

# 主成分分析

# 本章主要内容

1. 主成分分析基本原理
2. 主成分分析数学模型
3. 主成分分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 本章主要内容

1. 主成分分析基本原理
2. 主成分分析数学模型
3. 主成分分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 主成分分析基本原理

- 主成分分析, principal components analysis, 简称PCA
- 作用: 是将多个指标化为少数几个综合指标的一种统计分析方法, 本质就是降维。

举个例子, 学生的学习成绩, 按照科目可以分为语文、数学、英语、物理、化学、历史、地理等, 那么描述一个学生的学习成绩的好坏, 需要多个科目的成绩, 即多个维度。

为了简化问题, 可以将多个科目成绩归纳为文科成绩和理科成绩这两个维度, 这个就表示主成分分析。

本质: 通过主成分分析, 将反映学生学习成绩的多个维度 (多个科目的成绩) 降维成两个维度。

# 主成分分析基本原理

例如，反映城市综合发展水平有12个指标，其中包括：

**8个社会经济指标**，分别为：非农业人口数（万人）、工业总产值（万元）、货运总量（万吨）、批发零售住宿餐饮业从业人数（万人）、地方政府预算内收入（万元）、城乡居民年底储蓄余额（万元）、在岗职工人数（万人）、在岗职工工资总额（万元）；

**4个城市公共设施水平的指标**：人均居住面积（平方米）、每万人拥有公共汽车数（辆）、人均拥有铺装道路面积（平方米）和人均公共绿地面积（平方米）。

为了简化问题，可以将这12个指标归纳为：**城市规模及经济水平、城市基础设施和城市人均居住面积**这3个指标，这个就表示主成分分析。

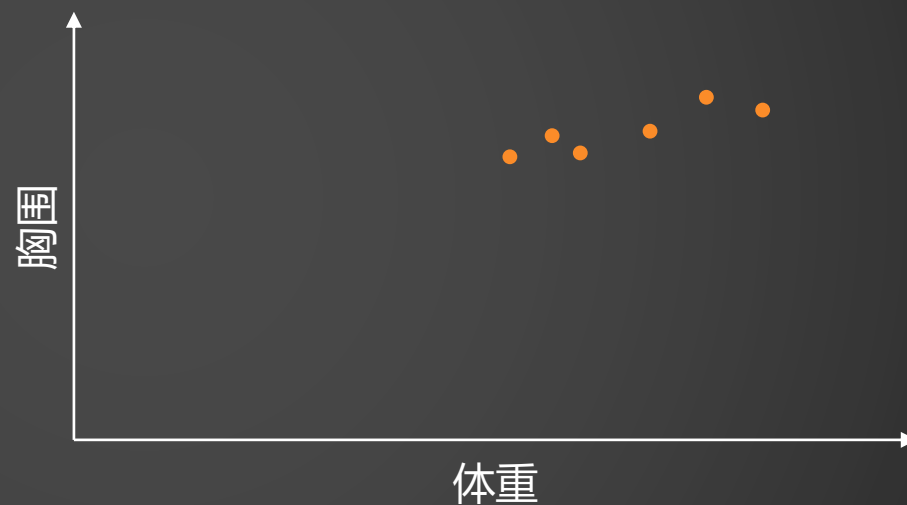
本质：通过主成分分析，将反映城市综合发展水平的12个维度降维成3个维度。

# 主成分分析：小例子

例如，有一些关于体重和胸围的数据。

体重	胸围
41	72
34	71
49	77
36	67
45	80
31	66

体重与胸围散点图



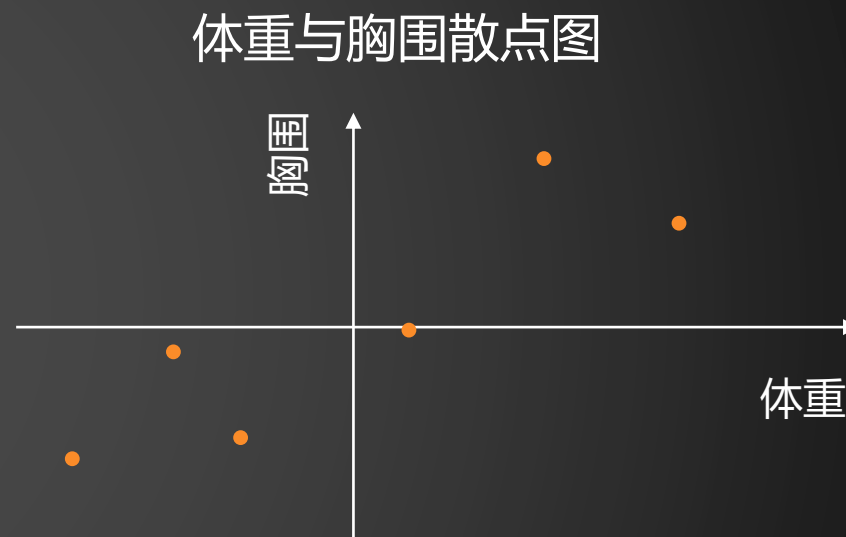
## 主成分分析：小例子

为了使数据更加直观，将这些数据中心化，即每个数据减去它们的均值。

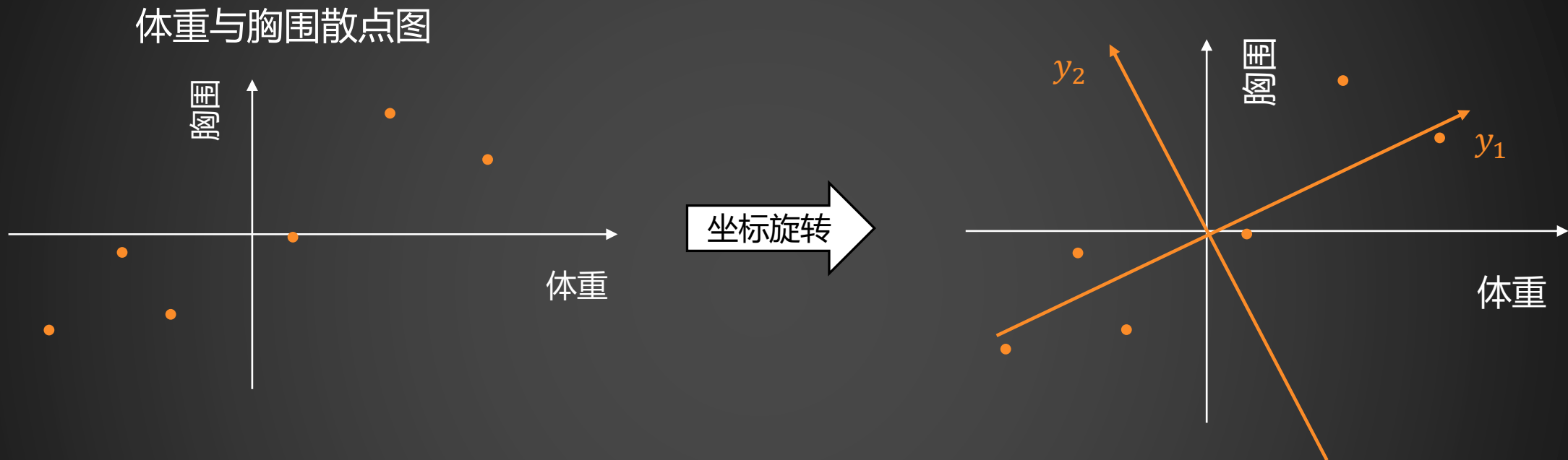
体重	胸围
41	72
34	71
49	77
36	67
45	80
31	66
39.33	72.17



