中山大學
SUN YAT-SEN UNIVERSITY

实验设计与数据分析

第十一讲：主成分分析

肖叶，副教授
环境科学与工程学院
中山大学

广州，2023

主成分分析

- 主成分分析的原理
- 主成分的性质
- 主成分分析的步骤
- 主成分分析的应用实例

主成分分析

- 学科定位：
  - 一种多元统计分析方法
- 又写作：
  - 主成份分析
  - Principal Component Analysis， PCA

主成分分析

- 处理对象：
  - p个变量的n组样本数据
- 处理实质：
  - 多个变量→少数几个综合变量

	1	2	…	n
X_1	X_{11}	X_{12}	$\cdots$	X_{1n}
X_2	X_{21}	X_{22}	$\cdots$	X_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
X_p	X_{p1}	X_{p2}	$\cdots$	X_{pn}

主成分分析的原理

- 如果p个变量是相互独立的：
  - 可以化成单变量问题逐一进行处理
- 现实中，各变量之间：
  - 总是存在不同程度的相关关系
- 研究策略：
  - 采取“降维”的方法
- 研究思路：
  - 利用全部p个变量构造新的q个综合变量（$q \leq p$），例如采用线性组合的方式；
  - 新的q个变量尽可能多地反映原p个变量的统计特性；
  - 新的q个变量是相互独立的

主成分分析的原理

- 举例说明1
  - **数据**
  - 4个变量：身高X_1（cm）、体重X_2（kg）、胸围X_3（cm）、坐高X_4（cm）
  - 30个样本： 30名同学的相关数据
  - **主成分分析结果（1）**
  - 降维与各主成分的贡献率：
  - 确定第一和第二两个主成分就可以描述原有四个变量
  - 第一主成分贡献率88.5%，第二主成分贡献率7.8%

主成分分析的原理

- 举例说明1
  - **主成分分析结果（2）**
  - 主成分与原始变量的关系：
  - 第一主成分：

$$+0.497X_1' + 0.515X_2' + 0.481X_3' + 0.507X_4'$$

  - 第二主成分：

$$-0.543X_1' + 0.210X_2' + 0.724X_3' - 0.368X_4'$$

  - 以上表达式中用的是原始变量经过标准化后的变量

主成分分析的原理

举例说明1

- **主成分分析结果（3）**
- 第一主成分的解释（ $+0.497X_1' + 0.515X_2' + 0.481X_3' + 0.507X_4'$ ）：
- 数学形式上的特点：
  - 4个分量均为正值，均在0.5左右
- 反映如下事实：
  - 一个身材魁梧的学生，他的4个变量取值都大，第一主成分的得分就高；
  - 反之，一个身材瘦小的学生，他的4个变量取值都小，第一主成分的得分就低
- 可以定义为：
  - 大小因子（ size factor）

主成分分析的原理

- 举例说明1
  - **主成分分析结果（4）**
  - 第二主成分的解释（$-0.543X_1' + 0.210X_2' + 0.724X_3' - 0.368X_4'$）：
  - 数学形式上的特点：
    - 身高、坐高对应的分量为负值，体重、胸围对应的分量为正值
  - 反映如下事实：
    - 学生身材的胖瘦情况
  - 可以定义为：
    - 形状因子（ shape factor）

主成分分析的原理

- 举例说明2
  - 一项十分著名的工作是美国统计学家斯通(stone)在1947年关于国民经济的研究。
  - 利用美国1929-1938年各年的数据，整理出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

主成分分析的原理

- 举例说明2
  - 在进行主成分分析后：
  - 用3个新变量（即3个主成分）取代了原17个变量；
  - 这种取代达到了97.4%的精度；
  - 根据经济学知识，斯通给3个新变量分别命名为总收入F1、总收入变化率F2和经济发展或衰退趋势F3；
  - 更有意思的是，这3个变量其实都可以直接测量

