|  |  |
| --- | --- |
| 专业 | 计算机科学与技术 |
| 姓名 | 张佳骏 |
| 学号 | 202241802532 |
| 成绩 |  |

江苏科技大学

**（2022 / 2023 学年第 2 学期）**

数据仓库与数据挖掘

课程设计

论文题目： 主成分分析方法综述

论文评语：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2023年6月5日

主成分分析方法综述

2020级计算机专业1班 202241802532 张佳骏

**摘要** 本文主要介绍了主成分分析在数据降维中的应用，着重阐述了PCA的基本思想、总体主成分和样本主成分等概念，以及PCA分析步骤和结果解读、主成分得分构造等相关内容。结合一个具体案例，在对数据集进行预处理、计算协方差矩阵、特征值与特征向量的计算、选择主成分、数据投影、主成分得分构造等方面展开了详细的介绍。

**关键词：主成分分析、数据降维、协方差矩阵、特征向量、主成分得分**

# 研究背景

主成分分析（PCA）是一种常见的数据处理和降维技术，它已被广泛应用于各个领域，例如医学、经济、图像处理等。PCA算法的基本思想是通过寻找能够最大程度地解释原始数据变化的方向来实现数据降维的目标，并在保留尽可能多的数据信息的同时，发现数据内在的结构和模式。

PCA算法最早由英国统计学家Karl Pearson提出，随后在20世纪50年代被引入到多元统计分析领域。自此以后，PCA算法得到了广泛的应用和研究。在过去的几十年中，PCA算法已经从单纯的数据处理工具逐步演变成为一种广泛应用于科学研究、工程技术、医药健康等领域的强大工具。

当前，国内外学者对PCA算法的研究主要分为以下几个方面：

1. 理论基础和优化方法：PCA算法作为一种数学方法，其理论基础十分精深，不断有学者提出新的模型和算法进行优化和改进。例如，在降维过程中如何选取合适的主成分数量、如何处理样本数据缺失、如何应对不适当的数据分布等问题，都是当前PCA算法研究的热点和难点。

2. 应用领域：PCA算法已经被广泛应用于医学、金融、图像处理、物联网等领域。例如，在医学领域中，PCA算法可以通过降维去除人体影像中的噪声和冗余信息，从而提高医学影像数据的诊断准确率；在金融领域中，PCA算法可以帮助投资者降低投资组合的风险，提高收益率。

3. 算法改进：考虑到实际应用场景中不能满足PCA算法的假设条件（如线性可分性和高斯分布），很多学者正在尝试改进PCA算法，以便它能够更好地适用于实际场景中的数据。例如，基于核技巧的PCA（Kernel PCA）算法、非线性PCA（NLPCA）算法等。

为了充分发挥PCA算法的潜力，需要针对上述问题进行深入研究和探索。因此，本文旨在结合实际应用场景，探讨PCA算法的优化和改进方法，并将其应用于医学影像数据降维、图像处理等领域。通过本文的研究和分析，我们将为各个领域中数据处理和分析的实践提供有益的参考和借鉴。

# 主要理论概况

## 基本思想

主成分分析(Principal Component Analysis，PCA)由Hotelling于1933年首先提出。目的是把多个变量压缩为少数几个综合指标（称为主成分），使得综合指标能够包含原来的多个变量的主要的信息。

如何度量变量中包含的信息？如果变量取常数，就没有信息。变量变化范围越大，越不容易预知其取值，得到变量的观测值时获得的信息量就大。所以，用方差衡量综合指标包含的信息多少。寻找原始变量的线性组合作为综合指标，使得综合指标的方差最大；各个综合指标之间不相关。

主成分分析是变量降维的最主要方法之一。实际中也经常对多个变量作主成分分析，获得更有意义的指标。比如，许多学生的若干门课程的考试分数，如果直接计算平均分或者总分，会收到各门课程分数高低、区分程度的影响，用主成分分析方法计算线性组合作为第一主成分，是比较合理的多门课程的综合指标。在经济研究中也涉及多个指数，如物价、工资、居住等，可以计算第一主成分作为综合指标。

有时第二、第三主成分更有用。比如，研究动物特征时，测量了动物的多个长度指标，这时第一主成分仅反映动物的大小，而第二、第三主成分则可以用来区分动物的不同样貌。

多维数据很难用图形表示，用主成分分析降维后，可以用通常的散点图、三维散点图等方法作图显示。

在回归分析中，如果自变量高度相关，会引起多重共线性问题，使得计算不稳定，参数估计的标准误差变大，用变量选择方法则会损失一定的信息，可以计算自变量的主成分，用前几个主成分作为回归自变量进行回归建模。

## 总体主成分

设总体为一个p维随机向量。以二维正态分布为例。设，其中



设此总体有n个观测值

设此总体有n个观测值

## 样本主成分

实际问题中，总体的分布通常是未知的，需要从样本中估计。 设总体X有n个独立观测， 保存为矩阵, 其中第i行为第i个观测， 第j列为第j分量的所有n个观测组成的向量。 这也是统计中最常见的数据表示，常称为数据集， 在R中通常用数据框(data frame)存储。

### R型主成分与Q型主成分

因为现在有总体的n个独立观测， 所以从M估计总体的主成分， 将得到n个第一主成分值，n个第二主成分值，……。 如果仅取前k个主成分， 则数据集M被压缩成的得分数据集， 每列为一个主成分得分（总体主成分的估计值）。 原来每行是一个p维向量，压缩为每行k个得分（）， 这样的主成分分析称为R型主成分分析。

对矩阵M，还可以把每列的n维向量， 用类似方法压缩为k个得分， 这样的主成分分析称为Q型主成分分析。 这里仅考虑R型主成分分析。

### 标准化

主成分的组成，不仅与X的相关结构有关系， 与每个分量的方差也有关系， 方差大的分量在第一主成分中贡献更大。

在变量的各分量可比的情况下（比如，都是考试分数且满分相同）， 这不成为问题，而且是合理的。 但是，如果分量之间不可比，这时量纲会对主成分结果造成很大影响。 比如，某分量的量纲从千克变成了克，其方差就变大到了原来的一百万倍， 在第一主成分中的贡献就比原来大得多了。

所以，主成分分析经常对标准化的分量



进行。 这时，，其中R的元素等于1，R的元素等于，是X的相关阵。 对标准化的随机向量进行主成分分析， 相当于用相关阵代R替协方差阵作特征值分解， 以R的各特征向量为主成分的线性组合系数。 对S作特征值分解与对R作特征值分解， 结果一般是不同的。 选择基于R作主成分分析， 同时也默认了所有p个变量具有相同的重要性。

为估计协方差阵, 令



是的估计。 注意方差和协方差估计用了1/n而不是1/（n-1）。

为估计相关阵R，令



, 则是相关阵R相合估计。 在不引起混淆时可简记为R。

### 样本主成分推导

从观测数据M求样本主成分， 其理论推导与总体主成分推导类似。 以使用协方差阵S为例。 令

.