体重	胸围
1.67	-0.17
-5.33	-1.17
9.67	4.83
-3.33	-5.17
5.67	7.83
-8.33	-6.17



## 主成分分析：小例子



坐标旋转后，这些数据点的大小主要体现在 $y_1$ 上，即新的变量 $y_1$ 就可以表示原始数据点的绝大部分信息。

这样，就可以原先两个变量（体重： $x_1$ ，胸围： $x_2$ ）代表的信息用一个变量（ $y_1$ ）来表示，即二维变成一维，这就是降维的过程。

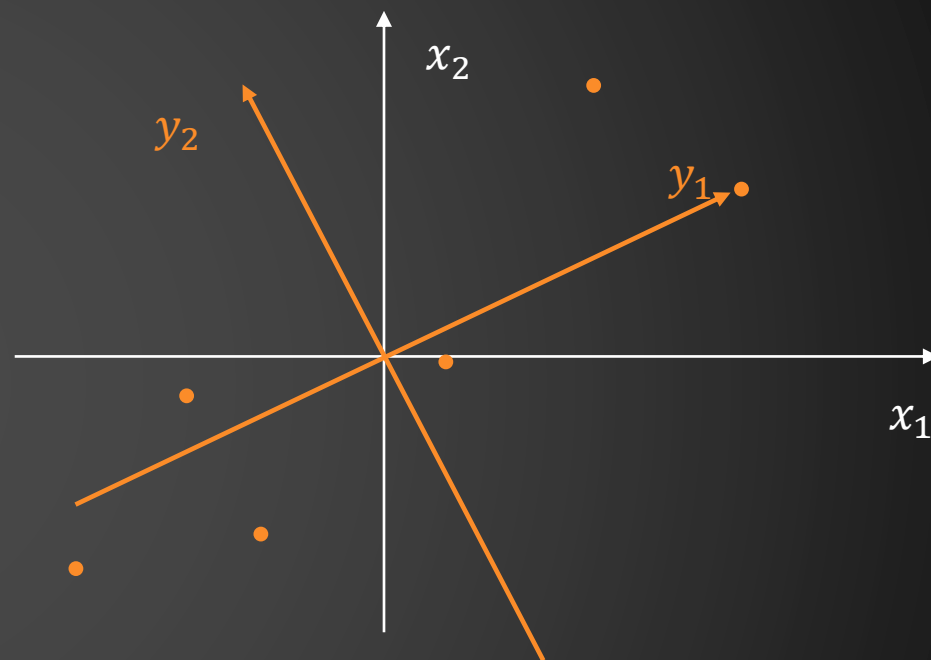


# 主成分的几何意义

- 主成分分析的过程其实就是坐标系旋转的过程，坐标旋转的公式如下。

$$\begin{cases} Y_1 = X_1 \cos\theta + X_2 \sin\theta \\ Y_2 = -X_1 \sin\theta + X_2 \cos\theta \end{cases}$$

- $Y_1, Y_2$  表示原始变量  $X_1, X_2$  的主成分。



说明：

- 主成分就是原来变量的线性组合，有几个原始变量就有几个主成分。
- 多维变量的情形类似，只不过是一个高维椭球，无法直观地观察。

# 本章主要内容

1. 主成分分析基本原理
2. 主成分分析数学模型
3. 主成分分析案例：全国35个中心城市综合发展水平分析（SPSS）

## 主成分分析的数学模型

从数学上来说，主成分分析是将原始的 $p$ 个变量进行线性组合，作为新的变量。

设原始变量为 $X_1, X_2, \dots, X_p$ ，这 $p$ 个指标构成 $p$ 维随机向量， $X = (X_1, X_2, \dots, X_p)'$ 。

考虑以下线性变换：

$$\begin{cases} Y_1 = X_1 \cos \theta + X_2 \sin \theta \\ Y_2 = -X_1 \sin \theta + X_2 \cos \theta \end{cases} \rightarrow \begin{cases} Y_1 = u_{11}X_1 + u_{21}X_2 + \dots + u_{p1}X_p \\ Y_2 = u_{12}X_1 + u_{22}X_2 + \dots + u_{p2}X_p \\ \dots\dots\dots \\ Y_p = u_{1p}X_1 + u_{2p}X_2 + \dots + u_{pp}X_p \end{cases}$$

接下来，要考虑这个线性变换的约束条件是什么？即系数 $u_i$ 与 $Y_i$ 要满足什么条件？

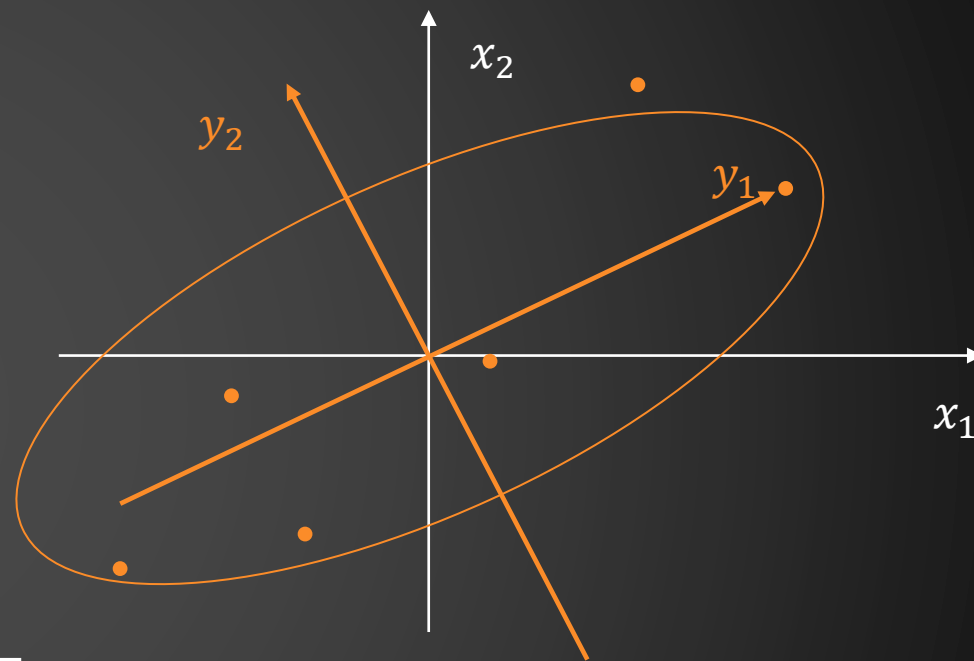
# 主成分分析的数学模型

考虑坐标旋转的线性变换。

$$\begin{cases} Y_1 = X_1 \cos\theta + X_2 \sin\theta \\ Y_2 = -X_1 \sin\theta + X_2 \cos\theta \end{cases}$$

观察以上式子，满足以下三个条件：

1. 设  $u_1 = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$ ，则  $u_1' u_1 = 1$ ，即  $u_i' u_i = 1 (i = 1, 2)$
2.  $Y_i$  与  $Y_j$  相互无关 ( $i \neq j, i, j = 1, 2$ )，因为  $Y_1$  与  $Y_2$  相互垂直。
3.  $Y_1$  是  $X_1, X_2$  的一切满足原则1的线性组合中方差最大者， $Y_2$  是与  $Y_1$  不相关的  $X_1, X_2$  所有线性组合中方差最大者。



## 主成分分析的数学模型

$$\begin{cases} Y_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\ Y_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\ \cdots \cdots \\ Y_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p \end{cases}$$

理论上，这个线性变换可以任意地对原始变量进行线性变换，不同的线性变换得到的综合变量 $Y$ 也不同，因此，为了取得更好的效果，将线性变换约束在以下三个原则之下。

1.  $u_i' u_i = 1 (i = 1, 2, \cdots, p)$
2.  $Y_i$ 与 $Y_j$ 相互无关 ( $i \neq j; i, j = 1, 2, \cdots, p$ )
3.  $Y_1$ 是 $X_1, X_2, \cdots, X_p$ 的一切满足原则1的线性组合中方差最大者； $Y_2$ 是与 $Y_1$ 不相关的 $X_1, X_2, \cdots, X_p$ 的所有线性组合中方差最大者； $\cdots \cdots Y_p$ 是与 $X_1, X_2, \cdots, X_{p-1}$ 都不相关的 $X_1, X_2, \cdots, X_p$ 的所有线性组合中方差最大者。



