.018	-0.085
AQ：建筑质量	0.002	0.363	0.039	0.100
AH：建筑高度	0.026	0.368	0.001	0.011
IC：不规则系数	0.002	0.039	-0.017	0.959

$$F_1 = -0.070LA - 0.058FA + 0.260Dist + 0.253Metro + 0.039BE - 0.229OR - 0.254SR + 0.257Loc + 0.002AQ + 0.026AH + 0.002IC$$

$$F_2 = -0.035LA + 0.014FA - 0.024Dist + 0.005Metro + 0.353BE - 0.020OR - 0.018SR + 0.058Loc + 0.363AQ + 0.368AH + 0.039IC$$

$$F_3 = 0.528LA + 0.519FA - 0.042Dist - 0.047Metro - 0.071BE + 0.051OR + 0.055SR - 0.018Loc + 0.039AQ + 0.001AH - 0.017IC$$

$$F_4 = -0.036LA + 0.006FA + 0.102Dist + 0.083Metro - 0.021BE + 0.163OR - 0.064SR - 0.085Loc + 0.100AQ + 0.011AH + 0.959IC$$

总结

- 当变量较多，信息高度重叠时，数据降维可以通过提取少量的主成分或公因子，保留原始变量的绝大多数信息，从而简化数据。
- 主成分/公因子之间相互独立，便于开展后续分析。
- 有利于更好地理解数据：因子命名、可视化……

	主成分分析	**因子分析**
基本原理	**把原始变量转化成主成分（X→PC）**	**以公因子解释原始变量（F→X）**
应用优势	侧重单纯的**数据简化**	侧重发现有意义的因子化**数据结构**
是否有假设	**无** （数据未通过适宜性检验也可分析，但影响主成分替代原始变量的效果）	**有**，假设“变量背后有潜在因子结构”且满足特殊因子不相关等特定条件 （数据须先通过适宜性检验）
是否需旋转	一般**不旋转** 旋转可增强解释性，但也破坏信息度	**要旋转** 旋转可使因子意义明确，以便于命名
求解方法	简单易操作	除主成分法，还有极大似然法等