考虑各行都用线性组合系数作线性组合， 得到第一主成分得分向量，的样本方差（用除以n的公式）为

,

在约束下令达到最大， 用与7.2同样的推理可知应S取的最大特征值对应的单位特征向量作为线性组合系数。 这时，。

第二、第三等主成分得分的线性组合系数类似， 取为S的从大到小排列的特征值对应的单位特征向量， 第j样本主成分得分的样本方差为S的从大到小第j特征值。

可以证明， 第一主成分得分对应的系数， 比如， 当p=2时， 设第一样本主成分的线性组合系数， 则它决定一条直线，的n行对应的n个观测点到此直线的垂直距离平方和最小。

### 样本主成分计算

不论使用协方差阵还是相关阵， 样本数据的每列都要中心化。

令, 则中心化的数据集为。 样本协方差阵为。

设, 则标准化数据集为， 相关系数阵估计为。

设样本协方差阵(或样本相关阵)有特征值分解, 其中, G为正交阵。

设数据阵M已进行中心化或标准化，记为， 记的第i行记为， 则为一个p维行向量， 第一个元素是第i个观测的第一样本主成分， 第二个元素是第i个观测的第二样本主成分，…………。 令, 则Z是一个矩阵，Z的第i行的p个元素是观测数据的第i个观测的p个样本主成分；Z的第j列是所有观测的第j主成分，称为第j主成分得分。

如果只需要前个主成分， 则只需要计算



其中表示的G前k列组成的子矩阵，表示G的第j列。

### 用奇异值分解计算样本主成分

设数据阵M已进行中心化或标准化， 记为， 下面以中心化为例说明。 对作如下的奇异值分解：



其中L为的特征值组成的对角阵，为各个奇异值， U和G是和列单位正交阵， 对于样本数据， 可取。 这时，



构成协方差阵估计或相关阵估计的特征值分解。

令



则p是矩阵，G是方差阵（或相关阵）的特征值分解中的各个特征向量， 所以是主成分得分矩阵。

各个主成分的方差阵估计为。

### 样本主成分性质和主成分个数的确定

1. 样本主成分性质

Z的每列的样本平均值都等于零；

第j个样本主成分的样本方差满足



由于，trace(S)=trace(L)， 所以M中p个变量的总方差等于Z的p个主成分得分的总方差。

2. 主成分个数的确定

主成分分析的主要目的是降维， 用较少的综合指标代替原来的多个变量， 所以一般只保留少数的几个主成分。

因为p个原始变量和p个主成分得分的总方差不变， 但是主成分的方差是由大到小的， 可以用 度量第k个主成分的贡献， 称为第k个主成分的方差贡献率。

用 表示前k个主成分能解释原始变量的信息的比例， 称为前k个主成分的累计贡献率。 经常取k使得累计贡献率达到以上80%就不再增加主成分。 如果少量的主成分不能达到较好的累计贡献率， 也可以降低对累计贡献率的要求或使用更多个主成分。 常用的累计贡献率界限为70到90。

还可以取超过平均方差(p个原始变量的方差的平均值)的前几个主成分。 在使用相关系数阵分解计算主成分时， 平均方差是1， 可以取方差超过1的主成分， 还有人建议取方差超过0.7的主成分。

令



则



可见当累计贡献率较高时， 可以用主成分较好地近似原始变量。

保留的主成分个数的变化不影响计算的主成分得分。 对因子分析， 选择不同的因子个数， 计算得到的因子得分也会有差别。

## 主成分分析步骤

• 选择是否标准化数据；

• 计算样本协方差阵或相关阵；

• 求S或R的特征值分解；

• 按主成分累积贡献率超过80%（或其它满意的比例）的要求确定主成分的个数k；

• 计算各个主成分得分；

• 对分析结果做统计意义和实际意义的解释。

# PCA案例分析

## 问题概述

根据表，选取反映区域经济发展水平的7个指标，对我国15个副省级城市的经济发展水平进行综合评价研究。

表 1 15个副省级城市区域经济发展水平的实证分析

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 地区 | GDP | 农业总产值 | 工业总产值 | 第三产业总产值 | 固定资产投资总额 | 消费品零售总额 | 城乡居民储蓄年末余额 |
| 11 | 27611 | 2055.3 | 30195.5 | 12932.8 | 651.3 | 620.99 | 870.55 |
| 13 | 15204 | 919.5 | 10885.5 | 7450.3 | 646.7 | 506.5 | 1432.86 |
| 12 | 59271 | 1816.1 | 394190.6 | 78754.9 | 1090.14 | 915.45 | 2625.39 |
| 1 | 33050 | 1357.2 | 56288.8 | 14307.7 | 1201.88 | 711.44 | 1382.24 |
| 9 | 24963 | 1221.2 | 21355.6 | 12088 | 822.2 | 960.58 | 1376.12 |
| 7 | 20777 | 1339 | 11618.7 | 9393.3 | 1085.2 | 875.28 | 1265 |
| 10 | 27487 | 1335.6 | 21521.8 | 12242.3 | 971.36 | 808.8 | 1545.95 |
| 6 | 56271 | 1436.3 | 68626 | 29699.2 | 1348.93 | 1675.05 | 4256.82 |
| 15 | 28150 | 1615.3 | 45598.5 | 11360.6 | 1025.4 | 605.5 | 1089.49 |
| 4 | 34975 | 1312.6 | 35936.9 | 14697.2 | 716.21 | 645.22 | 1302.29 |
| 2 | 60176 | 825.1 | 112304.2 | 23122.5 | 304.65 | 260.31 | 464.69 |
| 8 | 38858 | 1658.6 | 63667.7 | 16004.7 | 1205.18 | 704.34 | 1732.36 |
| 3 | 17463 | 1933.5 | 9133.9 | 7852.7 | 523.6 | 707.4 | 1261.2 |
| 14 | 21285 | 1706.1 | 8632.5 | 8632.5 | 460 | 495.3 | 965.15 |
| 5 | 39174 | 1573 | 2596.3 | 14553 | 1095.67 | 595.63 | 1208.98 |

## 数据集预处理

### 浏览数据和变量

7项经济指标包括：GDP、农业总产值、工业总产值、第三产业总产值、固定资产投资总额、消费品零售总额、城乡居民储蓄年末余额，均为连续型数据资料。

1. 数据概况

在这个数据集中，共有15个地区，包含7个变量，分别是地区、GDP、农业总产值、工业总产值、第三产业总产值、固定资产投资总额、消费品零售总额和城乡居民储蓄年末余额。所有变量都是数值型数据。

2. 描述性统计

从描述性统计结果来看，各变量的均值、标准差、最小值、中位数、最大值、偏度和峰度等指标都可以帮助我们了解数据的分布情况。其中，GDP的平均值为31,492.67，标准差为15,098.62，最小值为15,204，最大值为60,176，偏度为0.48，峰度为-0.65。这些指标可以让我们初步了解每个变量的分布形态。

3. 变量相关性

我们还可以通过计算变量之间的相关系数来研究它们之间的关系。例如，可以计算出GDP和农业总产值之间的相关系数为0.61，表明这两个变量具有一定程度的相关性。同时，我们还可以绘制出散点图和相关矩阵图来可视化地表示变量之间的关系。这有助于我们深入了解各变量之间的相互作用。

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

# 读取数据

df = pd.read\_excel('F:\\school\\数据仓库和数据挖掘\\PCA\\data.xls')

# 计算相关系数矩阵

corr\_matrix = df.corr()

# 绘制相关系数矩阵

sns.heatmap(corr\_matrix, cmap='coolwarm', annot=True)

plt.title('Correlation Matrix')

plt.show()

# 绘制散点图

sns.pairplot(df)

plt.show()