主成分分析的原理

举例说明2

- 斯通将3个主成分与实际测量的总收入I、总收入变化率ΔI以及时间t因素做相关分析

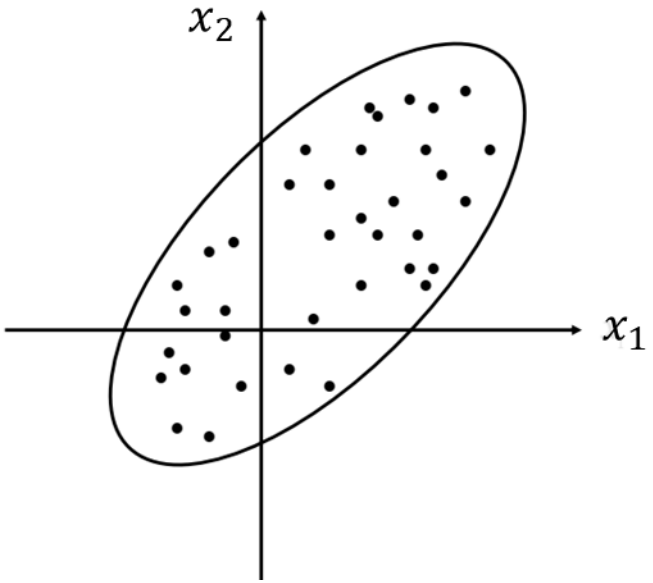
	F1	F2	F3	I	ΔI	t
F1	1					
F2	0	1				
F3	0	0	1			
I	0.995	-0.041	0.057	1		
ΔI	-0.056	0.948	-0.124	-0.102	1	
t	-0.369	-0.282	-0.836	-0.414	-0.112	1

主成分分析的原理

- 主成分分析的几何解释：
  - 为了方便，在二维空间中讨论主成分的几何意义。
  - 设有n个样本，每个样本有两个观测变量x_1和x_2，在由变量x_1和x_2所确定的二维平面中，n个样本点所散布的情况如椭圆状。
  - 有何特点？

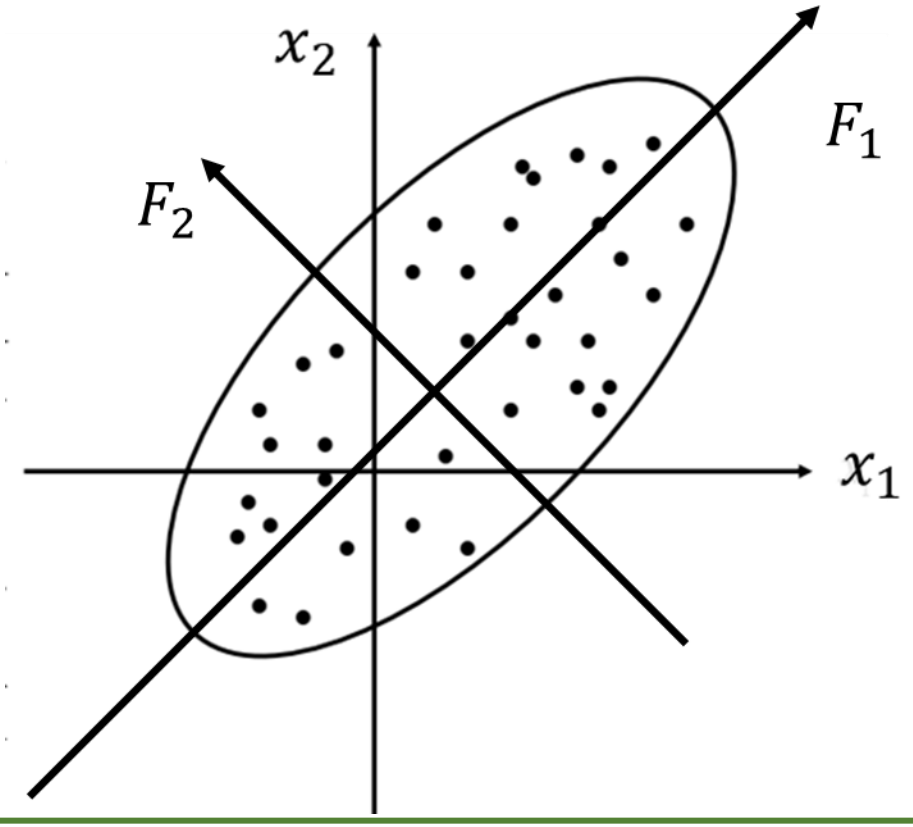
主成分分析的原理

- 主成分分析的几何解释：
  - n个样本点无论是沿着x_1轴方向或x_2轴方向都具有较大的离散性
  - 离散程度可以分别用x_1的方差和x_2的方差定量表示。
  - 显然，如果只考虑x_1和x_2 中的任何一个，那么包含在原始数据中的信息将会有较大的损失。



