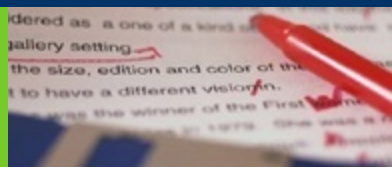
主成分分析法

主讲人：泰山教育　　小石老师

主成分分析法简介



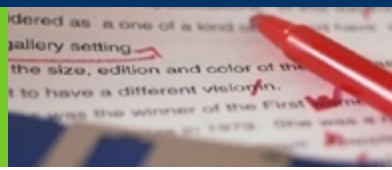
视频参考文献：　　张文霖
主成分分析在SPSS 中的操作应用

主成分分析（Principal Component Analysis，PCA），将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。

在实际课题中，为了全面分析问题，往往提出很多与此有关的变量（或因素），因为每个变量都在不同程度上反映这个课题的某些信息。

主成分：由原始指标综合形成的几个新指标。依据主成分所含信息量的大小成为第一主成分，第二主成分等等。

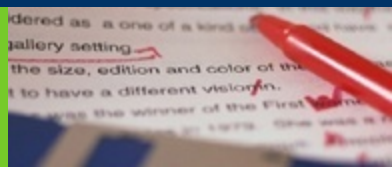
主成分分析法简介



一项十分著名的工作是美国的统计学家斯通(stone)在1947年关于国民经济的研究。他曾利用美国1929一1938年各年的数据，得到了17个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。

在进行主成分分析后，竟以97.4%的精度，用三新变量就取代了原17个变量。根据经济学知识，斯通给这三个新变量分别命名为总收入F1、总收入变化率F2和经济发展或衰退的趋势F3。

主成分分析法简介



主成分与原始变量之间的关系：

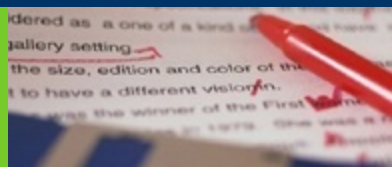
（1）主成分保留了原始变量绝大多数信息。

（2）主成分的个数大大少于原始变量的数目。

（3）各个主成分之间互不相关。

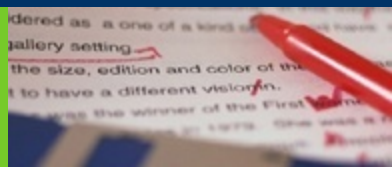
（4）每个主成分都是原始变量的线性组合。

主成分分析法简介



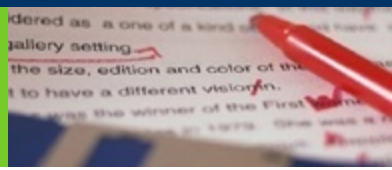
- 假设我们所讨论的实际问题中，有p个指标，我们把这p个指标看作p个随机变量，记为X_1，X_2，…，X_p，主成分分析就是要把这p个指标的问题，转变为讨论p个指标的线性组合的问题，而这些新的指标F_1，F_2，…，F_k(k≤p），按照保留主要信息量的原则充分反映原指标的信息，并且相互独立。
- 这种由讨论多个指标降为少数几个综合指标的过程在数学上就叫做降维。主成分分析通常的做法是，寻求原指标的线性组合F_i。

操作实例



地区	GDP	人均GDP	农业增加值	工业增加值	第三产业增加值	固定资产投资	基本建设投资	社会消费品零售总额	海关出口总额	地方财政收入
辽宁	5458.2	13000	14883.3	1376.2	2258.4	1315.9	529	2258.4	123.7	399.7
山东	10550	11643	1390	3502.5	3851	2288.7	1070.7	3181.9	211.1	610.2
河北	6076.6	9047	950.2	1406.7	2092.6	1161.6	597.1	1968.3	45.9	302.3
天津	2022.6	22068	83.9	822.8	960	703.7	361.9	941.4	115.7	171.8
江苏	10636	14397	1122.6	3536.3	3967.2	2320	1141.3	3215.8	384.7	643.7
上海	5408.8	40627	86.2	2196.2	2755.8	1970.2	779.3	2035.2	320.5	709
浙江	7670	16570	680	2356.5	3065	2296.6	1180.6	2877.5	294.2	566.9
福建	4682	13510	663	1047.1	1859	964.5	397.9	1663.3	173.7	272.9
广东	11770	15030	1023.9	4224.6	4793.6	3022.9	1275.5	5013.6	1843.7	1202
广西	2437.2	5062	591.4	367	995.7	542.2	352.7	1025.5	15.1	186.7

求指标对应的系数



解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	7.220	72.205	72.205	7.220	72.205	72.205
2	1.235	12.346	84.551	1.235	12.346	84.551
3	.877	8.769	93.319			
4	.547	5.466	98.786			
5	.085	.854	99.640			
6	.021	.211	99.850			
7	.012	.119	99.970			
8	.002	.018	99.988			
9	.001	.012	100.000			
10	-1.534E-16	-1.534E-15	100.000			

提取方法：主成份分析。

成份矩阵[a]

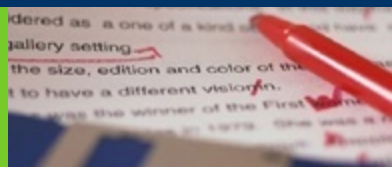
	成份	
	1	2
GDP	.949	.195
人均GDP	.112	-.824
农业增加值	-.109	.677
工业增加值	.978	-.005
第三产业增加值	.986	.070
固定资产投资	.983	-.068
基本建设投资	.947	-.024
社会消费品零售总额	.977	.176
海关出口总额	.800	-.051
地方财政收入	.954	-.128

