2．在 matlab 中提供了直接计算主成分的命令：

（1）.princomp

功能：主成分分析

格式：PC=princomp(X)

[PC,SCORE,latent,tsquare]=princomp(X)

说明：[PC,SCORE,latent,tsquare]=princomp(X)对数据矩阵 X 进行主成分分析，给出各主成分(PC)、所谓的 Z-得分(SCORE)、X 的方差矩阵的特征值(latent)和每个数据点的 HotellingT2 统计量(tsquare)。

（2）.pcacov

功能：运用协方差矩阵进行主成分分析

格式：PC=pcacov(X)

[PC,latent,explained]=pcacov(X)

说明：[PC,latent,explained]=pcacov(X)通过协方差矩阵 X 进行主成分分析，返回主成分(PC)、协方差矩阵 X 的特征值(latent)和每个特征向量表征在观测量总方差中所占的百分数(explained)。

（3）.pcares

功能：主成分分析的残差

格式：residuals=pcares(X,ndim)

说明：pcares(X,ndim)返回保留 X 的 ndim 个主成分所获的残差。注意，ndim 是一个标量，必须小于 X 的列数。而且，X 是数据矩阵，而不是协方差矩阵。

主成分分析方法（举例）(2008-04-26 21:41:50)

标签：杂谈　　分类：归纳整理

3. 主成分分析方法应用实例

1）实例 1：流域系统的主成分分析（张超，1984）

表 3.5.1（点击显示该表）给出了某流域系统 57 个流域盆地的 9 项变量指标。其中，x1 代表流域盆地总高度（m），x2 代表流域盆地山口的海拔高度（m），x3 代表流域盆地周长（m），x4 代表河道总长度（m），x5 代表河道总数，x6 代表平均分叉率，x7 代表河谷最大坡度（度），x8 代表河源数， x9 代表流域盆地面积（km2）。

表 3.5.1　某 57 个流域盆地地理要素数据

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	760	5490	1.704	2.481	30	2.785	31.8	20	0.143
2	1891	4450	2.765	4.394	30	5.833	37.0	26	0.312
3	325	5525	1.500	2.660	36	3.042	21.1	25	0.162
…	…	…	…	…	…	…	…	…	…
35	847	7188	1.591	1.610	14	3.17	31.3	10	0.094

注：表中数据详见书本 87 和 88 页。

(1) 分析过程：

① 将表 3.5.1 中的原始数据作标准化处理，然后将它们代入相关系数公式计算，得到相关系数矩阵（表 3.5.2）。

表 3.5.2　相关系数矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	1.000								
x_2	-0.370	1.000							
x_3	0.619	-0.017	1.000						
x_4	0.657	-0.157	0.841	1.000					
x_5	0.474	-0.150	0.737	0.921	1.000				
x_6	0.074	-0.274	0.167	0.094	0.165	1.000			
x_7	0.607	-0.566	0.162	0.217	0.158	0.170	1.000		
x_8	0.481	-0.158	0.753	0.928	0.999	0.181	0.164	1.000	
x_9	0.689	-0.016	0.910	0.937	0.788	0.071	0.158	0.799	1.000

② 由相关系数矩阵计算特征值，以及各个主成分的贡献率与累计贡献率（见表 3.5.3）。由表 3.5.3 可知，第一，第二，第三主成分的累计贡献率已高达 86.5%，故只需求出第一、第二、第三主成分 z1，z2，z3 即可。

表 3.5.3 特征值及主成分贡献率

主成分	特征值	贡献率（%）	累计贡献率（%）
1	5.043	56.029	56.029
2	1.746	19.399	75.428
3	0.997	11.076	86.504
4	0.610	6.781	93.285
5	0.339	3.778	97.061
6	0.172	1.907	98.967
7	0.079	0.8727	99.840
8	0.014	0.1556	99.996
9	0.0004	0.0042	100.00

③ 对于特征值 λ_1=5.043，λ_2=1.746，λ_3=0.997 分别求出其特征向量 e_1，e_2，e_3，再用公式 $l_{ij}=p(z_i,x_j)=\sqrt{\lambda_i}e_{ij}(i,j=1,2,\Lambda\ ,p)$ 计算各变量 x_1，x_2，…，x_9 在主成分 z_1，z_2，z3 上的载荷

（表 3.5.4）。

表 3.5.4 主成分载荷

原变量	主成分			占方差的百分数（%）
	Z_1	Z_2	Z_3	
x_1	0.75	-0.38	-0.36	83.05
x_2	-0.25	0.82	-0.08	73.20
x_3	0.89	0.19	0.00	82.19
x_4	0.97	0.14	-0.03	96.63
x_5	0.91	0.18	0.16	88.26
x_6	0.20	-0.36	0.86	89.97
x_7	0.35	-0.80	-0.25	83.19
x_8	0.92	0.17	0.16	89.90
x_9	0.93	0.22	-0.10	92.16