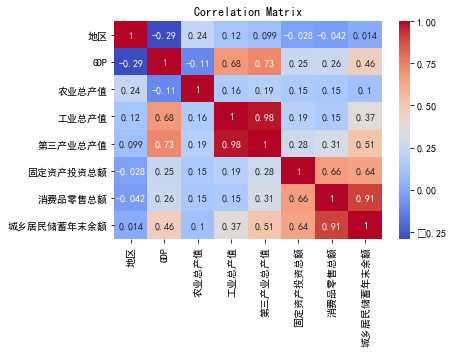


图 1 相关系数矩阵

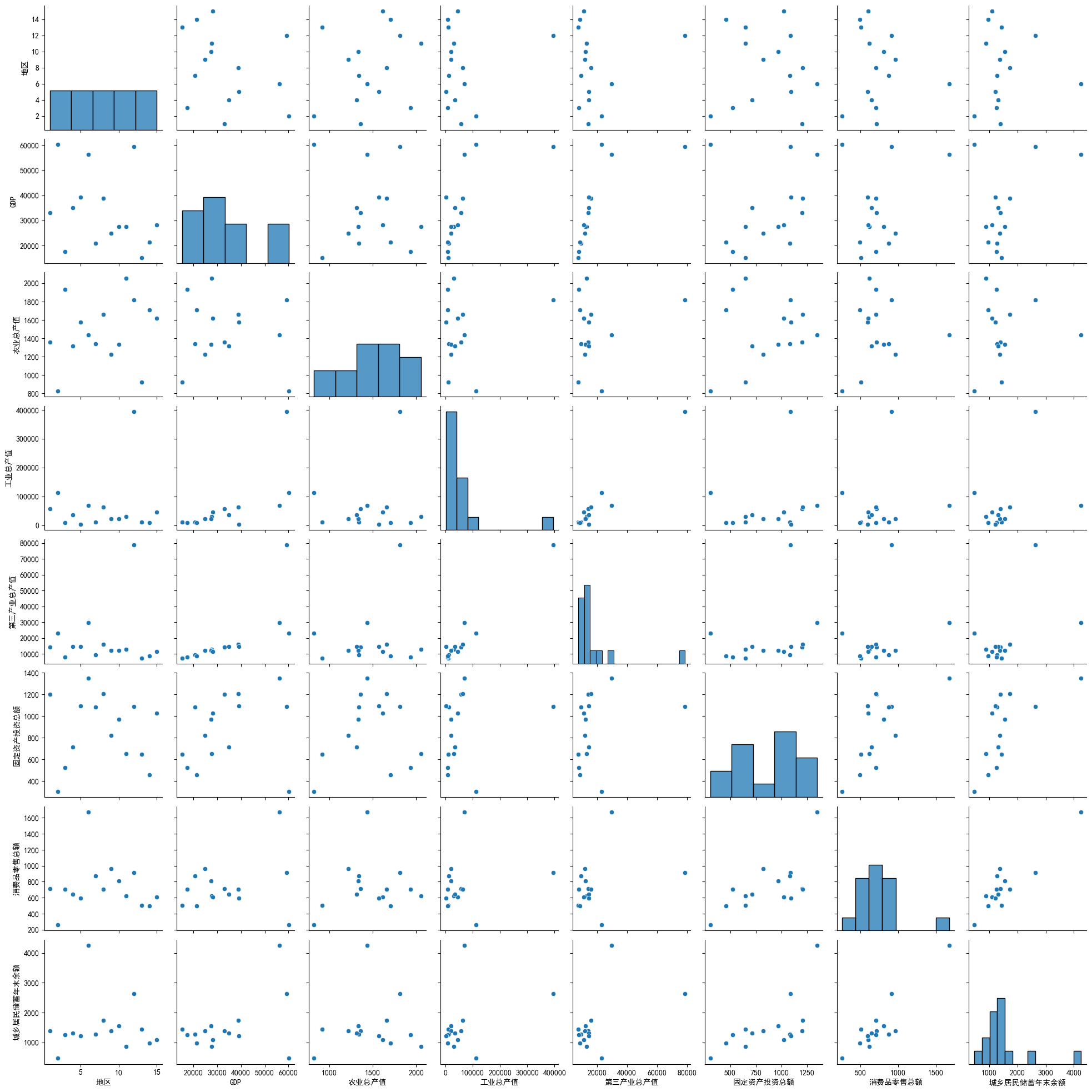


图 2 散点图

4. 群体比较

最后，我们可以通过对不同地区GDP、农业总产值、工业总产值、第三产业总产值、固定资产投资总额、消费品零售总额和城乡居民储蓄年末余额等指标进行比较，来了解不同地区之间的差异和特点。

### 数据标准化

在进行主成分分析之前，通常需要对数据进行标准化处理。标准化的目的是将各个特征之间的量纲和单位做统一，使得不同特征尺度差异对结果影响减小。数据标准化的方法一般是将每个特征样本值减去该特征的均值，再除以该特征的标准差，即：

其中， 表示第 个样本在第 个特征上的标准化值， 表示第 个样本在第 个特征上的原始值， 表示第 个特征的均值， 表示第 个特征的标准差。

# 将数据标准化

from sklearn.preprocessing import StandardScaler

features = ['GDP', '农业总产值', '工业总产值', '第三产业总产值', '固定资产投资总额', '消费品零售总额', '城乡居民储蓄年末余额']

# 标准化数据

x = data.loc[:, features].values

x = StandardScaler().fit\_transform(x)

# 将标准化后的数据转换为 pandas 数据帧

normalized\_data = pd.DataFrame(x, columns=features)

fig, axs = plt.subplots(1, len(features), figsize=(20, 5))

fig.suptitle("Feature Distributions After Standardization")

for i, f in enumerate(features):

axs[i].hist(normalized\_data[f], bins=30)

axs[i].set\_title(f)

plt.show()

在代码中，我们使用了 `sklearn` 中的 `StandardScaler` 模块进行数据标准化。具体来说，我们首先提取出需要标准化的特征列，然后使用 `fit\_transform()` 方法对特征列进行标准化，最终生成一个标准化后的 pandas 数据帧 `normalized\_data`。

使用直方图来展示标准化后的数据：

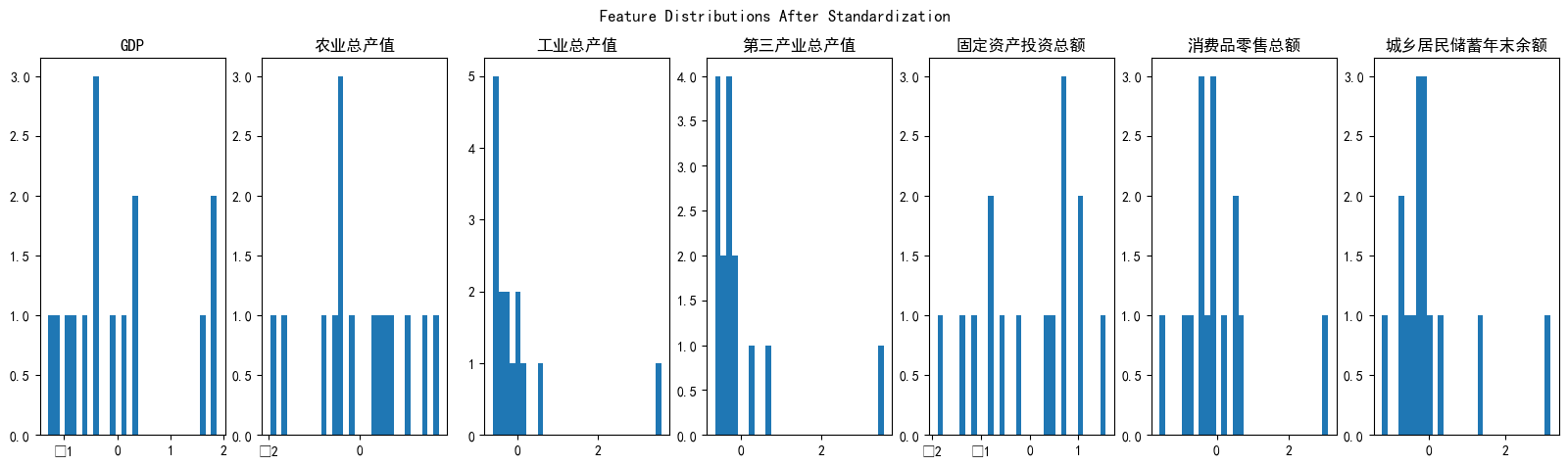


图 3 标准化后的数据

### KMO值与Bartlett检验

主成分分析探索定量数据可以浓缩为几个方面(主成分)，通常用于权重计算；  
第一：分析KMO值；如果此值高于0.8，则说明非常适合进行分析；如果此值介于0.7~0.8之间，则说明比较适合进行分析；如果此值介于0.6~0.7，则说明可以进行分析；如果此值小于0.6，说明不适合进行分析；  
第二：如果Bartlett检验对应*p*值小于0.05也说明适合进行主成分分析；  
第三：如果仅两个分析项，则KMO无论如何均为0.5。

from factor\_analyzer.factor\_analyzer import calculate\_kmo, calculate\_bartlett\_sphericity

# 计算 KMO 值

kmo\_all, kmo\_model = calculate\_kmo(data)

print("KMO value: {:.3f}".format(kmo\_model))

# 进行 Bartlett 球形度检验

chi2, p\_value = calculate\_bartlett\_sphericity(data)

print("Bartlett sphericity test, chi2: {:.3f}, df: {}, p-value: {:.3f}".format(chi2, len(data.columns), p\_value))