## 主成分的求解：拉格朗日乘子法

设 $p$ 维随机向量 $X = (X_1, X_2, \dots, X_p)'$ , 均值:  $E(X) = 0$ , 协方差阵 $D(X) = \Sigma > 0$

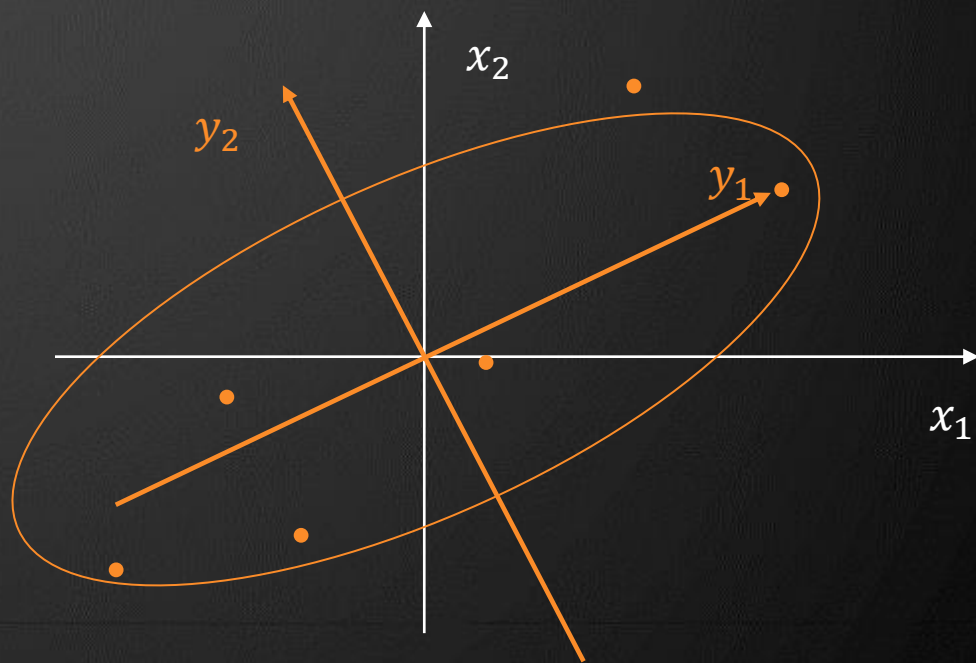
将 $Y_1 = u_{11}X_1 + u_{21}X_2 + \dots + u_{p1}X_p$ 写成矩阵相乘的形式,

有 $Y_1 = u_1'X$ , 其中 $u_1 = (u_{11}, u_{21}, \dots, u_{p1})'$

求解第一主成分 $Y_1 = u_1'X$ 的问题, 其实就是求以下问题:

$$\begin{cases} \max Var(Y_1) \\ s.t. \quad u_1'u_1 = 1 \end{cases}$$

所以, 将求解主成分的问题转化成了一个条件极值问题。



## 主成分的求解：拉格朗日乘子法

对于这个条件极值问题，一般采用拉格朗日乘子法求解，令

$$\varphi(u_1) = \text{Var}(u_1'X) - \lambda(u_1'u_1 - 1) = u_1'\Sigma u_1 - \lambda(u_1'u_1 - 1)$$

分别对参数 $u_1$ 和 $\lambda$ 求偏导数，并令其等于零，有

$$\begin{cases} \frac{\partial \varphi}{\partial u_1} = 2(\Sigma - \lambda I)u_1 = 0 \\ \frac{\partial \varphi}{\partial \lambda} = u_1'u_1 - 1 = 0 \end{cases}$$

由上式可得

$$|\Sigma - \lambda I| = 0$$

所以，求解主成分其实就是求解协方差阵 $\Sigma$ 的特征值和特征向量的问题。

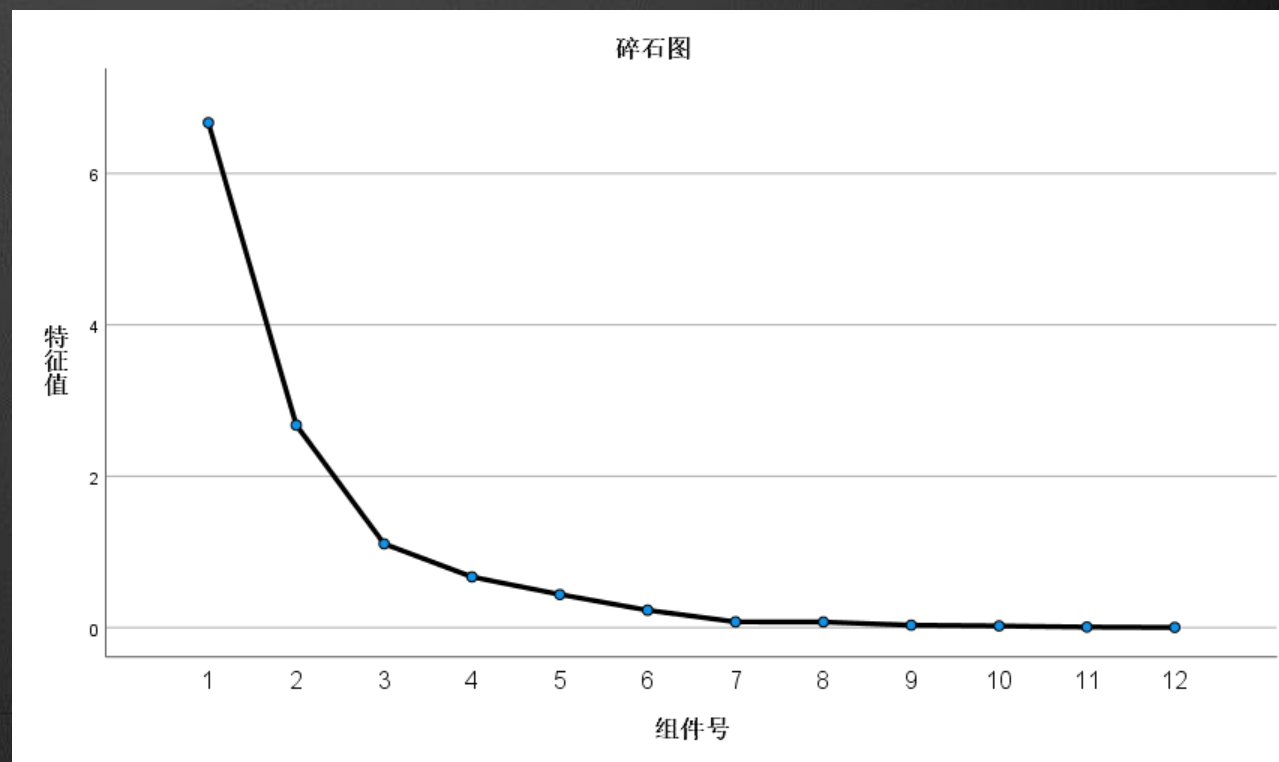
## 主成分的求解：一个重要的定理

定理：设随机向量  $X = (X_1, X_2, \dots, X_p)'$ ，协方差矩阵为  $\Sigma$ ， $\Sigma$  的特征值为  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ ， $u_1, u_2, \dots, u_p$  为特征根对应的单位正交特征向量，则  $X$  第  $i$  个主成分为：

$$Y_i = u_i' X$$

证明：略。

结论：由原始数据的协方差矩阵出发，  
可以求出主成分！





# 贡献率

定义：称 $\alpha_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$  ( $k = 1, 2, \cdots, p$ )为主成分 $Y_k$ 的方差贡献率，简称贡献率。

