主成分分析算法

主成分分析算法（Principal component analysis，PCA)）是一种无监督学习算法，针对没有响应变量而仅有特征变量的数据集，其主要作用就是降维。在很多时候，各个特征变量之间可能会存在较多的信息重叠，或者说相关性比较强，比如线性回归分析中的多重共线性关系，而当我们的样本观测值数较少，但是选取的变量过多的话，就会产生高维数据及其带来的“维度灾难”问题，也可以理解成是模型的自由度太小，进而造成构建效果欠佳。这时候就需要用到降维的方法，即针对有过多特征变量的数据集，在尽可能不损失信息或者少损失信息的情况下，将多个特征变量减少为少数几个潜在的主成分，这几个主成分可以高度概括数据中的信息，这样，既减少了变量个数，又能最大程度的保留原有特征变量中的信息。

目录

CONTENTS

PART 01

主成分分析算法的基本原理

主成分分析算法的基本原理

主成分分析是一种降维分析的统计过程，该过程通过正交变换将原始的 n 维数据集变换到一个新的被称做主成分的数据集中，也就是将众多的初始特征变量整合成少数几个相互无关的主成分特征变量，而这些新的特征变量尽可能地包含了初始特征变量的全部信息，然后用这些新的特征变量来代替以前的特征变量进行分析。比如在线性回归分析算法中，我们可能会遇到样本示例个数小于变量个数，即高维数据情形，或者原始特征变量之前存在较强相关性造成多重共线性的情况，那么我们完成可以先进行主成分分析，以提取的主成分作为新的特征变量，再进行线性回归分析等有监督学习。