主成分分析的原理

- 主成分分析的几何解释：
  - 如果将x_1轴和x_2轴先平移，再同时按逆时针方向旋转某个角度，可得到新坐标轴F_1和F_2
  - F_1和F_2是两个新变量。



主成分分析的原理

主成分分析的几何意义：

- 坐标平移旋转变换的目的：
  - 使得n个样品点在F_1轴方向上的离散程度最大，即F_1的方差最大。
  - 变量F_1代表了原始数据的绝大部分信息，在研究某具体问题时，即使不考虑变量F_2也无损大局。
- 经过上述变换，原始数据的大部分信息集中到F_1轴上，对数据中包含的信息起到了浓缩作用。
- 二维平面上的各点的方差大部分都归结在 F_1 轴上，而 F_2 轴上的方差很小。
- F_1和F_2称为原始变量x_1和x_2的综合变量。 F简化了系统结构，抓住了主要矛盾。
- F_1， F_2除了可以对包含在x_1， x_2中的信息起着浓缩作用之外，还具有不相关的性质，这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。

主成分分析的原理

- 取一组参数$\alpha_1, \alpha_2, \cdots, \alpha_p$，考虑变量$X_1, X_2, \cdots, X_p$的线性组合：
  - 设综合变量：

$$Z = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p$$

  - 采用矩阵和向量的形式表示：

$$Z = \boldsymbol{\alpha}' \mathbf{X}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

主成分分析的原理

- 要使综合变量Z最大限度地反映原p个变量的信息，参数α 应使得Z 满足以下4个条件：
  - 综合变量 Z 的方差最大；
  - 如果把数据看作p维空间中的n个点，使这n个点到超平面Z的距离和最小；
  - 以Z为因变量，$X_1, X_2, \cdots, X_p$为自变量进行回归，其残差平方和最小；
  - 综合变量Z与变量$X_1, X_2, \cdots, X_p$的相关系数的平方和最小。

主成分分析的原理

- 理论上可以证明：
  - 前3个条件：
  - 变量$X_1, X_2, \cdots, X_p$的协方差矩阵的特征值问题；
  - 第4个条件：
  - 变量$X_1, X_2, \cdots, X_p$的相关矩阵的特征值问题

主成分分析的原理

- 设S为变量$X_1, X_2, \cdots, X_p$的协方差矩阵
- S的p个非负特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$:
  - 方差最大的综合变量Z_1:
  - 由最大的特征值λ_1对应的特征向量$\boldsymbol{\alpha}_1^*$为系数来构造，

$$Z_1 = \boldsymbol{\alpha}_1^{*\prime}\mathbf{X} = \alpha_{11}X_1 + \alpha_{21}X_2 + \cdots + \alpha_{p1}X_p$$

  - 综合变量Z_1被称为:

    第一主成分（first principal component），其方差为λ_1

  - 第一主成分:

    可以反映原来p个变量的大部分信息，但还并不是全部的

主成分分析的原理

- 设S为变量$X_1, X_2, \cdots, X_p$的协方差矩阵
- S的p个非负特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$:
  - 作为第一主成分Z_1的补充，还可以考虑第二主成分Z_2；
  - Z_1、Z_2是独立的;
  - 利用次最大特征值λ_2对应的特征向量$\boldsymbol{\alpha}_2^*$为系数可以构造第二主成分:

$$Z_2 = \boldsymbol{\alpha}_2^{*\prime}\mathbf{X} = \alpha_{12}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{p2}X_p$$

  - 第二主成分的方差为λ_2
  - 依此类推可以获得第三主成分，……，第p个主成分

主成分分析

- 主成分分析的原理
- 主成分的性质
- 主成分分析的步骤
- 主成分分析的应用实例

主成分的性质

各主成分之间的关系

- 主成分之间彼此独立
- 主成分的方差之间存在如下关系：

$$Var(Z_1) = \lambda_1 \geq Var(Z_2) = \lambda_2 \geq \cdots\cdots \geq Var(Z_p) = \lambda_p$$