贡献率描述的是某个主成分提取了原始变量的多少信息，贡献率越大，该主成分对原始变量的解释力就越强。

称 $\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i}$ 为主成分 $Y_1, Y_2, \cdots, Y_m$ 的累计贡献率。

累计贡献率描述的是前几个主成分总共提取了原始变量的多少信息。

通常，使得选取的前 $m$ 个主成分的累计贡献率达到85%以上即可，即

$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i} \geq 85\%$$

这样，既能使损失的信息不多，又能达到减少变量的目的。

# 贡献率

实际应用中，一般用工具如SPSS来做主成分分析。SPSS会自动给出每个主成分的贡献率，以及前几个主成分的累计贡献率，如下图所示。

总方差解释						
成分	总计	初始特征值		提取载荷平方和		
		方差百分比	累积 %	总计	方差百分比	累积 %
1	6.671	55.589	55.589	6.671	55.589	55.589
2	2.675	22.293	77.882	2.675	22.293	77.882
3	1.107	9.224	87.105	1.107	9.224	87.105
4	.670	5.581	92.686			
5	.437	3.642	96.329			
6	.229	1.909	98.238			
7	.076	.631	98.869			
8	.073	.612	99.482			
9	.032	.264	99.746			
10	.021	.179	99.925			
11	.007	.056	99.981			
12	.002	.019	100.000			

提取方法：主成分分析法。

## 主成分得分

$$\begin{cases} Y_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\ Y_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\ \dots\dots\dots \\ Y_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p \end{cases}$$

计算主成分得分，就是将原始变量 $X$ 的数据代入上述线性变换的方程，得到 $Y$ 值的过程。

所以，需要先求出 $X$ 前面的系数 $u_1, u_2, \dots, u_p$ ，其中 $u_1 = (u_{11}, u_{21}, \dots, u_{p1})'$ ，即主成分系数。

下面，说说如何求主成分系数。

# 如何计算主成分的系数？

用SPSS做主成分分析，会得出一个成分矩阵，即因子载荷矩阵（后面会讲），如下图所示。

第*i*个主成分的系数：成分矩阵中的第*i*列的每个元素除以第*i*个特征根的平方根 $\sqrt{\lambda_i}$

接着，就可以写出各个主成分用**标准化后的原始变量**表示的表达式，用Excel可以完成计算。

总方差解释						
成分	总计	初始特征值		提取载荷平方和		
		方差百分比	累积 %	总计	方差百分比	累积 %
1	6.671	55.589	55.589	6.671	55.589	55.589
2	2.675	22.293	77.882	2.675	22.293	77.882
3	1.107	9.224	87.105	1.107	9.224	87.105
4	.670	5.581	92.686			
5	.437	3.642	96.329			
6	.229	1.909	98.238			
7	.076	.631	98.869			
8	.073	.612	99.482			
9	.032	.264	99.746			
10	.021	.179	99.925			
11	.007	.056	99.981			
12	.002	.019	100.000			

提取方法：主成分分析法。

成分矩阵 <sup>a</sup>			
	组件		
	1	2	3
x1	.878	-.325	.143
x2	.854	.254	.265
x3	.830	-.210	.330
x4	.789	-.203	-.403
x5	.956	.062	.130
x6	.984	-.032	-.058
x7	.933	-.207	-.143
x8	.971	-.039	-.163
x9	.059	.465	.727
x10	.207	.898	-.131
x11	.243	.927	-.052
x12	.241	.698	-.359

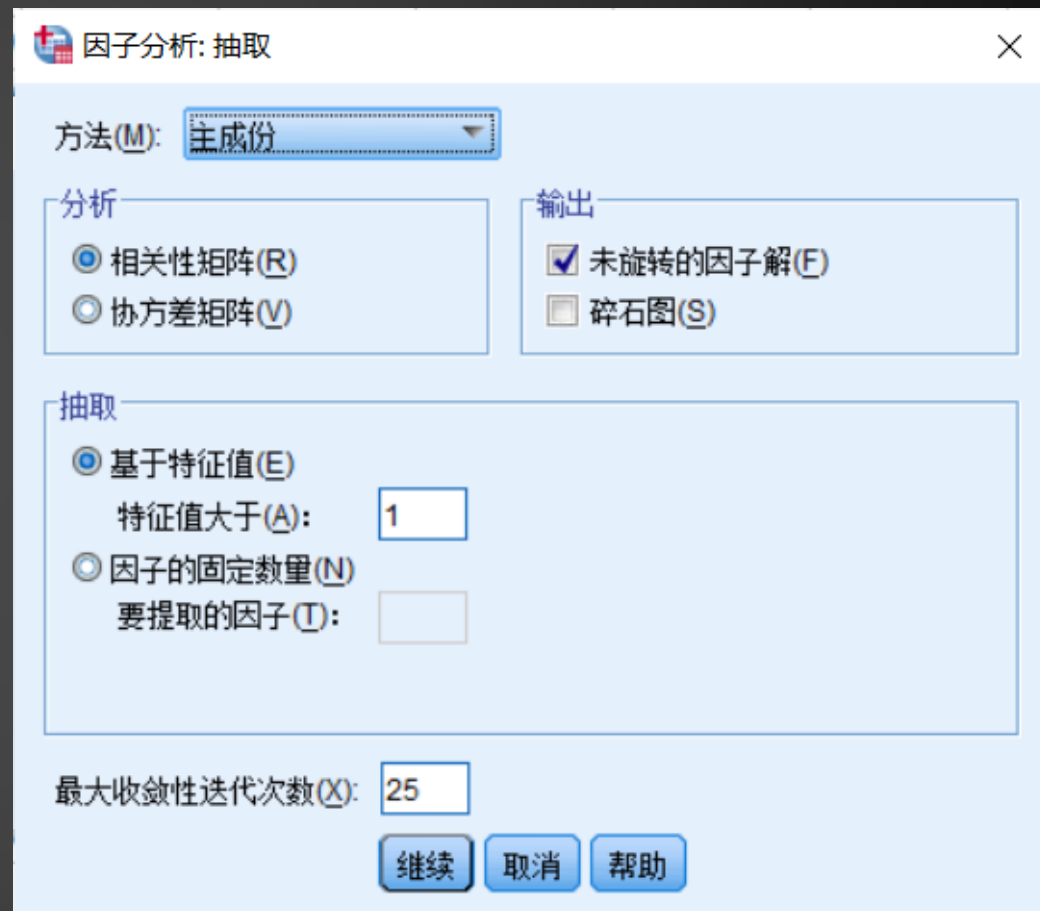
## 由相关阵出发求解主成分

通过前面的讲解，我们知道可以通过原始数据的协方差阵来求解主成分。

而，原始变量 $X_1, X_2, \dots, X_p$ 的相关矩阵其实就是对原始变量标准化后的协方差矩阵，所以也可以由相关矩阵出发求解主成分。

究竟是选择从协方差阵，还是从相关矩阵出发求解主成分？

当原始数据的数量级差异较大或者量纲不同时，说明原始数据需要进行标准化，要选择从相关矩阵出发求解主成分，否则，选择从协方差阵出发求解主成分。



因子分析: 抽取

方法(M): 主成份

分析

- ☒ 相关性矩阵(R)
- ☐ 协方差矩阵(V)

输出

- ☒ 未旋转的因子解(F)
- ☐ 碎石图(S)

抽取

- ☒ 基于特征值(E)
  - 特征值大于(A): 1
- ☐ 因子的固定数量(N)
  - 要提取的因子(I):

最大收敛性迭代次数(X): 25

继续 取消 帮助

# 本章主要内容

1. 主成分分析基本原理
2. 主成分分析数学模型
3. 主成分分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 主成分分析案例：全国35个中心城市综合发展水平分析