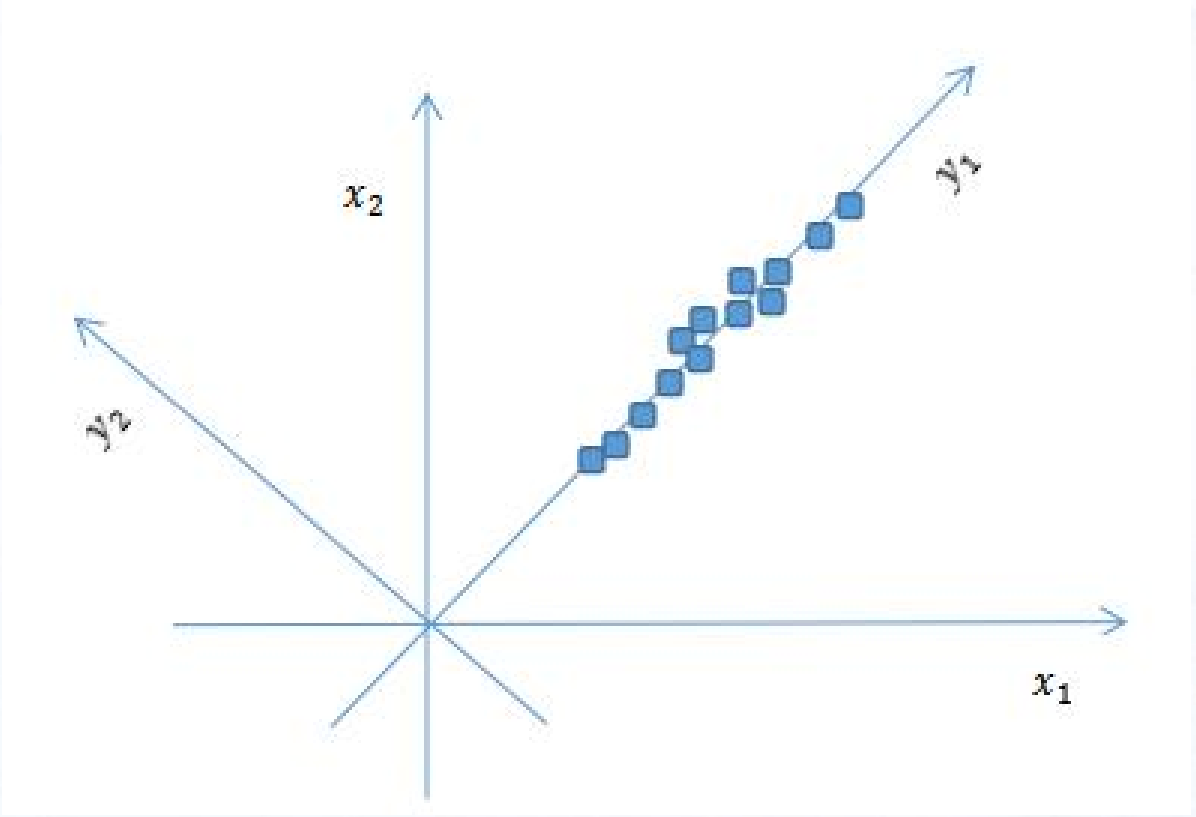
主成分分析算法的基本原理

当然，主成分分析本质上只是一种坐标变换，经过变换之后得到的主成分由于集合了多个原始特征变量的信息，所以其经济含义很可能不再清晰，无法有效就特征变量对响应变量的具体影响关系做出清晰解释，但这不妨碍其在降维方面的巨大优势，尤其是针对机器学习这一更多关注预测而不是关注解释的领域更是如此。

具体来说，主成分分析法从原始特征变量到新特征变量是一个正交变换（坐标变换），通过正交变换，将其原随机向量（分量间有相关性）转化成新随机向量（分量间不具有有相关性），也就是将原随机向量的协方差阵变换成对角阵。变换后的结果中，第一个主成分具有最大的方差值，每个后续的主成分在与前述主成分正交条件限制下也具有最大方差。降维时仅保存前 $m(m < n)$ 个主成分即可保持最大的数据信息量。

主成分分析算法的基本原理

如下图所示，原来特征空间内有两个特征变量x_1和x_2，样本观测值需要由这两个特征变量共同描述，但是我们进行正交变换（坐标变换）之后，将原特征变量x_1和x_2转化为新特征变量y_1和y_2，可以发现样本观测值几乎只用y_1一个特征变量，通过y_1的大小就可以进行描述。



主成分分析算法的数学概念

主成分分析算法的数学概念为：设有原始特征向量 $X=(X_1, X_2 \ldots\ldots, X_p)^T$，是一个$p$维随机特征向量，首先将其标准化 $ZX=(ZX_1, ZX_2 \ldots\ldots, ZX_p)^T$，使得每一变量的平均值为 0，方差为 1，之所以需要进行标准化，是因为如果变量之间的方差差别较大时，主成分分析就会被较大方差的变量所主导，使得分析结果严重失真。

然后考虑它的线性变换，如果样本示例个数 n 大于等于特征变量个数p（这也是大多数情况），则提取主成分F_i即为：

$$F_i = a_{1i} \times ZX_1 + a_{2i} \times ZX_2 \ldots\ldots + a_{pi} \times ZX_p$$

主成分分析算法的数学概念

也就是说进行坐标变换，将p维随机特征向量 X 转换成p维随机特征向量。但是这一坐标变换需满足如下优化条件：一方面是第一个主成分F_1最可能多地保留原始特征向量 X 的信息，实现途径是使的F_1的方差尽可能大；另一方面接下来的每一个主成分都要尽可能多地保留原始特征向量 X 的信息，但同时又不能跟前面已经提取的主成分的信息有所重叠，也就是说各个主成分之间是相互正交的，或者说需要满足 $cov(F_i, F_j) = 0$, 其中$j = 1,2,\ldots\ldots, i-1$。在满足上述条件下，各主成分的方差依次递减，不同的主成分之间相互正交（没有相关性），达到前几个主成分 F_i就可以代表原始特征向量 X 大部分信息的效果。

而如果样本示例个数 n 小于特征变量个数p，则只能提取 n − 1 个主成分，不然主成分之间就会产生严格多重共线性。

在主成分个数的具体确定方面，如果从尽可能保持原始特征变量信息的角度，最终选取的主成分的个数可以通过各个主成分的累积方差贡献率来确定，一般情况下以累积方差贡献率大于等于 85%为标准。如果单纯从降维的角度，可以直接限定提取的主成分的个数，以达到降维效果为目的保留主成分。

主成分的特征值

除了前面所述的方差贡献率之外，主成分的特征值也可以代表该主成分的解释能力，特征值是方差的组成部分，所有主成分的特征值加起来就是分析中主成分的方差之和，即主成分的“总方差”。

由于我们提取的各个主成分之间是完全不相关的，分析的是一个零相关矩阵，标准化为单位方差。比如针对10个原始特征变量我们提取了10个主成分，10个主成分的总方差就是10。