(2) 结果分析：

▲ 第一主成分 z1 与 x1，x3，x4，x5，x8，x9 有较大的正相关，可以看作是流域盆地规模

的代表；

▲ 第二主成分 z2 与 x2 有较大的正相关，与 x7 有较大的负相关，分可以看作是流域侵蚀状况的代表；

▲ 第三主成分 z3 与 x6 有较大的正相关，可以看作是河系形态的代表；

▲ 根据主成分载荷，该流域系统的 9 项要素可以被归纳为三类，即流域盆地的规模，流域侵蚀状况和流域河系形态。如果选取其中相关系数绝对值最大者作为代表，则流域面积、流域盆地出口的海拔高度和分叉率可作为这三类要素的代表。

主成分分析法

主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标。在实证问题研究中，为了全面、系统地分析问题，我们必须考虑众多影响因素。这些涉及的因素一般称为指标，在多元统计分析中也称为变量。因为每个变量都在不同程度上反映了所研究问题的某些信息，并且指标之间彼此有一定的相关性，因而所得的统计数据反映的信息在一定程度上有重叠。在用统计方法研究多变量问题时，变量太多会增加计算量和增加分析问题的复杂性，人们希望在进行定量分析的过程中，涉及的变量较少，得到的信息量较多。主成分分析正是适应这一要求产生的，是解决这类题的理想工具。

主成分分析法是一种数学变换的方法，它把给定的一组相关变量通过线性变换转成另一组不相关的变量，这些新的变量按照方差依次递减的顺序排列。在数学变换中保持变量的总方差不变，使第一变量具有最大的方差，称为第一主成分，第二变量的方差次大，并且和第一变量不相关，称为第二主成分。依次类推，I 个变量就有 I 个主成分。

1．主成分分析的基本原理

主成分分析：把原来多个变量划为少数几个综合指标的一种统计分析方法，是一种降维处理技术。）

记原来的变量指标为 x1，x2，…，xP，它们的综合指标——新变量指标为 z1，z2，…，zm（m≤p），则

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \Lambda + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \Lambda + l_{2p}x_p \\ \qquad\qquad \mathrm{M} \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \Lambda + l_{mp}x_p \end{cases}$$

z1，z2，…，zm 分别称为原变量指标 x1，x2，…，xP 的第一，第二，…，第 m 主成分，在实际问题的分析中，常挑选前几个最大的主成分。

① z_i与 z_j（i≠j；i，j=1，2，…，m）相互无关；

② z_1是 x_1，x_2，…，x_p的一切线性组合中方差最大者，z_2是与 z_1不相关的 x_1，x_2，…，x_p的所有线性组合中方差最大者；……；z_m是与 z_1，z_2，……，z_{m-1}都不相关的 x_1，x_2，…，x_p的所有线性组合中方差最大者。

■ 系数 lij 的确定原则（单击展开显示）

① zi 与 zj（i≠j；i，j=1，2，…，m）相互无关；

② z1 是 x1，x2，…，xP 的一切线性组合中方差最大者，z2 是与 z1 不相关的 x1，x2，…，xP 的所有线性组合中方差最大者；……；zm 是与 z1，z2，……，zm－1 都不相关的 x1，x2，…，xP 的所有线性组合中方差最大者。

■ 主成分分析的数学特征（单击展开显示）

找主成分就是确定原来变量 x_j（j=1，2，…，p）在诸主成分 z_i（i=1，2，…，m）上的**载荷** l_{ij}（i=1，2，…，m；j=1，2，…，p）。它们分别是 x_1，x_2，…，x_p的相关矩阵的 m 个较大的特征值所对应的特征向量。

2. 主成分分析的计算步骤

① 计算相关系数矩阵

$$R=\begin{bmatrix} r_{11} & r_{12} & \Lambda & r_{1p} \\ r_{21} & r_{22} & \Lambda & r_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ r_{p1} & r_{p2} & \Lambda & r_{pp} \end{bmatrix}$$

其中 $$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$$

② 计算特征值与特征向量

Ⅰ 解特征方程$|\lambda I-R|=0$，通常用雅可比法（Jacobi）求出特征值$\lambda_i\,(i=1,2,\Lambda\ ,p)$，并使其按大小顺序排列，即$\lambda_1\geq\lambda_2\geq\Lambda\ ,\geq\lambda_p\geq 0$；

Ⅱ 分别求出对应于特征值λ_i的特征向量$e_i\,(i=1,2,\Lambda\ ,p)$。这里要求$\|e_i\|$=1，即$\sum_{j=1}^{p}e_{ij}^2=1$，其中e_{ij}表示向量e_i的第 j 个分量。