- 方差大小客观描述了各个主成分的贡献（即所含原始信息的多少），因而$Z_1, Z_2, \cdots, Z_p$所包含的原始信息依次减少

主成分的性质

各个主成分的贡献

- 度量方式：贡献率与累积贡献率
- 贡献率：

$$p_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

  - p_k称为第k个主成分的贡献率
  - 它反映了第k个主成分提取全部原始信息的多少
- 累积贡献率：

$$G_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

  - G_k为前k个主成分的累积贡献率
  - 它反映了前k个主成分共同提取全部原始信息的多少

主成分的性质

- 累积贡献率与主成分个数的确定
  - 一般做法：
  - 如果前m个主成分的累积贡献率大于等于85%，则取m个主成分就能够反映全部p个变量的绝大部分信息了。

主成分分析

- 主成分分析的原理
- 主成分的性质
- 主成分分析的步骤
- 主成分分析的应用实例

主成分分析的步骤

- （1）对原始数据进行标准化处理
  - 标准化处理的目的：
  - 使得各变量的均值为0，方差为1
  - 这样得出的协方差矩阵与相关矩阵是一样的，满足第4个条件
  - 标准化处理的计算方法：
  - 设有p个变量，n个样本，观测值为X_{ij}（$i=1,2,\cdots,p$；$j=1,2,\cdots,n$），则标准化后的值为：

$$X'_{ij}=\frac{X_{ij}-\bar{X}_i}{S_i} \qquad \bar{X}_i=\frac{1}{n}\sum_{j=1}^{n}X_{ij} \qquad S_i=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_i\right)^2}$$

主成分分析的步骤

- （2）由标准化后的数据求得协方差矩阵
  - 相当于获得原始数据的相关矩阵
  - 计算方法：

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

$$r_{ij} = \frac{1}{n-1}\sum_{k=1}^{n} X'_{ik}X'_{jk}$$

$$(i = 1, 2, \cdots, p;\quad j = 1, 2, \cdots, p)$$

主成分分析的步骤

- （3）求协方差矩阵的特征值及其对应的特征向量
  - 获得：
  - 所有特征根

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

  - 特征根对应的特征向量

$$\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \cdots, \boldsymbol{\alpha_p}$$

主成分分析的步骤

- （4）确定主成分
  - 计算累积贡献率
  - 一般根据累积贡献率85%，确定要选取的主成分个数k
  - 根据k及每个 $\lambda_i(1 \leq i \leq p)$ 所对应的特征向量α_i，构造标准化变量的主成分
  - 转化为原始变量的主成分表达式

主成分分析的步骤

- (5) 计算主成分的分值
  - 将各样本所对应的变量值代入主成分表达式，计算各样本的主成分得分

- (6) 解释结论
  - 根据样本的主成分得分，结合专业知识进行综合研究，对结论进行解释与推断

主成分分析

- 主成分分析的原理
- 主成分的性质
- 主成分分析的步骤
- 主成分分析的应用实例

主成分分析的应用

- 主成分分析在环境数据处理中的作用：
  - 对环境系统中错综复杂的信息进行提取和精练，常用于环境质量分类和环境影响识别
  - 揭示变量之间的一些关系，而这些关系有时候是别的方法或已有的专业知识所难以预料的
- 举例说明
  - 河流监测数据分析
  - 黄山市生态区划

主成分分析的应用

- 例1：下表1为某河流9个断面的BOD_5、 COD、 DO的监测值，请根据此表进行主成分分析。

表1　断面水质监测数据

断面	1#	2#	3#	4#	5#	6#	7#	8#	9#
BOD_5（x_1）	1.59	3.03	6.15	18.99	3.32	3.65	2.75	3.59	5.99
COD（x_2）	2.46	4.62	8.55	18.97	6.11	6.89	3.46	4.60	8.79
DO（x_3）	7.86	7.68	7.06	4.32	6.65	7.76	7.60	7.33	7.06

主成分分析的应用

- 解：（1）计算原始数据的相关矩阵

表2　各变量的平均值与标准差

原始变量	平均值	标准差	标准化变量
BOD_5（x_1）	5.4511	5.2844	$x_1'=(x_1-5.4511)/5.2844$
COD（x_2）	7.1611	4.9248	$x_2'=(x_2-7.1611)/4.9248$
DO（x_3）	7.0356	1.0925	$x_3'=(x_3-7.0356)/1.0925$