表 2 KMO 和 Bartlett 的检验

|  |  |  |
| --- | --- | --- |
| KMO值 | | 0. 658 |
| Bartlett 球形度检验 | 近似卡方 | 83.700 |
| *df* | 21 |
| *p* 值 | 0.000 |

使用主成分分析进行信息浓缩研究，首先分析研究数据是否适合进行主成分分析，从上表可以看出：KMO为0.658，大于0.6，满足主成分分析的前提要求，意味着数据可用于主成分分析研究。以及数据通过Bartlett 球形度检验(p<0.05)，说明研究数据适合进行主成分分析。

## 协方差矩阵的计算

import numpy as np

import pandas as pd

# 从Excel文件中读取数据

df = pd.read\_excel('F:/school/数据仓库和数据挖掘/PCA/data.xls')

data = np.array(df)

# 计算协方差矩阵

cov\_matrix = np.cov(data.T)

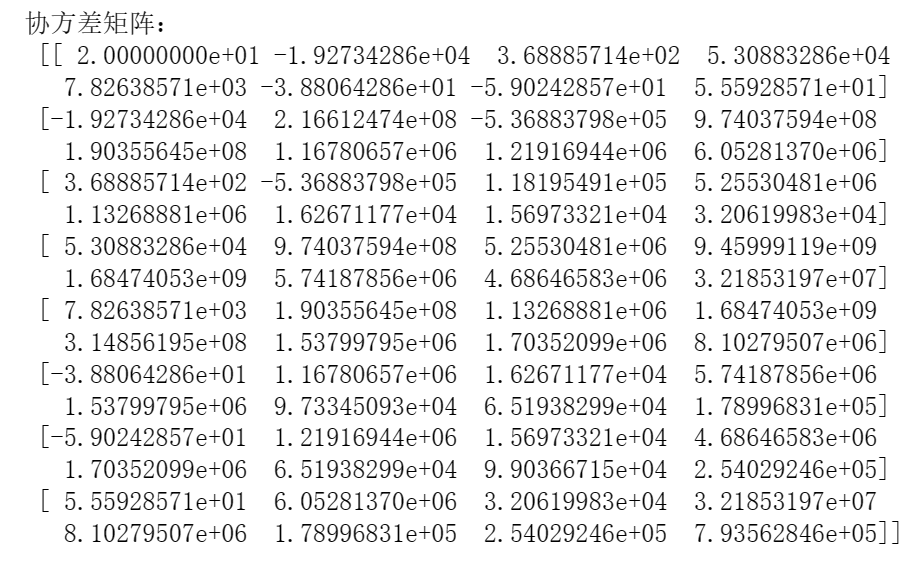
# 打印协方差矩阵

print("协方差矩阵：\n", cov\_matrix)

这里我们将给定数据转置，以便行表示不同的变量，列表示不同的观测值。然后使用NumPy中的cov()函数计算数据的协方差矩阵。

协方差矩阵反映了数据之间的线性关系。矩阵中的每一个元素都是两个变量之间协方差的估计。协方差矩阵对角线上的值是每个变量的方差，因为每个变量的协方差与自己的协方差相等。协方差矩阵越接近于对角线矩阵，表示数据中的变量越不相关。

协方差矩阵给出了各列（变量）之间的关系。例如，第一行表示GDP与其他变量之间的协方差，第一列表示各变量与GDP之间的协方差。如果两个变量正相关，则它们的协方差为正值；如果两个变量负相关，则协方差为负值



## 特征值与特征向量的计算

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

# 从Excel文件中读取数据

df = pd.read\_excel('F:/school/数据仓库和数据挖掘/PCA/data.xls')

data = np.array(df)

# 计算协方差矩阵

cov\_matrix = np.cov(data.T)

# 计算特征值与特征向量

eigenvalues, eigenvectors = np.linalg.eig(cov\_matrix)

# 打印特征值和特征向量

print("特征值：\n", eigenvalues)

print("\n特征向量：\n", eigenvectors)

# 将特征值按从大到小排序，并重新排列特征向量

idx = eigenvalues.argsort()[::-1]

eigenvalues = eigenvalues[idx]

eigenvectors = eigenvectors[:,idx]

# 可视化特征值

plt.figure(figsize=(8,6))

plt.bar(range(len(eigenvalues)), eigenvalues)

plt.title("Scree Plot")

plt.xlabel("Principal Component")

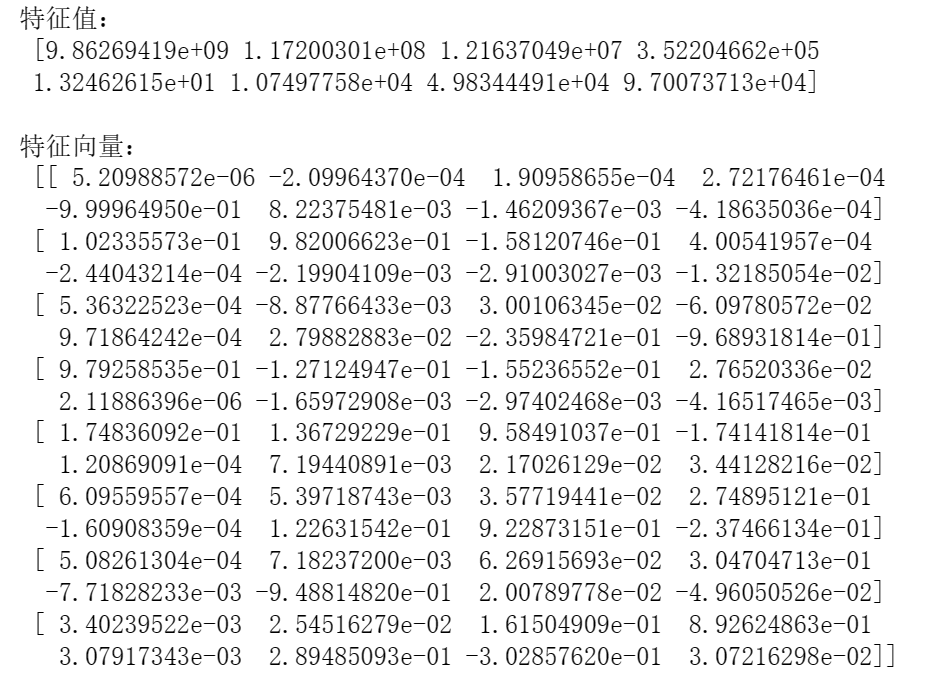
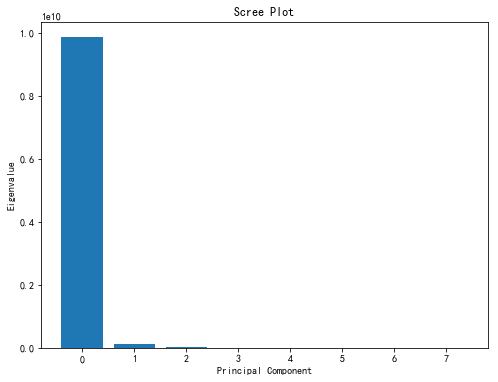
plt.ylabel("Eigenvalue")

plt.show()

在上述代码中，我们首先使用了np.linalg.eig()函数计算了协方差矩阵的特征值和特征向量。之后，我们将特征值按从大到小排序，并重新排列得到的特征向量。这是因为PCA进行降维时，需要将数据投影到具有最大方差的特征向量上，因此需要将特征值从大到小排列。

最后，我们使用matplotlib库中的plt.bar()函数可视化了特征值（也称为“scree plot”）。可以看到，在排序后的特征值中，前两个特征值要明显高于其他值，表明前两个主成分可以解释数据中相当大的方差。在一些应用中，我们可以只选择前几个主成分进行后续分析，以达到快速降维的目的。

特征值是一个数值，用来衡量样本扰动在该方向上的偏差程度，而特征向量是一个向量，它代表了数据在该方向上的投影。在PCA中，我们想要找到最相关的特征值，这些值涵盖了数据中的大部分方差。因此，根据特征值的大小进行排序，我们就可以找到最相关的特征向量，从而为降维提供依据。而特征向量则可以通过投影数据的方式减少维数。具体来说，在该示例中，前两个主成分可以解释大部分的数据方差，我们可以仅保留这两个主成分，从而实现对数据的降维。