选取反映城市综合发展水平的12个指标作为原始变量，其中包括8个社会经济指标和4个城市公共设施水平的指标。

- 8个社会经济指标：

非农业人口数（万人）、工业总产值（万元）、货运总量（万吨）、批发零售住宿餐饮业从业人数（万人）、地方政府预算内收入（万元）、城乡居民年底储蓄余额（万元）、在岗职工人数（万人）、在岗职工工资总额（万元）。

- 4个城市公共设施水平的指标：

人均居住面积（平方米）、每万人拥有公共汽车数（辆）、人均拥有铺装道路面积（平方米）、人均公共绿地面积（平方米）。

运用SPSS软件，对全国35个中心城市的综合发展水平进行主成分分析。



- x1：非农业人口数（万人）
- x2：工业总产值（万元）
- x3：货运总量（万吨）
- x4：批发零售住宿餐饮业从业人数（万人）
- x5：地方政府预算内收入（万元）
- x6：城乡居民年底储蓄余额（万元）
- x7：在岗职工人数（万人）
- x8：在岗职工工资总额（万元）
- x9：人均居住面积（平方米）
- x10：每万人拥有公共汽车数（辆）
- x11：人均拥有铺装道路面积（平方米）
- x12：人均公共绿地面积（平方米）

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	地区	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
2	深圳	122.39	52451037	6792.66	10.84	2908370	21994500	104.98	3259900	21	114.91	47.29	177.62
3	广州	493.32	40178324	28859.45	21.47	2747707	37273276	182.16	5247087	17	11.16	12.76	178.76
4	南京	391.67	25093816	14804.68	7.62	1364788	11336202	87.91	1950742	16	9.06	12.13	136.72
5	北京	830.8	38103630	30671.14	127.4	5925388	64413910	434.15	10989365	15	17.3	8.56	44.94
6	合肥	160.18	5348605	4640.84	3.39	358694	3592488	37.88	526577	17	14.11	15.72	28.74
7	贵阳	165.27	3569419	5317.55	5.75	403855	3449487	54.53	664234	16	9.37	3.11	105.35
8	银川	79.2	1464867	2127.17	1.65	122605	1930771	29.12	393035	15	9.26	10.43	40.21
9	厦门	83.74	13201500	3054.82	2.83	701456	3971559	54.78	1042111	20	15.5	8.15	26.44
10	南宁	167.99	2083763	5893.09	4.95	362435	4514961	50.79	668976	18	9.91	9.32	35.12
11	青岛	329.96	25588695	30552.6	6.72	1201398	9084693	104.55	1603305	15	14.78	11.41	35.78
12	乌鲁木齐	142.94	3110943	12754.02	3.94	409119	4203000	47.42	782873	19	22.89	6.49	20.53
13	大连	297.48	15468641	21081.47	6.6	1105405	13101986	82.13	1442215	14	13.79	6.24	40.21
14	长沙	205.83	5339304	10630.5	6.31	598930	7048500	60.04	1019924	18	10.09	9.1	29.1
15	福州	205.43	12889573	8250.39	4.69	674522	8762245	71.3	1073262	18	9.65	7.9	31.6
16	呼和浩特	97.81	2407794	4155.1	2	205779	2554496	28.9	407963	18	3.81	8.92	26.58
17	南昌	195.46	4149169	4454.45	3.62	314094	4828029	49.79	692717	17	7.37	7.67	23.98
18	哈尔滨	454.52	7215089	9517.8	24.99	763600	11536951	168.83	2102165	14	12.75	6.34	18.51
19	昆明	205.34	5809573	12337.86	7.07	601101	7085278	73.34	1045469	15	15.33	4.49	23.33
20	长春	313.05	15115270	10891.98	6.94	459709	8313564	89.7	1244167	15	11.87	7.03	18.75
21	宁波	168.81	26302862	13797.38	4.8	1394162	10596339	59.88	1418635	17	9.88	6.81	17.65
22	兰州	175.54	5215490	5580.8	3.7	205660	4683830	54.91	740661	15	10.33	6.3	11.22
23	济南	297.21	13185425	14354.4	6.6	761054	7583525	78.38	1256160	19	7.77	10.62	19.54
24	郑州	249.72	9270494	7846.91	8.77	658737	10484859	83.99	1137056	19	10.11	7.63	17.77
25	西宁	105.13	1148959	2037.15	1.24	84397	1749293	20.6	301364	17	11.47	4.92	14.2
26	杭州	263.67	32025226	16815.2	8.36	1503888	14664200	75.72	1867776	17	8.93	6.5	23.19
27	沈阳	440.6	10643612	14635.74	7.3	810889	14229575	101.7	1521548	15	9.32	6.7	28.36
28	石家庄	331.33	11981505	10008.48	8.07	493429	10444919	86.74	1067432	18	7.23	8.28	21.56
29	成都	386.23	9700976	28798.2	8.06	895752	14944197	124.03	1894496	17	8.95	10.17	25.59
30	海口	76.05	2025643	3304.4	2.72	122541	2843664	22.97	340392	20	5.09	7.07	15.79
31	太原	222.63	5183200	15248.11	2.43	333473	6601300	74.55	945212	16	5.06	7.88	20.58
32	上海	1041.39	103000000	63861	35.22	8992850	60546000	281.51	7686511	19	14.57	12.92	19.11
33	西安	312.88	6386627	9392	12.21	648037	12105607	113.73	1535896	15	7.32	4.48	8.82
34	天津	549.74	40496103	34679	15.38	2045295	18253200	174.5	3254148	18	7.99	7.23	17.45
35	武汉	474.98	13344938	16610.34	13.58	804368	12855341	136.08	1868350	17	6.87	4.16	8.34
36	重庆	753.92	15889928	32450.2	12.83	1615618	18965569	203.79	2535070	21	4.94	4.24	10.8



# SPSS主成分分析

主要操作步骤:

1. 读取数据: 读取Excel数据文件
2. SPSS菜单选择: 【分析】 - 【降维】 - 【因子分析】
3. 主成分分析相关选项: 选择变量、勾选相关系数、碎石图等
4. 解释分析结果
5. 主成分命名、主成分得分

# 因子分析

# 本章主要内容

1. 因子分析基本原理
2. 因子分析数学模型
3. 因子载荷的求解、因子旋转、因子得分
4. 因子分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 本章主要内容

1. 因子分析基本原理
2. 因子分析数学模型
3. 因子载荷的求解、因子旋转、因子得分
4. 因子分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 因子分析基本原理

因子分析, factor analysis, 由Charles Spearman于1904年首次提出的, 与主成分分析类似, 因子分析是要找出少数几个新的变量来代替原始变量, 本质就是降维。

例如, 心理学家瑟斯登对56项测验的得分进行因子分析, 得出了7种主要智力因子: 词语理解能力、语言流畅能力、计数能力、空间能力、记忆力、知觉速度和推理能力。

例如, 调查青年对婚姻家庭的态度, 抽取了 $n$ 个青年回答50个问题的问卷, 经过因子分析, 这些问题可以归纳为如下几个方面: 对相貌的重视、对孩子的观点、对老人的态度等, 每一个方面就是一个因子。

因子分析可以看作是主成分分析的推广和扩展, 但它对问题的研究更深入、更细致一些, 主成分分析可以看作是因子分析的一个特例。

# 因子分析：小例子

假设现在有一些中学生的身体指标数据，共有四个指标：  
身高（X1）、体重（X2）、胸围（X3）和坐高（X4）  
求出相关系数矩阵如下。

	X1	X2	X3	X4
X1	1			
X2	0.863162	1		
X3	0.732112	0.896506	1	
X4	0.920462	0.882731	0.782883	1

根据相关性，可以将X1、X4归为一个因子：大小因子  
可以将X2、X3归为另一个因子：胖瘦因子  
原先的4个变量简化为两个。

序号	X1	X2	X3	X4
1	148	41	72	78
2	139	34	71	76
3	160	49	77	86
4	149	36	67	79
5	159	45	80	86
6	142	31	66	76
7	153	43	76	83
8	150	43	77	79
9	151	42	77	80
10	139	31	68	74
11	140	29	64	74
12	161	47	78	84
13	158	49	78	83
14	140	33	67	77
15	137	31	66	73