- 相关矩阵**R**

$$\mathbf{R}=\begin{bmatrix}1 & 0.9776 & -0.9534\\ 0.9776 & 1 & -0.9326\\ -0.9534 & -0.9326 & 1\end{bmatrix}$$

主成分分析的应用

- 解：（2）计算特征值和特征向量

表3　特征值与特征向量

		特征向量			特征值（λ）	贡献率（%）（信息量）	累积贡献率（%）
		BOD_5	COD	DO			
主	Z_1	0.5817	0.5776	–0.5727	2.9093	96.97	96.97
成	Z_2	0.2248	0.5625	0.7957	0.0714	2.38	99.35
分	Z_3	-0.7817	0.5916	-0.1974	0.0194	0.65	100.00

主成分分析的应用

- 解：（3）确定主成分，并计算相应的分值
  - 标准化变量对应的主成分表达式

$$Z_1 = 0.5817x_1' + 0.5776x_2' - 0.5727x_3'$$
$$Z_2 = 0.2248x_1' + 0.5625x_2' + 0.7957x_3'$$
$$Z_3 = -0.7817x_1' + 0.5916x_2' - 0.1974x_3'$$

  - 原始变量对应的主成分表达式

$$Z_1 = 0.1097x_1 + 0.1177x_2 - 0.5239x_3 + 2.2451$$
$$Z_2 = 0.0424x_1 + 0.1141x_2 + 0.7283x_3 - 6.1727$$
$$Z_3 = -0.1482x_1 + 0.1198x_2 - 0.1792x_3 + 1.2134$$

表4　各断面的主成分得分

主成分	1#	2#	3#	4#	5#	6#	7#	8#	9#
Z_1	-1.409	-0.902	0.227	4.298	-0.155	-0.609	-1.028	-0.660	0.238
Z_2	-0.100	0.076	0.205	-0.057	-0.492	0.420	-0.126	-0.157	0.226
Z_3	-0.136	-0.058	0.061	-0.102	0.262	0.107	-0.142	-0.081	0.113

主成分分析的应用

- 解：（4）结论分析与解释
  - 主成分个数的确定：
    - 第一主成分的贡献率已达到96.97%，因此只考虑该主成分
  - 主成分物理意义的推断：
  - 第一主成分中各变量的系数绝对值大致相同，且BOD_5和COD系数为正，DO系数为负，所以认为第一主成分表示有机污染与水体自净能力的对比程度
  - 第一主成分的值越大，说明有机污染越严重；相反，该值越小，说明水体自净能力越强。

主成分分析的应用

- 解：（4）结论分析与解释
  - 第一主成分取值的使用：
  - 可以利用第一主成分的值来评价水环境质量的优劣：
  - 主成分分值越小，对应的水环境质量越好
  - 水质评价结果
  - 根据第一主成分得分， 9个断面的水质由好到坏依次为：1#， 7#， 2#，8#， 6#， 5#， 3#， 9#， 4#

主成分分析的应用

- 例2：黄山市七个区县的生态区划
  - 27项指标（27个变量）：年平均气温（℃），日照时数（小时），多年平均降水量（mm），森林覆盖率（%），公益林面积（公顷），野生动植物分布系数（%），自然保护区面积（公顷），林地面积（公顷），牧草地面积（公顷），水域面积（公顷），耕地（ha），园地（ha），居民点及工矿用地（ha），交通用地（ha），未利用土地（ha），人口自然增长率（‰），人口密度（人/km^2），单位土地产出（万元/km^2），农民年均纯收入（元），国内生产总值（万元），国内生产总值发展速度，第三产业比重（%），农业产值（万元），林业产值（万元），牧业产值（万元），渔业产值（万元），工业产值（万元）
  - 7个区县（7个样本）：屯溪区，黄山区，徽州区，歙县，休宁，黟县，祁门

主成分分析的应用

- 例2：相关矩阵前4个特征值及其对应的主成分

特征值λ	12.3270	5.9691	4.7745	2.1589
贡献率（%）	45.65556	22.10778	17.68333	7.995926
累计贡献率（%）	45.65556	67.76333	85.44667	93.44259
主成分	L1	L2	L3	L4