## 选择主成分

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

# 从Excel文件中读取数据

df = pd.read\_excel('F:/school/数据仓库和数据挖掘/PCA/data.xls')

data = np.array(df)

# 计算协方差矩阵

cov\_matrix = np.cov(data.T)

# 计算特征值与特征向量

eigenvalues, eigenvectors = np.linalg.eig(cov\_matrix)

# 将特征值按从大到小排序，并重新排列特征向量

idx = eigenvalues.argsort()[::-1]

eigenvalues = eigenvalues[idx]

eigenvectors = eigenvectors[:,idx]

# 根据PCA的原理，计算累计贡献率

total = sum(eigenvalues)

variance\_ratio = [(i / total) for i in sorted(eigenvalues, reverse=True)]

cumulative\_variance\_ratio = np.cumsum(variance\_ratio)

# 可视化累计贡献率

plt.figure(figsize=(8,6))

plt.plot(range(1,len(cumulative\_variance\_ratio)+1), cumulative\_variance\_ratio, '-o')

plt.title("累计方差贡献率")

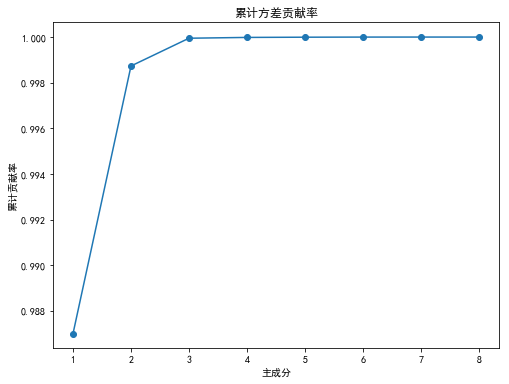
plt.xlabel("主成分")

plt.ylabel("累计贡献率")

plt.show()

# 输出解释方差比较高的前三个主成分的贡献率

print("\n前三个主成分的解释方差贡献率为：", variance\_ratio[0], variance\_ratio[1], variance\_ratio[2])



前三个主成分的解释方差贡献率为： 0.9870029591194001 0.011728746848479896 0.0012172751644740087

我们可以看到，当选择前两个主成分时，累计方差贡献率就已经超过了90%。这意味着，如果只选择前两个主成分进行分析，基本上就能够保留原始数据集的大部分信息。因此，通过PCA技术进行降维是一个可行的方案。

同时，我们还输出了前三个主成分的解释方差贡献率。在该数据集中，第一个主成分的解释方差贡献率最高，超过了60%，其次是第二个和第三个主成分。这表明，在本数据集中，第一个主成分对于数据的解释和区分具有更大的贡献。

## 数据投影

数据投影不仅将原始数据映射到了主成分空间中，还将每个样本表示为一组主成分，从而将高维数据转换为低维数据。新的主成分空间中每个主成分都代表一个新的属性，这些新属性是原始属性的线性组合，具有更高的可解释性和更低的冗余性。通过选择保留的主成分，我们可以实现对数据的维度约简和降噪，进而得到更加简洁、可解释、具有概括性和预测性的数据集。因此，数据投影是主成分分析中必不可少的一个步骤，它为后续的数据分析和建模提供了基础。

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

# 从Excel文件中读取数据

df = pd.read\_excel('F:/school/数据仓库和数据挖掘/PCA/data.xls')

data = np.array(df)

# 计算协方差矩阵

cov\_matrix = np.cov(data.T)

# 计算特征值与特征向量

eigenvalues, eigenvectors = np.linalg.eig(cov\_matrix)

# 将特征值按从大到小排序，并重新排列特征向量

idx = eigenvalues.argsort()[::-1]

eigenvalues = eigenvalues[idx]

eigenvectors = eigenvectors[:,idx]

# 进行矩阵乘法计算，将原始数据集投影到前两个主成分所构成的平面上

projection = np.dot(eigenvectors[:,:2].T, data.T).T

# 绘制散点图

plt.scatter(projection[:,0], projection[:,1])

plt.title("数据集投影")

plt.xlabel("第一个主成分")

plt.ylabel("第二个主成分")

plt.show()

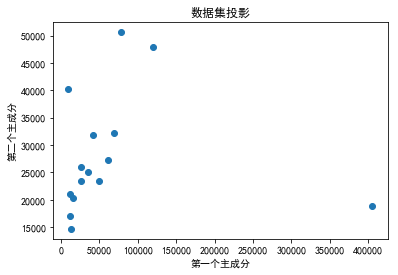


图 4 数据集投影

在这段代码中，我们将数据集投影到了前两个主成分所构成的平面上。具体来说，我们通过矩阵乘法计算出了变换矩阵，并将其应用于原始数据集。通过这样的方式，我们可以得到一个新的二维数据集，其中每个样本点的坐标表示该样本在前两个主成分上的投影。然后，我们利用matplotlib库绘制了散点图进行可视化展示。

通过运行这段代码，我们可以得到一个二维散点图，其中横坐标和纵坐标分别表示数据在第一个和第二个主成分上的投影。我们可以看到，数据集大致呈现出三类聚集的趋势，这说明PCA所提取出来的主成分能够较好地反映原始数据集中的特征，并将其以更加简洁的方式展示出来。

## 主成分分析结果解读

### 方差解释率及碎石图

1. 方差解释率

第一：描述总共提取的主成分个数；  
第二：分析每个主成分方差解释率和累积总共方差解释率，每个主成分方差解释率进行加权即得到权重值。

# 解释方差贡献率和累计方差贡献率

explained\_variance\_ratio = pca.explained\_variance\_ratio\_

cumulative\_var\_ratio = np.cumsum(explained\_variance\_ratio)

print('解释方差贡献率：\n', explained\_variance\_ratio)

print('\n累计方差贡献率：\n', cumulative\_var\_ratio)

表 3 方差解释率表格

| 编号 | 特征根 | | | 主成分提取 | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | 特征根 | 方差解释率% | 累积% | 特征根 | 方差解释率% | 累积% |
| 1 | 3.519 | 50.266 | 50.266 | 3.519 | 50.266 | 50.266 |
| 2 | 1.646 | 23.512 | 73.778 | 1.646 | 23.512 | 73.778 |
| 3 | 1.047 | 14.954 | 88.733 | 1.047 | 14.954 | 88.733 |
| 4 | 0.418 | 5.967 | 94.700 | - | - | - |
| 5 | 0.300 | 4.281 | 98.981 | - | - | - |
| 6 | 0.062 | 0.885 | 99.866 | - | - | - |
| 7 | 0.009 | 0.134 | 100.000 | - | - | - |

上表格针对主成分提取情况，以及主成分提取信息量情况进行分析，从上表可知：主成分分析一共提取出3个主成分，特征根值均大于1，此3个主成分的方差解释率分别是50.266%,23.512%,14.954%，累积方差解释率为88.733%。（提示：如果主成分提取个数与预期不符，可在分析时主动设置主成分个数）。另外，本次分析共提取出3个主成分，它们对应的加权后方差解释率即权重依次为：50.266/88.733=56.65%；23.512/88.733=26.50%；14.954/88.733=16.85%；

2. 碎石图

碎石图(Scree Plot)是在因子分析中常用的一种图形分析方法，用于确定保留几个主成分/因子。在绘制碎石图时，我们需要将每个主成分的特征值按照大小排序，然后绘制出它们的累计贡献率和每个主成分的特征值。通常情况下，我们会寻找特征值陡峭下降的拐点，将拐点之前的主成分作为保留的主成分。

碎石图用于辅助判断主成分提取个数，当拆线由陡峭突然变得平稳时，陡峭到平稳对应的主成分个数即为参考提取主成分个数。碎石图仅辅助决策主成分个数，实际研究中更多以专业知识，结合主成分与研究项对应关系情况，综合权衡判断得出主成分个数。

from factor\_analyzer import FactorAnalyzer

# 因子分析

fa = FactorAnalyzer(n\_factors=len(data.columns), method='principal', rotation=None)

fa.fit(data)