③ 计算主成分贡献率及累计贡献率

主成分z_i的贡献率为

$$\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}\qquad(i=1,2,\Lambda\ ,p)$$

累计贡献率为

$$\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}\qquad(i=1,2,\Lambda\ ,p)$$

一般取累计贡献率达 85—95%的特征值$\lambda_1,\lambda_2,\Lambda\ ,\lambda_m$所对应的第一、第二，…，第 m（m≤p）个主成分。

④ 计算主成分载荷

$$l_{ij} = p(z_i, x_j) = \sqrt{\lambda_i} e_{ij} (i, j = 1,2,\Lambda\ ,p)$$

计算各主成分的得分

$$Z = \begin{bmatrix} z_{11} & z_{12} & \Lambda & z_{1m} \\ z_{21} & z_{22} & \Lambda & z_{2m} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ z_{n1} & z_{n2} & \Lambda & z_{nm} \end{bmatrix}$$

1 主成分分析法的数学原理

设有个原始指标：，用来评价个单位，则共有个数据。这个原始指标之间往往存在着一定的相关性，主成分分析的目的是要将这些原始指标组合成新的不相关的指标：，以使各指标在整个经济过程中的作用容易解释，这些综合指标表现为原始指标的线性函数：

由于所组合成的新指标：彼此不相关，就使我们有可能从中选择主要成分，通过对主要成分的重点分析，达到综合评价的目的。

通过数学计算可以将个原始指标：的总方差分解为新的不相关的指标：的方差之和，并使第一个综合指标方差达到最大（贡献率最大），第二个综合指标方差达到次大，依此类推，一般前面几个综合指标即可包含总方差中的绝大部分，也就是说，主成分分析可以使原始指标的大部分方差“集中”于少数几个主成分综合指标上，通过对这几个主成分的分析来实现对总体的综合评价。

2 主成分分析法的计算步骤

主成分分析可分为五个主要步骤：

第一步，列出原始指标数值矩阵；

第二步，计算的相关矩阵；

第三步，计算相关矩阵的特征值和特征向量（即指标的系数）；

第四步，计算贡献率和累计贡献率，据以确定主成分的个数，并建立主成分方程；每个主成分的贡献率等于它的特征值除以原始指标个数，累计贡献率等于各主成分的贡献率顺序相加，根据一定的选择标准，如果前个主成分的累计贡献率大于或等于，则可选定这个主成分，根据特征向量建立这个主成分的线性方程：

第五步，解释各主成分的意义，并将各单位的原始指标数值代入方程中计算综合评价值进行分析比较（在多指标综合评价中，一般只需取第一个主成分作为全面反映各指标状况的综合指标，因为它综合原始指标信息的能力最强）。

3 主成分分析法的应用实例

[实例] 南通隆盛机电集团有限公司生产的一种新产品有 20 种型号，现通过 4 个技术指标进行综合评价，原始指标数值矩阵为：

应用 SPSS 统计分析软件可得：

的相关矩阵为：，

相关矩阵的特征值、贡献率、累计贡献率和特征向量为：

表1　　主成分、特征值、贡献率、累计贡献率和特征向量表

y	λ	贡献率	累计贡献率	l_1	l_2	l_3	l_4
y_1	2.920	0.730	0.730	0.1485	-0.5735	-0.5577	-0.5814
y_2	1.024	0.256	0.986	0.9544	-0.0984	0.2695	0.0824
y_3	0.049	0.012	0.998	0.2516	0.7733	-0.5589	-0.1624
y_4	0.007	0.002	1.000	-0.0612	0.2519	0.5513	-0.0793

从表 1 可见，前 2 个主成分的累计贡献率为 98.6%，如果舍弃其余的主成分，丢失的信息仅为 1.4%。若只选择第一主成分，其贡献率为 73%，也已经包含了原始指标数值的绝大部分信息，具有一定的代表性。该方程式为：

…….（1）

通过对原始指标数值标准化处理后代入方程式(1)得到各种型号产品的综合评价值为：

表2　　20种型号产品的综合评价值及排序表

产品型号	01	02	03	04	05	06	07	08	09	10
综合评价值	-2.01	-2.10	-3.03	-1.35	-1.76	-1.98	0.31	0.75	-1.40	-1.10
排　　序	**18**	**19**	**20**	**14**	**16**	**17**	**8**	**6**	**15**	**13**
产品型号	11	12	13	14	15	16	17	18	19	20
综合评价值	1.59	0.05	-0.16	0.89	0.23	3.81	1.78	0.99	0.27	0.32
排　　序	**3**	**11**	**12**	**5**	**10**	**1**	**2**	**4**	**9**	**7**