主成分分析的应用

- 例2：四个主成分得分计算式

$$L1=-0.1583X_1-0.2213X_2-0.1346X_3-0.2663X_4+0.2725X_5+0.2695X_6+0.1184X_7+0.2645X_8+0.1114X_9+0.133X_{10}+0.2549X_{11}+0.1356X_{12}+0.1151X_{13}+0.1989X_{14}+0.1898X_{15}-0.2168X_{16}-0.2449X_{17}-0.238X_{18}-0.2477X_{19}-0.0207X_{20}-0.2028X_{21}-0.0994X_{22}+0.1742X_{23}+0.2569X_{24}+0.1023X_{25}+0.0788X_{26}+0.1198X_{27}$$

$$L2=0.1873X_1+0.0963X_2-0.2792X_3-0.1289X_4+0.0666X_5+0.0475X_6-0.3271X_7+0.0291X_8+0.1303X_9+-0.052X_{10}+0.1629X_{11}+0.2987X_{12}+0.0754X_{13}+0.0307X_{14}+0.2167X_{15}+0.0155X_{16}+0.1544X_{17}+0.128_1{}_8+0.1174X_{19}+0.3806X_{20}+0.1996X_{21}+0.1291X_{22}+0.3169X_{23}-0.1104X_{24}+0.3229X_{25}+0.0249X_{26}+0.2907X_{27}$$

$$L3=0.274X_1+0.1837X_2+0.0784X_3+0.044X_4+0.0026X_5+0.0264X_6+0.0493X_7+0.0485X_8-0.3383X_9-0.3812X_{10}+0.0403X_{11}+0.1292X_{12}-0.4073X_{13}+0.0618X_{14}+0.0758X_{15}+0.1923X_{16}-0.015X_{17}-0.0482X_{18}-0.1588X_{19}-0.0531X_{20}-0.1398X_{21}-0.3101X_{22}-0.0298X_{23}-0.1105X_{24}+0.1566X_{25}+0.4361X_{26}+0.114X_{27}$$

$$L4=-0.062X_1+0.0279X_2+0.2646X_3-0.0777X_4+0.1451X_5+0.1875X_6-0.1534X_7+0.2255X_8-0.2521X_9+0.1051X_{10}+0.0268X_{11}-0.1961X_{12}-0.065X_{13}+0.4192X_{14}+0.0852X_{15}+0.3072X_{16}+0.2218X_{17}+0.2857X_{18}+0.1004X_{19}+0.1382X_{20}-0.1708X_{21}+0.3503X_{22}+0.0239X_{23}+0.1103X_{24}-0.0211X_{25}-0.0408X_{26}-0.2583X_{27}$$

主成分分析的应用

- 例2：四个主成分的解释：
  - 第一主成分的表达式
  - 与森林覆盖率、公益林面积、野生动植物分布系数、林地面积、耕地面积以及林业产值成较大的正相关，
  - 与人口自然增长率、人口密度、单位土地产出、农民平均纯收入、国内生产总值发展速度成较大的负相关。

主成分分析的应用

- 例2：四个主成分的解释：
  - 第一主成分所占贡献率达到45.66%，而且从与其相关的原始变量来看，相关度比较均匀，所以这是一个综合成分。
    - 分析正相关的因子可知，森林资源是黄山市的一个极其敏感的生态因子，黄山市的生态环境的维系在很大程度上依赖于丰富的森林资源。而另一较大正相关的耕地因子可以作为一种生态缓冲的利用方式。当耕地处于森林资源的外缘，对森林资源起到的是一个自然保护的作用。
    - 负相关的因子包括人口自然增长率、人口密度、单位土地产出、农民平均纯收入、国内生产总值发展速度等，这些都是社会经济因子，都是社会和经济发展的标志。应该注意到，黄山市不是一个工业城市，大部分人口都是农民，农村人口、农业经济是黄山市的敏感社会经济因子。
  - 当样本的第一主成分数值大时，意味着跟森林资源相关的各个变量值偏大，描述社会经济发展的变量取值偏小。所以，该主成分得分大的样本适合于作为生态核心区，其自然生态系统需要继续得到较好的维持。

主成分分析的应用

- 例2：四个主成分的解释：
  - 第二主成分的表达式
  - 与国内生产总值成最大的正相关，其次，与农业产值，牧业产值和工业产值也有较大的正相关。第二主成分反映了部分敏感的经济因子。
  - 与之成最大负相关的是自然保护区面积，但是每个区县中自然保护区的面积占土地面积的百分率都非常小。自然保护区的面积不合适作为一个关键生态因子。另一个正相关较大的因子是园地面积。园地不是黄山市的核心生态系统，适合作为起缓冲作用的生态系统。
  - 综合来说，第二主成分得分高的样本适合作为生态缓冲区或者生态 建设区。