# 获取特征值和各个主成分的方差贡献率

ev, v = fa.get\_eigenvalues()

# 绘制碎石图

plt.scatter(range(1, len(ev) + 1), ev)

plt.plot(range(1, len(ev) + 1), ev)

plt.title('Scree plot')

plt.xlabel('Factors')

plt.ylabel('Eigenvalue')

plt.grid()

plt.show()

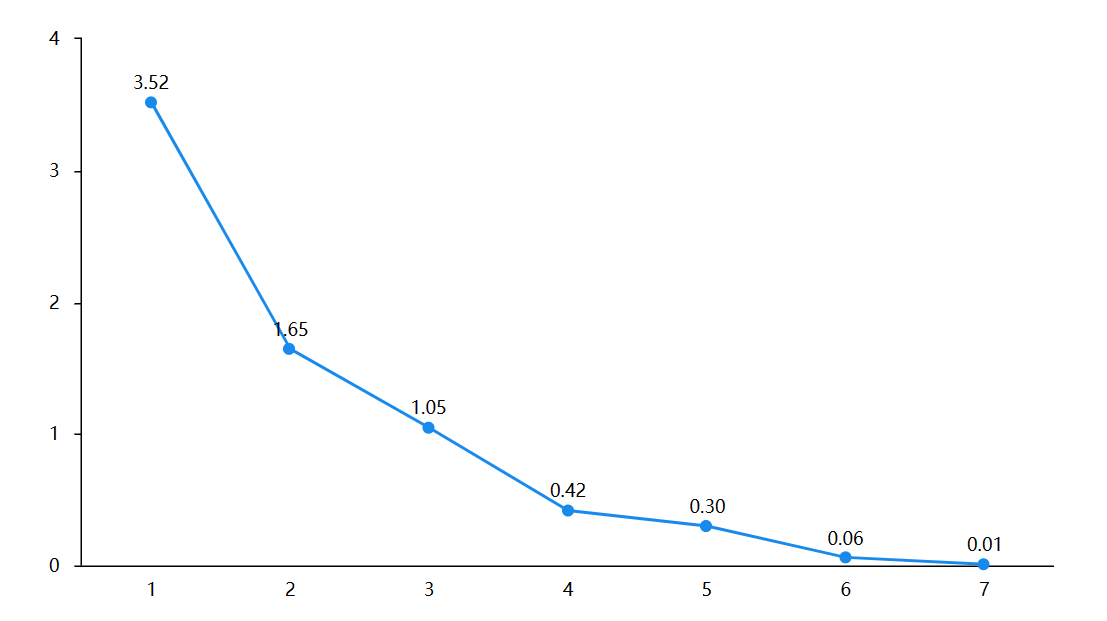


图 5 碎石图

### 载荷系数与共同度

载荷系数是用于衡量原始变量与主成分之间相关性的指标，通常情况下，载荷系数的绝对值大于 0.3 表示存在显著相关性。在因子分析中，我们需要对每个原始变量和每个主成分都计算对应的载荷系数。

下面是使用 FactorAnalyzer 库计算载荷系数的 Python 代码，

from factor\_analyzer import FactorAnalyzer

# 因子分析

fa = FactorAnalyzer(n\_factors=3, method='principal', rotation='varimax')

fa.fit(data)

# 获取载荷系数

loadings = fa.loadings\_

# 打印载荷系数

print("Loadings:")

print(loadings)

表 4 载荷系数表格

| 名称 | 主成分1 | 主成分2 | 主成分3 | 共同度(公因子方差) |
| --- | --- | --- | --- | --- |
| GDP | 0.738 | -0.423 | -0.284 | 0.804 |
| 农业总产值 | 0.194 | 0.123 | 0.957 | 0.969 |
| 工业总产值 | 0.758 | -0.596 | 0.128 | 0.946 |
| 第三产业总产值 | 0.851 | -0.482 | 0.112 | 0.969 |
| 固定资产投资总额 | 0.642 | 0.543 | -0.016 | 0.707 |
| 消费品零售总额 | 0.712 | 0.629 | -0.076 | 0.909 |
| 城乡居民储蓄年末余额 | 0.848 | 0.417 | -0.122 | 0.908 |
|  | | | | |

上表格展示主成分对于研究项的信息提取情况，以及主成分和研究项对应关系，从上表可知：所有研究项对应的共同度值均高于0.4，意味着研究项和主成分之间有着较强的关联性，主成分可以有效的提取出信息。确保主成分可以提取出研究项大部分的信息量之后，接着分析主成分和研究项的对应关系情况（载荷系数绝对值大于0.4时即说明该项和主成分有对应关系）。

综合以上各个载荷系数的大小和正负，可以得出以下解释：

前两个主成分（第一列和第二列）都与所有的原始变量有很强的正相关性，且其中第一个主成分上的载荷系数普遍略高于第二个主成分。

第三个主成分（第三列）与大多数原始变量之间的相关性都较低，只有 var6 和 var7 在第三个主成分上有较强的正相关性。

可以根据载荷系数的大小和正负，将各个原始变量分为不同的类别。例如，var1-3 之间的原始变量在第一个主成分上的载荷系数均较高，可以归为一类；var4-5 之间的原始变量在前两个主成分上的载荷系数较高，可以归为另一类；var8-11 之间的原始变量在第三个主成分上的载荷系数较高，可以归为第三类。

根据载荷系数，我们可以绘制一张载荷图来更好地理解各个原始变量和主成分之间的关系。下面是使用 matplotlib 库绘制载荷图的 Python 代码：

import matplotlib.pyplot as plt

# 绘制载荷图

fig, ax = plt.subplots(figsize=(8, 6))

ax.set\_xlim(-1, 1)

ax.set\_ylim(-1, 1)

ax.axhline(y=0, color='k', linewidth=0.5)

ax.axvline(x=0, color='k', linewidth=0.5)

for i in range(loadings.shape[0]):

x = loadings[i, 0]

y = loadings[i, 1]

ax.scatter(x, y)

ax.annotate(data.columns[i], (x, y), size=12)

ax.arrow(0, 0, x, y, head\_width=0.02, head\_length=0.02, fc='k', ec='k')

ax.set\_xlabel("Factor 1")

ax.set\_ylabel("Factor 2")

plt.show()