说明：这里仅展示前15条记录。

# 本章主要内容

1. 因子分析基本原理
2. 因子分析数学模型
3. 因子载荷的求解、因子旋转、因子得分
4. 因子分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 因子分析数学模型

- 为了了解学生的学习能力，观测了 $n$ 个学生 $p$ 个科目（语文、英语、政治、代数、几何等）的成绩（分数），用 $X_1, X_2, \dots, X_p$ 表示 $p$ 个科目， $X_{(t)} = (x_{t1}, x_{t2}, \dots, x_{tp})' (t = 1, 2, \dots, n)$ 表示第 $t$ 个学生的 $p$ 个科目的成绩。
- 对这些资料进行归纳分析，发现各个科目的成绩由两部分组成，用数学语言表达为：

$$X_i = a_i F + \varepsilon_i \quad (i = 1, 2, \dots, p)$$

其中，

- $F$ 表示**公共因子**，代表智力的高低，称为智力因子，对所有科目都起作用。
- 系数 $a_i$ 表示第 $i$ 个科目成绩在智力因子上的体现，称为**因子载荷**。
- $\varepsilon_i$ 表示特殊因子，即某个科目成绩特有的因子，例如写作能力、抽象思维能力、记忆力等。

我们的目标是求出因子载荷，根据因子载荷获得公共因子，这就是一个简单的因子模型。



## 因子分析数学模型

一个前提：需要说明的是，不论是主成分分析还是因子分析，都需要原始变量之间有较强的相关性，才能够进行降维，即将多个变量综合为少数几个变量，所以变量间较强的相关性是进行因子分析的前提。

设有 $n$ 个样本，每个样本有 $p$ 个特征（变量）， $X = (X_1, X_2, \dots, X_p)'$ ，且这 $p$ 个特征之间有较强的相关性。

为了便于研究，将样本数据标准化，则标准化后的数据均值为0，方差为1，将标准化后的原始变量用 $X$ 表示，用 $F_1, F_2, \dots, F_m (m < p)$ 表示标准化后的公共因子，则对于每一个特征，有

$$X_i = a_i F + \varepsilon_i \longrightarrow X_1 = a_{11}F_1 + a_{12}F_2 + \dots + a_{1m}F_m + \epsilon_1$$

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \epsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \epsilon_2 \\ \dots\dots\dots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \epsilon_p \end{cases}$$

写成矩阵形式：

$$X = AF + \epsilon$$

其中，

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$$

模型中的系数 $a_{i1}, a_{i2}, \dots, a_{im}$ 相当于每个公共因子的权重，由于历史原因，系数 $a_{ij}$ 被叫作因子载荷，矩阵A称为因子载荷矩阵。

模型中的3个基本假定：

1. 公共因子 $F_i (i = 1, 2, \dots, m)$ 互不相关；
2. 特殊因子互不相关；
3. 公共因子与特殊因子互不相关。

接着，需要了解一下因子载荷的统计意义。

对于 $X_i$ 与其任意公共因子 $F_j$ ，有

$$\begin{aligned} cov(X_i, F_j) &= cov\left(\sum_{j=1}^m a_{ij} F_j + \epsilon_i, F_j\right) \\ &= cov\left(\sum_{j=1}^m a_{ij} F_j, F_j\right) + cov(\epsilon_i, F_j) \\ &= a_{ij} \end{aligned}$$

即 $a_{ij}$ 是 $X_i$ 与 $F_j$ 的协方差，又因为 $X_i$ 与 $F_j$ 都是经过标准化的变量，所以 $a_{ij}$ 也是 $X_i$ 与 $F_j$ 的相关系数。

# 因子分析数学模型

右图是SPSS给出的成分矩阵，其实就是因子载荷矩阵。

成分矩阵中的数值就是因子载荷 $a_{ij}$

从成分矩阵可以看出，得到3个公共因子。

- 因子1：与变量x6,x8,x5,x7,x1,x2,x3,x4的相关性较强。
- 因子2：与变量x11,x10,x12的相关性较强。
- 因子3：与变量x9的相关性较强。

根据成分矩阵，可以结合实际情况，进行因子命名。

成分矩阵 <sup>a</sup>			
	成分		
	1	2	3
x6	.984	-.032	-.058
x8	.971	-.039	-.163
x5	.956	.062	.130
x7	.933	-.207	-.143
x1	.878	-.325	.143
x2	.854	.254	.265
x3	.830	-.210	.330
x4	.789	-.203	-.403
x11	.243	.927	-.052
x10	.207	.898	-.131
x12	.241	.698	-.359
x9	.059	.465	.727

提取方法：主成分分析法。

a. 提取了 3 个成分。

# 因子分析数学模型

由前面知道，因子载荷 $a_{ij}$ 表示变量 $X_i$ 与公共因子 $F_j$ 的相关性，则用所有公共因子的因子载荷的平方和表示变量 $X_i$ 的**共同度**，定义如下：

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2$$

$h_i^2$ 称为变量 $X_i$ 的**共同度**，表示所有公共因子对原始变量 $X_i$ 的解释度，或者说公共因子从变量 $X_i$ 中提取出信息的大小，共同度越大，说明因子分析的效果越好。

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \epsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \epsilon_2 \\ \dots\dots\dots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \epsilon_p \end{cases}$$

公因子方差		
	初始	提取
x1	1.000	.898
x2	1.000	.864
x3	1.000	.843
x4	1.000	.825
x5	1.000	.934
x6	1.000	.974
x7	1.000	.934
x8	1.000	.971
x9	1.000	.748
x10	1.000	.867
x11	1.000	.921
x12	1.000	.675

提取方法：主成分分析法。

## 因子分析数学模型

以上是从横向看因子载荷矩阵，得到共同度，反映了所有公共因子对原始变量的解释度。

还可以从纵向看因子载荷矩阵，则有

$$g_j^2 = a_{1j}^2 + a_{2j}^2 + \cdots + a_{mj}^2 (j = 1, 2, \cdots, m)$$

$g_j^2$ 称为公共因子 $F_j$ 的**方差贡献**，反映了公共因子 $F_j$ 对所有原始变量 $X$ 的解释度，这个值越大，说明这个公共因子对原始变量 $X$ 的影响越大，可以根据方差贡献提取出前几个最有影响力的公共因子。

从前面可以看出，因子分析的关键在于求因子载荷矩阵，有了因子载荷矩阵，可以确定提取几个公共因子，以及根据因子载荷对因子进行命名。

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \epsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \epsilon_2 \\ \dots\dots\dots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \epsilon_p \end{cases}$$