主成分分析的应用

- 例2：四个主成分的解释：
  - 第三主成分
  - 没有正相关较大的生态因子。
  - 与牧草地面积，水域面积，居民点及工矿用地，渔业产值成较大的负相关。充沛的水资源不仅是黄山市特有的自然资源、景观环境的重要组成部分，还是维系生态环境的重要生态因子。因此，基于充沛的水资源发展起来的渔业在这一主成分中也成为一个敏感因子。这一主成分可以归纳为水资源因子。

主成分分析的应用

- 例2：四个主成分的解释：
  - 第四主成分的贡献率只有8%，前三个主成分的累计贡献率已经达到85%以上。该主成分可以不加以考虑。
    - 为了减少信息的损失量，特别选取第四主成分，这样信息量达到93.44%
  - 第四主成分与交通用地，人口自然增长率和第三产业比值有较大的正相关
  - 这三个因子值的增大是社会发展和城市发展的要素

主成分分析的应用

- 例2：各个样本的四个主成分得分

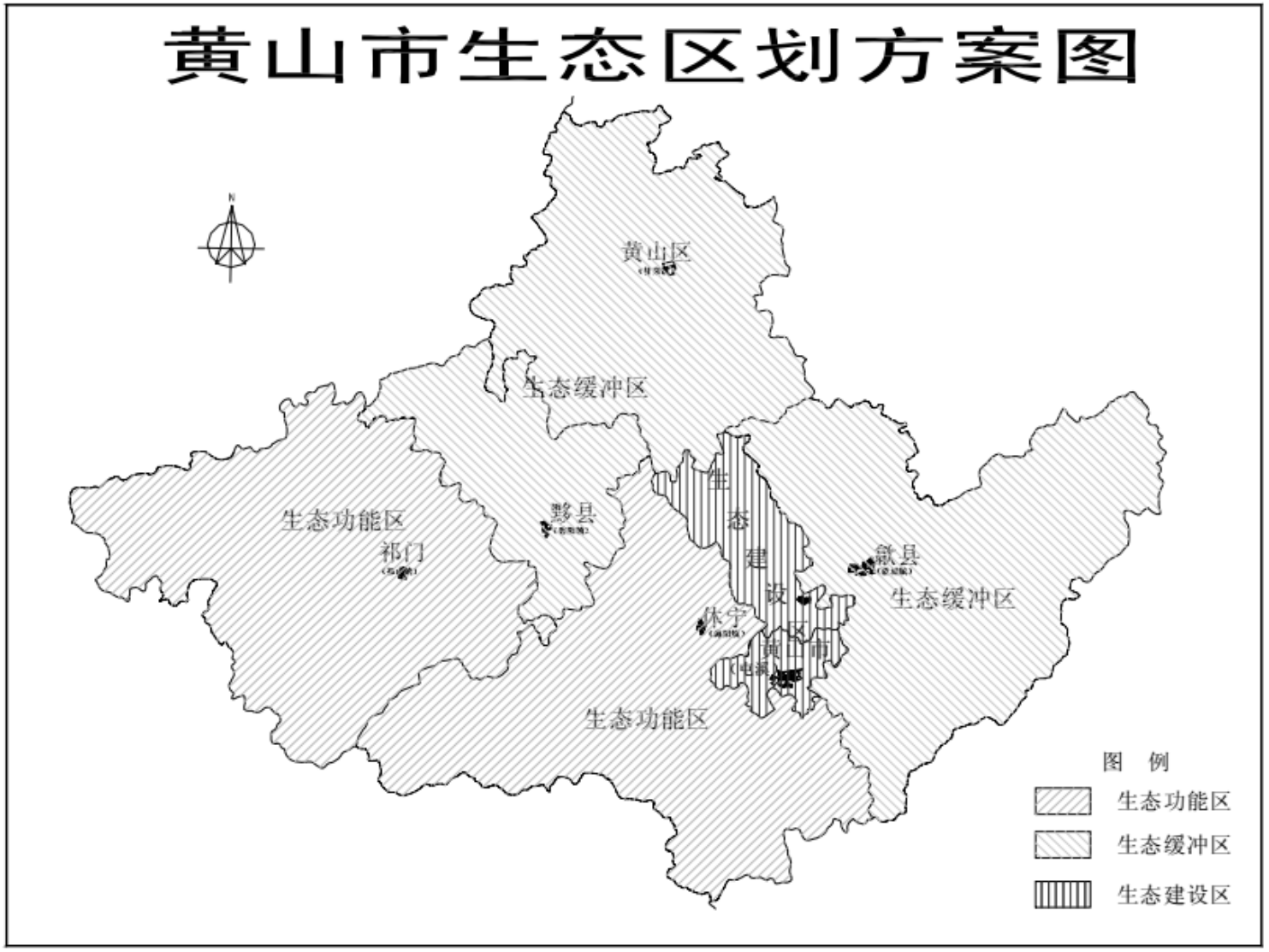
区县 主成分	屯溪区 1	黄山区 2	徽州区 3	歙县 4	休宁 5	黟县 6	祁门 7
L1	-6.5551	2.0263	-2.6555	2.056	2.7105	-0.4248	2.8426
L2	1.5533	-0.3828	-1.7053	4.2192	1.0748	-2.5615	-2.1976
L3	-0.5222	-4.7582	0.9384	1.0281	1.1709	0.8169	1.3261
L4	1.5326	-0.2261	-1.7856	-1.3626	1.1905	-1.0671	1.7183