主成分的特征值

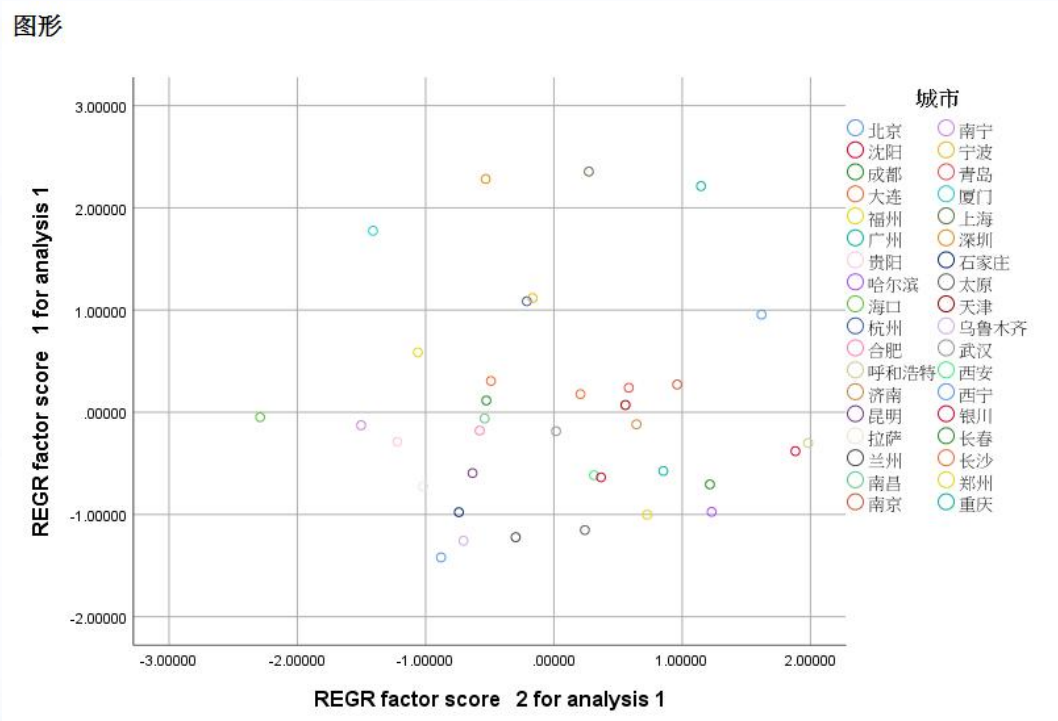
比如在某次分析中第一个主成分的特征值为6.61875，其方差贡献率就是66.19%（6.61875/10），或者说该主成分能够解释总方差的66.19%；第二个主成分的特征值为1.47683，，其方差贡献率就是14.77%（1.47683/10），或者说该主成分能够解释总方差的14.77%。

特征值越大解释能力越强，通常情况下只有特征值大于1的主成分是有效的，因为平均值就是1（结合前面讲解的10个主成分的总方差就是10来理解），如果某主成分的特征值低于1，则说明该主成分对于方差的解释还没有达到平均水平，所以不建议保留。该方法也可以用来作为“确定主成分个数”的判别标准。

样本示例的主成分得分

每个样本示例主成分的具体取值，称为主成分得分。以下图片即为主成分得分的图形展示。（图片来源：《SPSS数据挖掘与案例分析应用实践》杨维忠著，机械工业出版社，第11章 城镇居民消费支出结构研究及政策启示）。

该研究针对某年度中国大中城市城镇居民消费支出科目提取了两个主成分，第一主成分（纵轴）代表食品、家庭设备用品及服务、交通和通讯、教育文化娱乐服务、居住、杂项商品和服务；第二主成分（横轴）代表衣着、医疗保健，图中直观的展示了各大中城市在两个主成分方面的优势（短板），可以通过划分为四个象限（（0，0）为原点）的方式进行解释，比如位于第一象限的有上海、广州、北京、南京、青岛、大连、天津，表示这7个城市在消费支出的两个主成分方面都领先其他城市。



主成分载荷

主成分载荷表示每个变量对于主成分的影响，常用特征向量矩阵来表示，以下图片即为一个示例。（图片来源：《Stata统计分析从入门到精通》杨维忠、张甜著，清华大学出版社，第11章 主成分分析与因子分析）。