# 本章主要内容

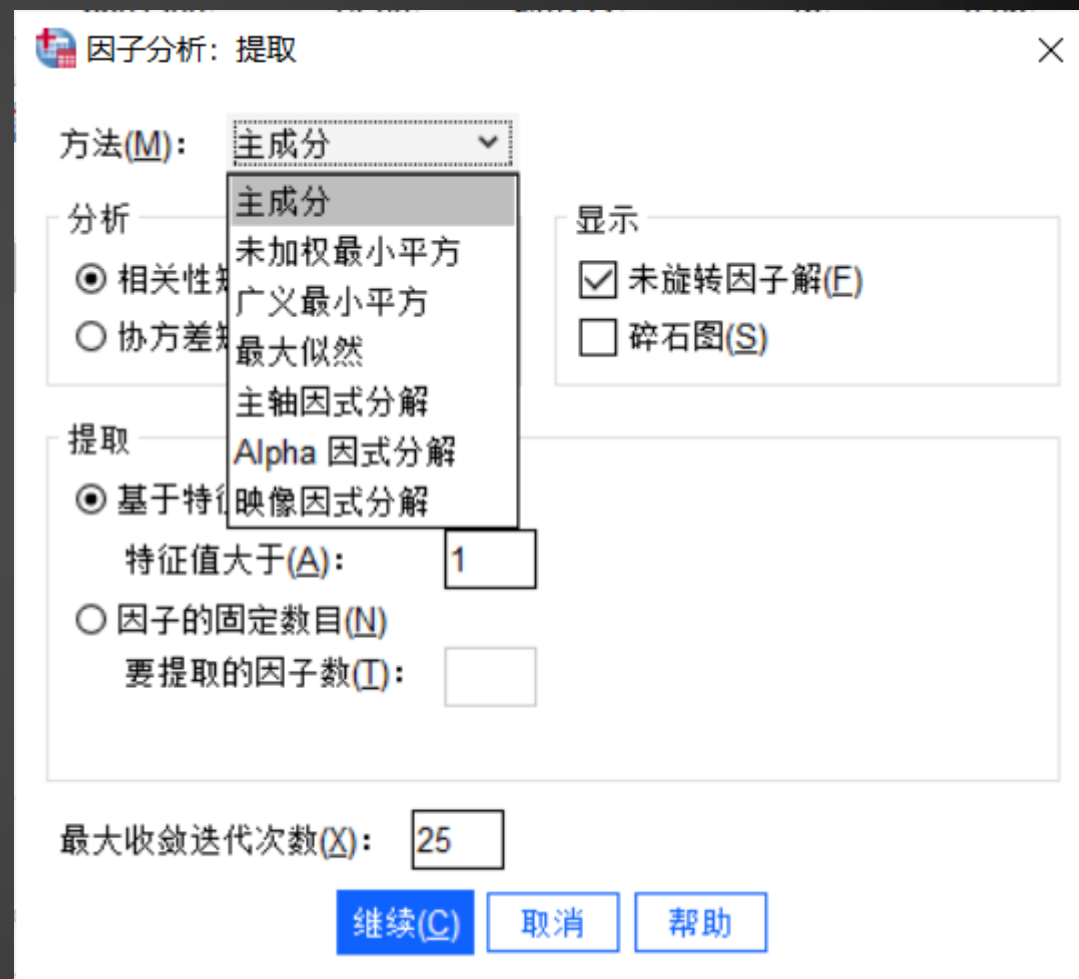
1. 因子分析基本原理
2. 因子分析数学模型
3. 因子载荷的求解、因子旋转、因子得分
4. 因子分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 因子载荷的求解

**主成分分析：**进行因子分析之前，先对数据进行一次主成分分析，然后将前几个主成分作为公共因子。

**主轴因子法：**假定 $m$ 个公共因子只能解释原始变量的部分方差，利用公共因子方差来代替相关矩阵主对角线上的元素1，并以这个新得到的矩阵（调整相关矩阵）为出发点，分别求解特征根与特征向量，从而得到因子解。

**极大似然估计法：**假定公共因子 $F$ 和特殊因子 $\varepsilon$ 服从正态分布，则能够得到因子载荷和特殊因子方差的极大似然估计，即利用极大似然估计法来估计因子载荷矩阵。



因子分析：提取

方法(M): 主成分

分析

- ☒ 相关性矩阵
- ☐ 协方差矩阵

提取

- ☒ 基于特征值
- ☐ 基于因子方差

特征值大于(A): 1

☐ 因子的固定数目(N)

要提取的因子数(I):

显示

- ☒ 未旋转因子解(E)
- ☐ 碎石图(S)

最大收敛迭代次数(X): 25

继续(C) 取消 帮助



# 因子旋转

找到公共因子后，需要知道每一个公共因子的意义，以便对实际问题进行分析。但是，有时候，得到的初始因子解各主因子的典型代表变量不是很突出，容易使因子的意义含糊不清。

成分矩阵 <sup>a</sup>			
	成分		
	1	2	3
x6	.984	-.032	-.058
x8	.971	-.039	-.163
x5	.956	.062	.130
x7	.933	-.207	-.143
x1	.878	-.325	.143
x2	.854	.254	.265
x3	.830	-.210	.330
x4	.789	-.203	-.403
x11	.243	.927	-.052
x10	.207	.898	-.131
x12	.241	.698	-.359
x9	.059	.465	.727

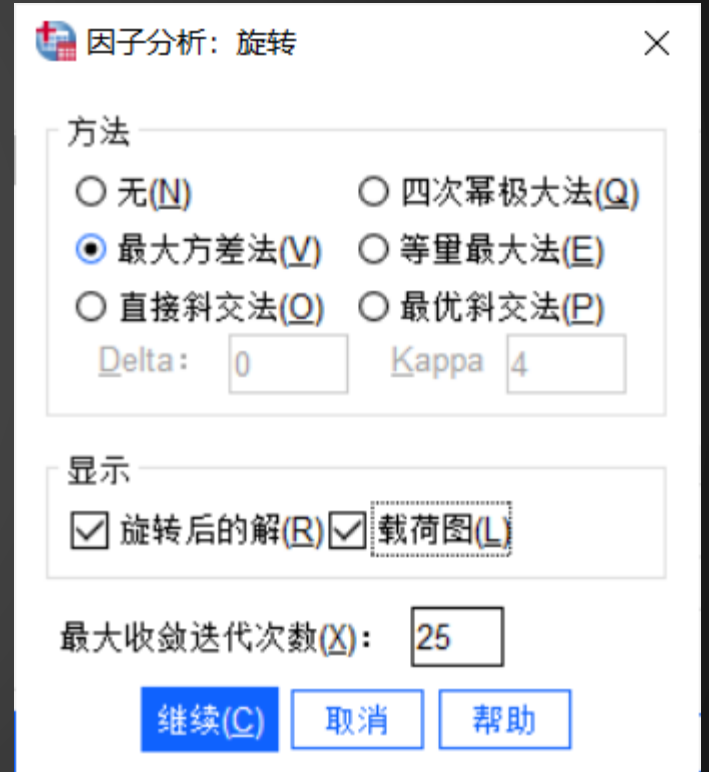
旋转后的成分矩阵 <sup>a</sup>			
	成分		
	1	2	3
x6	.970	.174	-.053
x8	.952	.199	-.155
x7	.947	.030	-.191
x5	.934	.194	.155
x1	.929	-.183	.039
x3	.870	-.147	.253
x2	.806	.309	.344
x4	.791	.091	-.437
x11	.068	.921	.259
x10	.034	.914	.175
x12	.092	.809	-.106
x9	.010	.205	.840

# 因子旋转

因子旋转分为正交旋转和斜交旋转。

正交旋转：对因子载荷矩阵，不断右乘正交矩阵，使得因子载荷方差达到最大，所以也叫最大方差法。

斜交旋转：指斜交因子模型对应的因子旋转法，SPSS中选择直接斜交法。



因子分析：旋转

方法

☐ 无(N) ☐ 四次幂极大法(Q)

☒ 最大方差法(V) ☐ 等量最大法(E)

☐ 直接斜交法(O) ☐ 最优斜交法(P)

Delta: 0 Kappa 4

显示

☒ 旋转后的解(R) ☒ 载荷图(L)

最大收敛迭代次数(X): 25

继续(C) 取消 帮助

# 因子得分

- 因子得分，就是公共因子 $F_1, F_2, \dots, F_m$ 在每一个样本点上的得分。
- 利用因子得分，可以看出样本在哪个因子上的得分较高，即样本属于什么类别。
- 计算方法：用回归的思想求出线性组合系数的估计值。

因子分析：因子得分

☒ 保存为变量(S)

方法

☒ 回归(R)

☐ 巴特利特(B)