主成分分析的应用

例2：黄山市生态区划方案

- 在主成分分析基础上，结合聚类分析和定性分析得出
- 生态建设区：
  - 屯溪区和徽州区
  - 属于典型城市生态系统，工作重点放在生态环境的重建和恢复上
- 生态缓冲区：
  - 歙县，黟县划为生态农业区
  - 黄山区划为生态旅游区
- 生态功能区：
  - 祁门和休宁
  - 重点保护森林生态系统和重要水源地

主成分分析的应用


黄山市生态区划方案图
黄山区
生态缓冲区
生态功能区
祁门
黟县
生
态
建
设
区
休宁
黄
山
市
歙县
生态缓冲区
生态功能区
图 例
生态功能区
生态缓冲区
生态建设区

主成分分析概念小结

- 什么是主成分?
  - 主成分是对变量共性的提取
- 用两个变量的情况举例说明
  - 原始变量及其包含的信息量:
    - 设变量X_1，X_2的相关程度为r（不妨假设$r \geq 0$），可以理解为X_1和X_2的原始数据中包含的信息量均为1，总信息量为2，r是共性部分。
  - 主成分、特征根及其包含的信息量:
    - 主成分是对总信息量的再分配。第一主成分包含的信息量为$1+r$，第二主成分包含信息量为$1-r$，总信息量仍为2。
    - 两个主成分对应的特征根分别为：λ_1“=”$1+r$和λ_2“=”$1-r$

主成分分析概念小结

- 用两个变量的情况举例说明
  - 变量之间相关程度与主成分分析的关系：
  - 第一主成分包括的信息随两变量相关程度（共性）的增大而增大
  - 当两变量完全相关时，
    - λ_1“=”2和λ_2“=”0，即第一主成分包括的了X_1，X_2的全部信息
  - 当两变量不相关时，
    - λ_1“=”λ_2“=”1，第一、二主成分包括X_1，X_2的信息各占一半
  - 当变量间的相关关系不明显时，作主成分分析意义不大

主成分分析概念小结

- 主成分是原变量的线性函数，是对原变量信息的综合：
  - 每个主成分可以看作是一个综合指标，该综合指标能够根据原变量具体估计出来，并可参与进一步的分析
- 主成分分析中最大的困难是对主成分的解释
  - 如果主成分得不到切合实际的解释，也就失去了主成分分析的意义
  - 对主成分的解释常伴有主观性
- 主成分分析本身往往并不是目的，而是达到目的的一种手段。因此，它经常被用在大型研究项目的某个中间环节里。例如，
  - 把它用在多元回归中，便产生了主成分回归，这种回归具有一些优良性质，可以克服原变量的多重共线性
  - 它还可以用于综合评价、聚类分析、判别分析等

中山大學
SUN YAT-SEN UNIVERSITY

谢谢！