提取方法 ：主成分分析法。

a．已提取了 2 个成份。

成分矩阵中的数据除以主成分相对应的特征值开平方根便得到两个主成分中每个指标所对应的系数。

求指标对应的系数

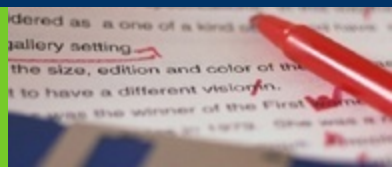


$$F_1 = 0.353ZX_1 + 0.042ZX_2 - 0.041ZX_3 + 0.364ZX_4 + 0.367ZX_5 + 0.366ZX_6 + 0.352ZX_7 + 0.364ZX_8 + 0.298ZX_9 + 0.355ZX_{10}$$

$$F_2 = 0.175ZX_1 - 0.741ZX_2 + 0.609ZX_3 - 0.004ZX_4 + 0.063ZX_5 - 0.061ZX_6 - 0.022ZX_7 + 0.158ZX_8 - 0.046ZX_9 - 0.115ZX_{10}$$

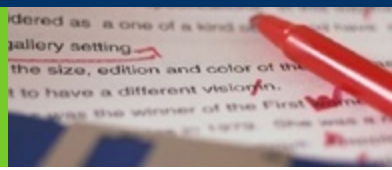
$$F = (72.2/84.5)\ F_1 + (12.3/84.5)\ F_2$$

归一化后的原始数据



	zGDP	z人均GDP	z农业增加值	z工业增加值	z第三产业增加值	z固定资产投资	z基本建设投资	z社会消费品零售总额	z海关出口总额	z地方财政收入
辽宁	-0.35567	-0.31866	2.83364	-0.54141	-0.31477	-0.41279	-0.65088	-0.1317	-0.42652	-0.34163
山东	1.13739	-0.45836	-0.16853	1.08599	0.93403	0.75886	0.82066	0.62994	-0.26383	0.33158
河北	-0.17434	-0.72561	-0.26638	-0.51807	-0.44478	-0.59863	-0.46588	-0.37096	-0.57134	-0.65313
天津	-1.36309	0.61486	-0.45913	-0.96497	-1.33289	-1.15012	-1.10481	-1.21788	-0.44141	-1.07049
江苏	1.16261	-0.17484	-0.22802	1.11186	1.02515	0.79656	1.01245	0.6579	0.05932	0.43872
上海	-0.37015	2.52544	-0.45861	0.08619	0.07525	0.37526	0.02907	-0.31578	-0.06018	0.64756
浙江	0.29289	0.04886	-0.3265	0.20888	0.31771	0.76838	1.11921	0.37889	-0.10914	0.1931
福建	-0.58327	-0.26616	-0.33028	-0.7933	-0.62796	-0.83602	-1.00702	-0.6225	-0.33344	-0.74715
广东	1.49513	-0.10968	-0.24998	1.63866	1.67316	1.64314	1.37701	2.14062	2.7752	2.22425
广西	-1.24151	-1.13585	-0.34621	-1.31382	-1.3049	-1.34464	-1.1298	-1.14852	-0.62867	-1.02283

求指标对应的系数



$$F_1 = 0.131ZX_1 + 0.015ZX_2 - 0.015ZX_3 + 0.135ZX_4 + 0.137ZX_5 + 0.136ZX_6 + 0.131ZX_7 + 0.135ZX_8 + 0.111ZX_9 + 0.132ZX_{10}$$

$$F_2 = 0.158ZX_1 - 0.667ZX_2 + 0.548ZX_3 - 0.004ZX_4 + 0.056ZX_5 - 0.055ZX_6 - 0.020ZX_7 + 0.142ZX_8 - 0.041ZX_9 - 0.104ZX_{10}$$

$$F = (72.2/84.5)\ F_1 + (12.3/84.5)\ F_2$$

方法一结果

城市	F1	F2	F	排名
广东	5.224739	0.114592	4.478657	1
江苏	2.254315	0.234636	1.959442	2
山东	1.962522	0.500242	1.749029	3
浙江	1.160716	-0.19308	0.963062	4
上海	0.296827	-2.35794	-0.09077	5
辽宁	-1.24298	1.960091	-0.77534	6
河北	-1.35286	0.408853	-1.09565	7
福建	-1.97451	-0.06651	-1.69594	8
天津	-3.04194	-1.00948	-2.7452	9
广西	-3.28683	0.408604	-2.7473	10

方法二结果

城市	F1	F2	F	排名
广东	1.942896	0.10049	1.673905	**1**
江苏	0.837921	0.20962	0.746189	**2**
山东	0.729458	0.448675	0.688464	**3**
浙江	0.431365	-0.17473	0.342875	**4**
辽宁	-0.46171	1.76376	-0.13679	**5**
上海	0.109227	-2.12134	-0.21644	**6**
河北	-0.50276	0.368541	-0.37555	**7**
福建	-0.73386	-0.05851	-0.63526	**8**
广西	-1.22134	0.369522	-0.98907	**9**
天津	-1.13119	-0.90602	-1.09832	**10**

Thank You !