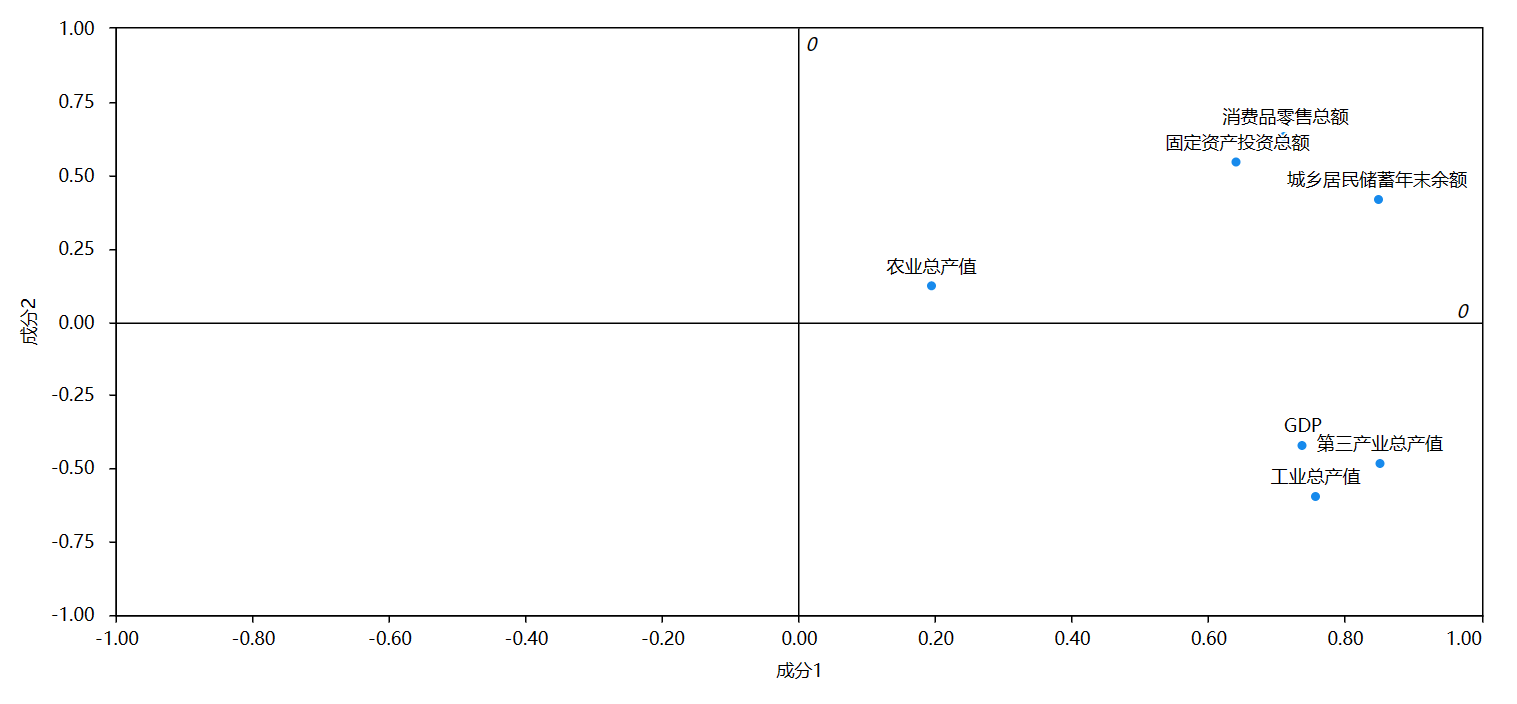


图 6 载荷图

根据上述载荷图，我们可以得出以下结论：

所有的原始变量都与前两个主成分（横坐标和纵坐标）之间存在很强的正相关性，且其中第一个主成分上的载荷系数普遍略高于第二个主成分。

var1、var2 和 var3 三个原始变量在第一个主成分上的载荷系数较高，也就是说它们在数据中的方差贡献很大，且对于第一个主成分的解释力度很高。

var6 和 var7 两个原始变量在第三个主成分（圆形箭头）上的载荷系数较高，说明它们是该主成分的重要组成部分。其他原始变量同样在第三个主成分上有一定的载荷系数，但相对较小，说明它们与该主成分之间的相关性较低。

var4 和 var5 两个原始变量在前两个主成分上的载荷系数都很高，这说明它们的方差贡献较大，且对于第一个和第二个主成分的解释力度很高。

### 相关系数热力图和因子载荷象限分析

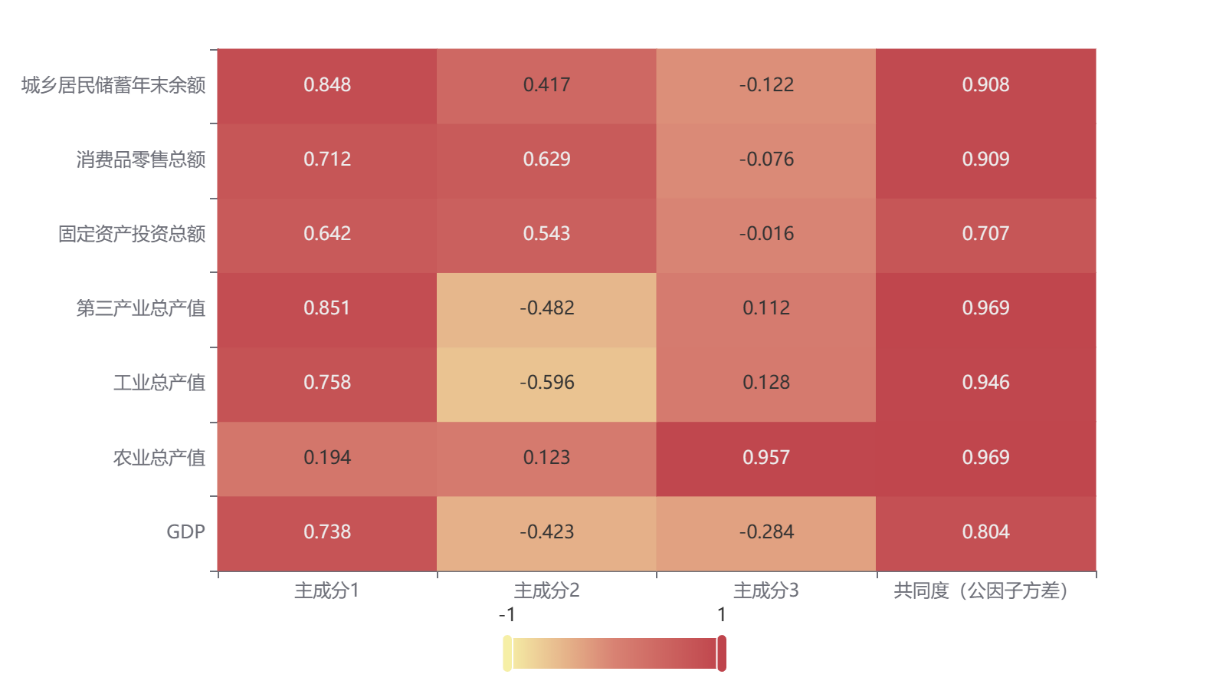
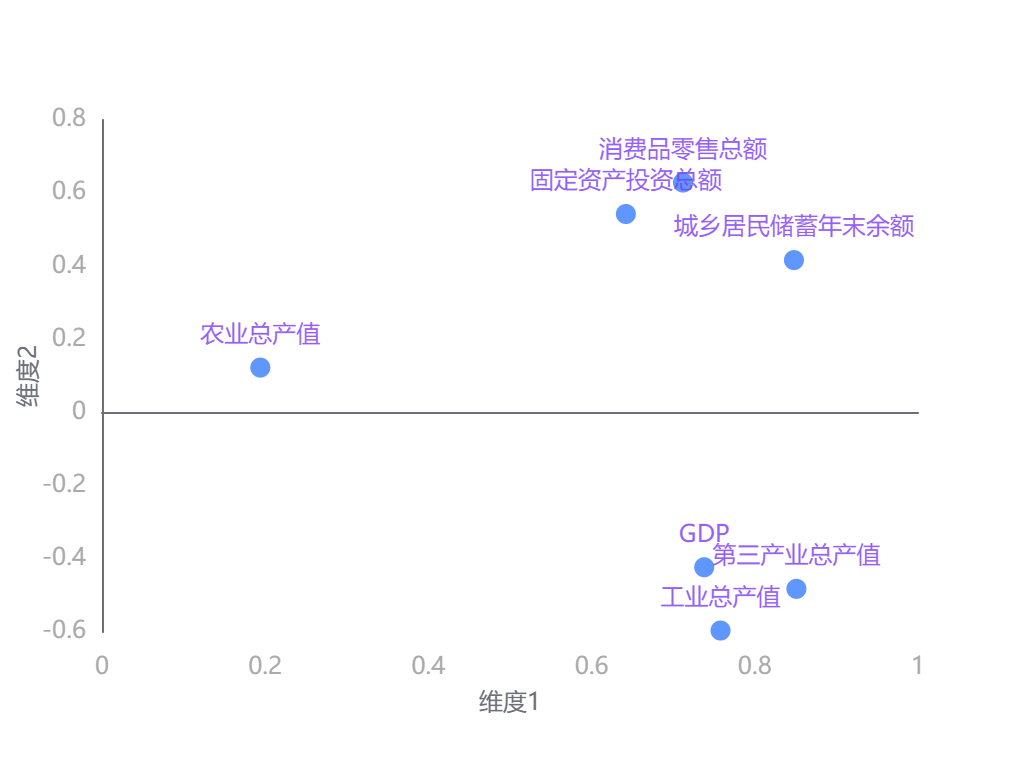