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
V2	0.3669	-0.0292	0.1501	0.0770	-0.4435	-0.4078	0.0228
V3	0.2243	-0.3380	0.6741	0.2771	0.4064	-0.0283	0.2359
V4	0.3760	-0.0203	-0.0404	0.1164	-0.2705	-0.4969	0.0681
V5	0.3705	0.0179	0.0527	-0.0671	0.3307	-0.0240	-0.8547
V6	0.3510	-0.1240	0.1737	0.0699	-0.4773	0.7470	-0.0310
V7	0.3595	-0.0411	-0.2491	-0.0309	0.4535	0.1048	0.3765
V8	0.3669	0.0081	-0.3220	-0.0573	0.0097	0.0834	0.0282
V9	0.3590	0.0807	-0.3474	-0.1394	0.1280	0.0474	0.2086
V10	0.0489	0.7182	0.0121	0.6840	0.0740	0.0850	0.0077
V11	0.1341	0.5871	0.4493	-0.6347	0.0152	0.0004	0.1468

Variable	Comp8	Comp9	Comp10	Unexplained
V2	-0.3880	-0.2994	0.4821	0
V3	0.2699	-0.0737	0.0649	0
V4	0.1808	0.3612	-0.5960	0
V5	-0.0965	-0.0368	-0.0628	0
V6	-0.0742	0.0401	-0.1803	0
V7	-0.6171	0.2547	-0.0440	0
V8	0.5108	0.4043	0.5710	0
V9	0.2938	-0.7302	-0.2070	0
V10	0.0060	-0.0175	0.0268	0
V11	0.0318	0.0987	-0.0108	0

图中 Comp1~Comp10 代表提取的 10 个主成分，V2~V11 代表 10 个原始变量，针对主成分 1，其在 V2~V11 上的载荷分别是 0.3669、0.2243、0.0489、0.1341。需要说明的是，每个主成分荷载的列式平方和为 1，如针对主成分 1（Comp1），即有：

$$0.3669^2 + 0.2243^2 + \cdots + 0.0489^2 + 0.1341^2 = 1$$

PART 02

数据准备

数据准备

本节我们以“数据11.1”为例进行讲解，其中记录的是《中国2021年1-3月份地区主要能源产品产量统计》，数据摘编自《中国经济景气月报202104》。该数据文件中共有21个变量，分别是V1~V21，分别代表地区、汽油万吨、煤油万吨、柴油万吨、燃料油万吨、石脑油万吨、液化石油气万吨、石油焦万吨、石油沥青万吨、焦炭万吨、煤气亿立方米、火力发电量亿千瓦小时、水力发电量亿千瓦小时、核能发电量亿千瓦小时、风力发电量亿千瓦小时、太阳能发电量亿千瓦小时、原煤万吨、原油万吨、天然气亿立方米、煤层气亿立方米、液化天然气万吨。

我们下面我们针对汽油万吨、煤油万吨、柴油万吨、燃料油万吨、石脑油万吨、液化石油气万吨、石油焦万吨、石油沥青万吨、焦炭万吨、煤气亿立方米，即V2~V10共10个变量开展主成分分析。

载入分析所需要的模块和函数

在进行分析之前，我们首先载入分析所需要的模块和函数，读取数据集并进行观察。

示例

参阅教材内容

变量设置及数据处理

示例

参阅教材内容

特征变量相关性分析

示例

参阅教材内容

PART 03

主成分分析算法示例

主成分提取及特征值、方差贡献率计算

示例

参阅教材内容

绘制碎石图观察各主成分特征值

示例

参阅教材内容

绘制碎石图观察各主成分方差贡献率

示例

参阅教材内容

绘制碎石图观察主成分累积方差贡献率

示例

参阅教材内容

绘制二维图形展示样本示例在前两个主成分上的得分

示例

参阅教材内容

绘制三维图形展示样本示例在前三个主成分上的得分

示例

参阅教材内容

输出特征向量矩阵，观察主成分载荷

示例

参阅教材内容

PART 04

习　题

习题

继续使用“数据11.1”，针对火力发电量亿千瓦小时、水力发电量亿千瓦小时、核能发电量亿千瓦小时、风力发电量亿千瓦小时、太阳能发电量亿千瓦小时、原煤万吨、原油万吨、天然气亿立方米、煤层气亿立方米、液化天然气万吨等变量，即V12~V21共11个变量开展主成分分析。

(1) 载入分析所需要的库和模块

(2) 变量设置及数据处理

(3) 特征变量相关性分析

(4) 主成分提取及特征值、方差贡献率计算

习题

(5) 绘制碎石图观察各主成分特征值

(6) 绘制碎石图观察各主成分方差贡献率

(7) 绘制碎石图观察主成分累积方差贡献率

(8) 计算样本示例的主成分得分

(9) 绘制二维图形展示样本示例在前两个主成分上的得分

(10) 绘制三维图形展示样本示例在前三个主成分上的得分

输出特征向量矩阵，观察主成分载荷。

感谢聆听

THANKS