☐ 安德森-鲁宾(A)

☒ 显示因子得分系数矩阵(D)

继续(C) 取消 帮助

FAC1_1	FAC2_1	FAC3_1
-.11720	5.19494	1.26656
1.12881	1.25556	-.57746
-.00438	.85425	-.61598
3.37437	.48928	-3.04015
-.72484	.35088	-.00940
-.66483	.31204	-.96159
-.89265	.20699	-.91478
-.61329	.01060	.99804
-.63835	-.02452	.17277
.16030	-.03748	-.15274
-.57159	-.11420	.73609
-.04028	-.12416	-.88330
-.43922	-.13093	.29792
-.37920	-.13531	.32644
-.79228	-.19365	.25895
-.61737	-.21760	-.11294
.15119	-.22219	-1.58707
-.38996	-.23111	-.77692

# 本章主要内容

1. 因子分析基本原理
2. 因子分析数学模型
3. 因子载荷的求解、因子旋转、因子得分
4. 因子分析案例：全国35个中心城市综合发展水平分析（SPSS）

# 因子分析案例：全国35个中心城市综合发展水平分析

选取反映城市综合发展水平的12个指标作为原始变量，其中包括8个社会经济指标和4个城市公共设施水平的指标。

- 8个社会经济指标：

非农业人口数（万人）、工业总产值（万元）、货运总量（万吨）、批发零售住宿餐饮业从业人数（万人）、地方政府预算内收入（万元）、城乡居民年底储蓄余额（万元）、在岗职工人数（万人）、在岗职工工资总额（万元）。

- 4个城市公共设施水平的指标：

人均居住面积（平方米）、每万人拥有公共汽车数（辆）、人均拥有铺装道路面积（平方米）、人均公共绿地面积（平方米）。

运用SPSS软件，对全国35个中心城市的综合发展水平进行因子分析。

- x1：非农业人口数（万人）
- x2：工业总产值（万元）
- x3：货运总量（万吨）
- x4：批发零售住宿餐饮业从业人数（万人）
- x5：地方政府预算内收入（万元）
- x6：城乡居民年底储蓄余额（万元）
- x7：在岗职工人数（万人）
- x8：在岗职工工资总额（万元）
- x9：人均居住面积（平方米）
- x10：每万人拥有公共汽车数（辆）
- x11：人均拥有铺装道路面积（平方米）
- x12：人均公共绿地面积（平方米）

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	地区	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
2	深圳	122.39	52451037	6792.66	10.84	2908370	21994500	104.98	3259900	21	114.91	47.29	177.62
3	广州	493.32	40178324	28859.45	21.47	2747707	37273276	182.16	5247087	17	11.16	12.76	178.76
4	南京	391.67	25093816	14804.68	7.62	1364788	11336202	87.91	1950742	16	9.06	12.13	136.72
5	北京	830.8	38103630	30671.14	127.4	5925388	64413910	434.15	10989365	15	17.3	8.56	44.94
6	合肥	160.18	5348605	4640.84	3.39	358694	3592488	37.88	526577	17	14.11	15.72	28.74
7	贵阳	165.27	3569419	5317.55	5.75	403855	3449487	54.53	664234	16	9.37	3.11	105.35
8	银川	79.2	1464867	2127.17	1.65	122605	1930771	29.12	393035	15	9.26	10.43	40.21
9	厦门	83.74	13201500	3054.82	2.83	701456	3971559	54.78	1042111	20	15.5	8.15	26.44
10	南宁	167.99	2083763	5893.09	4.95	362435	4514961	50.79	668976	18	9.91	9.32	35.12
11	青岛	329.96	25588695	30552.6	6.72	1201398	9084693	104.55	1603305	15	14.78	11.41	35.78
12	乌鲁木齐	142.94	3110943	12754.02	3.94	409119	4203000	47.42	782873	19	22.89	6.49	20.53
13	大连	297.48	15468641	21081.47	6.6	1105405	13101986	82.13	1442215	14	13.79	6.24	40.21
14	长沙	205.83	5339304	10630.5	6.31	598930	7048500	60.04	1019924	18	10.09	9.1	29.1
15	福州	205.43	12889573	8250.39	4.69	674522	8762245	71.3	1073262	18	9.65	7.9	31.6
16	呼和浩特	97.81	2407794	4155.1	2	205779	2554496	28.9	407963	18	3.81	8.92	26.58
17	南昌	195.46	4149169	4454.45	3.62	314094	4828029	49.79	692717	17	7.37	7.67	23.98
18	哈尔滨	454.52	7215089	9517.8	24.99	763600	11536951	168.83	2102165	14	12.75	6.34	18.51
19	昆明	205.34	5809573	12337.86	7.07	601101	7085278	73.34	1045469	15	15.33	4.49	23.33
20	长春	313.05	15115270	10891.98	6.94	459709	8313564	89.7	1244167	15	11.87	7.03	18.75
21	宁波	168.81	26302862	13797.38	4.8	1394162	10596339	59.88	1418635	17	9.88	6.81	17.65
22	兰州	175.54	5215490	5580.8	3.7	205660	4683830	54.91	740661	15	10.33	6.3	11.22
23	济南	297.21	13185425	14354.4	6.6	761054	7583525	78.38	1256160	19	7.77	10.62	19.54
24	郑州	249.72	9270494	7846.91	8.77	658737	10484859	83.99	1137056	19	10.11	7.63	17.77
25	西宁	105.13	1148959	2037.15	1.24	84397	1749293	20.6	301364	17	11.47	4.92	14.2
26	杭州	263.67	32025226	16815.2	8.36	1503888	14664200	75.72	1867776	17	8.93	6.5	23.19
27	沈阳	440.6	10643612	14635.74	7.3	810889	14229575	101.7	1521548	15	9.32	6.7	28.36
28	石家庄	331.33	11981505	10008.48	8.07	493429	10444919	86.74	1067432	18	7.23	8.28	21.56
29	成都	386.23	9700976	28798.2	8.06	895752	14944197	124.03	1894496	17	8.95	10.17	25.59
30	海口	76.05	2025643	3304.4	2.72	122541	2843664	22.97	340392	20	5.09	7.07	15.79
31	太原	222.63	5183200	15248.11	2.43	333473	6601300	74.55	945212	16	5.06	7.88	20.58
32	上海	1041.39	103000000	63861	35.22	8992850	60546000	281.51	7686511	19	14.57	12.92	19.11
33	西安	312.88	6386627	9392	12.21	648037	12105607	113.73	1535896	15	7.32	4.48	8.82
34	天津	549.74	40496103	34679	15.38	2045295	18253200	174.5	3254148	18	7.99	7.23	17.45
35	武汉	474.98	13344938	16610.34	13.58	804368	12855341	136.08	1868350	17	6.87	4.16	8.34
36	重庆	753.92	15889928	32450.2	12.83	1615618	18965569	203.79	2535070	21	4.94	4.24	10.8



# SPSS因子分析

主要操作步骤:

1. 读取数据
2. SPSS菜单: 【分析】 - 【降维】 - 【因子分析】。
3. 因子分析选项: 变量选择、相关性检验、因子抽取方法、因子旋转、因子得分等
4. 解释分析结果

# 因子分析与主成分分析的异同

- 相同点:

1. 主成分分析和因子分析的核心思想都是降维，即用较少的新变量代替原来较多的旧变量。

- 不同点:

1. 主成分分析中的主成分个数与原始变量个数是一样的，即有几个变量就有几个主成分，只不过最后我们确定了少数几个主成分而已。
2. 而因子分析则需要事先确定要找几个成分，也称为因子(factor)，然后将原始变量综合为少数的几个因子，以再现原始变量与因子之间的关系，一般来说，因子的个数会远远少于原始变量的个数。
3. 因子分析可以看作是主成分分析的推广和扩展，而主成分分析则可以看作是因子分析的一个特例。