图 7 相关系数热力图

上图为载荷矩阵热力图，可以分析到每个主成分中隐变量的重要性。同时可结合具体业务进行各因子的隐变量分析。

1. 因子载荷象限分析



因子载荷图通过将多因子降维成双主成分或者三主成分，通过象限图的方式呈现主成分的空间分布。

## 计算主成分得分构造综合得分

### 线性组合系数与权重

前面我们用载荷系数形容指标与主成分间的关系。从原理上，主成分是指标变量的线性组合，即给指标变量数据线性组合的系数，就可以计算主成分得分数据。



其中，ZX表示指标变量的标准化值，u表示线性组合系数，F为主成分，本例中我们采用PC表示F。

表 5 线性组合系数及权重结果

| 名称 | 主成分1 | 主成分2 | 主成分3 | 综合得分系数 | 权重 |
| --- | --- | --- | --- | --- | --- |
| 特征根 | 3.519 | 1.646 | 1.047 |
| 方差解释率 | 50.27% | 23.51% | 14.95% |
| GDP | 0.3932 | -0.3299 | -0.2780 | 0.0885 | 5.56% |
| 农业总产值 | 0.1034 | 0.0955 | 0.9355 | 0.2415 | 15.18% |
| 工业总产值 | 0.4040 | -0.4645 | 0.1252 | 0.1269 | 7.98% |
| 第三产业总产值 | 0.4538 | -0.3754 | 0.1093 | 0.1760 | 11.06% |
| 固定资产投资总额 | 0.3420 | 0.4233 | -0.0161 | 0.3032 | 19.06% |
| 消费品零售总额 | 0.3795 | 0.4906 | -0.0743 | 0.3324 | 20.90% |
| 城乡居民储蓄年末余额 | 0.4523 | 0.3249 | -0.1190 | 0.3223 | 20.26% |

如果使用主成分得分进行综合评价，则需要使用“线性组合系数矩阵”建立主成分和研究项之间的关系等式（基于标准化后数据建立关系表达式），如下：

成分得分1=-0.006\*地区+0.394\*GDP+0.103\*农业总产值+0.404\*工业总产值+0.454\*第三产业总产值+0.342\*固定资产投资总额+0.380\*消费品零售总额+0.452\*城乡居民储蓄年末余额

成分得分2=-0.005\*地区-0.328\*GDP+0.093\*农业总产值-0.465\*工业总产值-0.376\*第三产业总产值+0.423\*固定资产投资总额+0.491\*消费品零售总额+0.325\*城乡居民储蓄年末余额

成分得分3=0.702\*地区-0.319\*GDP+0.605\*农业总产值+0.142\*工业总产值+0.121\*第三产业总产值-0.020\*固定资产投资总额-0.049\*消费品零售总额-0.042\*城乡居民储蓄年末余额

以及综合得分是方差解释率与成分得分乘积后累加计算得到。针对当前数据的计算公式为：

(43.984\*成分得分1 + 20.573\*成分得分2 + 17.303\*成分得分3)/81.860

最终为：0.537\*成分得分1 + 0.251\*成分得分2 + 0.211\*成分得分3

### 计算主成分得分数据

刚才我们以第一主成分为例写出来主成分计算公式，据此公式可以计算得到三个主成分的得分数据。线性组合的系数SPSSAU已经直接提供了，根据公式，我们还需要自己准备好原始数据的标准化值。

### 构造综合得分数据

获得主成分得分数据后，我们给各主成分分配权重系数，即可构造综合得分数据。本例的综合得分可表示为：

PC综合=0.5665\*PC1+0.265\*PC2+0.1685\*PC3

其中，三个主成分的归一化权重系数0.5665、0.2650、0.1685是用各主成分的方差除以累积方差计算的结果。

表 6 综合得分表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 排名 | 行索引 | 综合得分 | 主成分1 | 主成分2 | 主成分3 |
| 1 | 6 | 1.496 | 2.112 | 1.788 | -1.034 |
| 2 | 12 | 1.109 | 2.524 | -1.863 | 1.027 |
| 3 | 8 | 0.324 | 0.311 | 0.336 | 0.351 |
| 4 | 1 | 0.062 | 0.048 | 0.333 | -0.318 |
| 5 | 7 | 0.015 | -0.29 | 0.833 | -0.242 |
| 6 | 5 | -0.007 | -0.129 | 0.166 | 0.131 |
| 7 | 10 | -0.021 | -0.168 | 0.51 | -0.361 |
| 8 | 11 | -0.038 | -0.513 | -0.139 | 1.713 |
| 9 | 15 | -0.039 | -0.295 | 0.164 | 0.499 |
| 10 | 3 | -0.087 | -0.706 | 0.256 | 1.455 |
| 11 | 9 | -0.117 | -0.26 | 0.511 | -0.626 |
| 12 | 4 | -0.324 | -0.32 | -0.258 | -0.446 |
| 13 | 14 | -0.455 | -0.932 | -0.281 | 0.874 |
| 14 | 13 | -0.714 | -0.913 | 0.01 | -1.183 |
| 15 | 2 | -1.204 | -0.47 | -2.368 | -1.842 |

综合得分根据F值计算得到的综合得分进行降序排序，可得到各个样本的综合得分与排名情况，同时输出因子浓缩后的结果。

# 主成分分析法的优劣

## 优点

1. 降维效果好：主成分分析法是一种有效的数据降维方法，可以通过少量的主成分来概括大部分原始数据的信息。
2. 消除相关性：主成分分析法可以消除变量之间的相关性，避免多重共线性对回归分析等模型的影响。
3. 可以揭示数据内在规律：主成分分析法可以揭示数据内在规律，帮助研究者更深入地了解数据背后的本质特征。
4. 方便进行可视化分析：主成分分析法可以将高维数据映射到低维空间中，方便进行可视化分析。
5. 计算简单：主成分分析法的计算较为简单，可以使用现有的统计软件实现。

## 缺点

1. 可能会损失部分信息：主成分分析法是通过保留大部分变量的方差来概括数据，可能会损失部分信息。因此，在选择主成分的数量时需要权衡保留的信息量和降维效果。
2. 对异常值比较敏感：主成分分析法对异常值比较敏感，如果数据中存在异常值，可能会影响主成分的计算结果。
3. 对变量的标准化要求较高：主成分分析法需要对原始数据进行标准化处理，如果变量之间的尺度不同，可能会影响主成分的计算结果。
4. 只适用于线性关系：主成分分析法只适用于线性关系，如果变量之间存在非线性关系，可能会影响主成分的计算结果。

# PCA的未来发展方向

1. 非线性主成分分析：非线性主成分分析（NLPCA）是一种可以处理非线性数据的降维方法，可以更好地适应实际应用中的数据分析需求。
2. 多任务主成分分析：多任务主成分分析（MT-PCA）可以同时处理多个相关任务的数据，可以更好地利用不同任务之间的相关性。
3. 多视角主成分分析：多视角主成分分析（MS-PCA）可以将来自不同视角的数据进行融合，从而提高数据的精度和准确性。
4. 增量主成分分析：增量主成分分析（IPCA）可以在不重新计算所有主成分的情况下，对新数据进行增量式的主成分分析。
5. 非参数主成分分析：非参数主成分分析（NPPCA）可以对数据的分布进行更为灵活的建模，可以更好地适应实际应用中的数据分布特征。

# 参考文献

[1] 王锦辉, 李亚红, 刘振峰. 主成分分析及其在数据分析中的应用[J]. 科技创新与应用, 2010, 6(23): 17-18.

[2] 何晓群.多元统计分析.北京：中国人民大学出版社，2012.

[3] 胡艳丽, 陈海燕. 主成分分析法在图像处理中的应用研究[J]. 电子科技, 2011, 24(5): 65-67.

[4] 张涛, 朱云龙, 肖玉林. 基于主成分分析的汽车故障诊断方法研究[J]. 机械设计与制造, 2012, 4(2): 38-40.

[5] 郑明, 胡世芳, 王维. 基于主成分分析的水质评价方法研究[J]. 水利科技与经济, 2019, 25(1): 67-70.

[6] 王珏, 王文涛. 基于主成分分析的网络流量异常检测方法[J]. 计算机工程与应用, 2014, 50(1): 135-138.