Hans Journal of Data Mining 数据挖掘, 2021, 11(4), 203-216
Published Online October 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.114019

Hans汉斯

主成分分析的几个应用理解及R语言实践

赵　霖，李裕梅

北京工商大学数学与统计学院，北京

收稿日期：2021年8月19日；录用日期：2021年9月23日；发布日期：2021年9月30日

摘　要

主成分分析是一种常用的简化数据集的技术，也是一种应用广泛的多元统计分析方法。在各高校开设的多门课程中，主成分分析的理论都是重点内容。但在教学过程中，我们发现其理论与实际应用间还有很多值得理解、挖掘和实践验证的地方。本文针对主成分分析过程进行回顾，并主要探讨其几个应用情况，即基于主成分分析的数据降维、基于主成分分析的综合评价、基于主成分分析的关键特征确定、基于主成分分析的样本聚类等几个方面。我们详细梳理并部分推导和补充每个应用的理论过程，整理各种应用在一些文献里的使用场景，给出我们对各个应用的R语言实践代码和相应分析等。这些应用的理论过程和R语言实践，有助于对主成分分析进行深刻理解和融会贯通，为主成分分析的学习和使用提供重要的参考。

关键词

主成分分析，数据降维，综合评价，关键特征确定，样本聚类，R语言实践

Several Applications of Principal Component Analysis and Corresponding R Language Practice

Mu Zhao, Yumei Li

School of Mathematics and Statistics, Beijing Technology and Business University, Beijing

Received: Aug. 19th, 2021; accepted: Sep. 23rd, 2021; published: Sep. 30th, 2021

Abstract

Principal component analysis is a commonly used technology to simplify data sets, it is also one kind of widely used methods of multivariate statistical analysis. In many courses offered by colleges, the theory of principal component analysis is a key content. But in the teaching process, we

文章引用: 赵霖, 李裕梅. 主成分分析的几个应用理解及 R 语言实践[J]. 数据挖掘, 2021, 11(4): 203-216.
DOI: 10.12677/hjdm.2021.114019

find that there are a lot of things which are worth understanding, exploring and verifying by practice between the theory and the practical applications. This paper reviews the process of principal component analysis, and mainly discusses about its several applications, namely the data dimension reduction based on principal component analysis, comprehensive evaluation based on principal component analysis, determination of the key features based on principal component analysis, samples clustering based on principal component analysis. We detailedly reorganize, partially derivate and complete the theoretical process of each application, organize some usage scenario of applications in literature, give codes of R language practice and the corresponding analysis for each application. The theoretical process and R language practice of these applications are helpful for some researchers to further understand and master the principal component analysis, and also provide important reference for the initial learner to learn and use it.

Keywords

Principle Component Analysis, Dimensionality Reduction, Comprehensive Evaluation, Determination of the Key Features, Samples Clustering, R Language Practice

1. 主成分分析的主要步骤和相关概念

主成分分析，从一个原始数据矩阵 X^0 出发， X^0 是 m 行 n 列的。其中，每一行是一个样本，每一列是一个变量，即共有 m 个样本、n 个变量。

主要步骤如下[1]：

1) 将原始数据做标准化处理，即令： $X_i = \dfrac{X_i^0 - E\left(X_i^0\right)}{\sqrt{D\left(X_i^0\right)}}$ ，得到 X；

2) 建立变量的相关系数阵 R，即 X 的协方差矩阵；

3) 求 R 的特征根为 $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ ，相应的特征向量为 $T_1, T_2, \cdots, T_n$ ；

4) 得到 n 个主成分为 $Y_i = T_i' X'$ ， $i = 1, 2, \cdots, n$ 。

上面步骤涉及到的主要概念如下：

1) 载荷矩阵：就是上面的特征向量矩阵 $T = \left(T_1, T_2, \cdots, T_n\right)$ ，是 n 行 n 列的；

2) 主成分得分： $Y_i = T_i' X'$ ，即样本在第 i 个主成分上的得分，也是数据 X 经过第 i 个特征向量变换后的数据；

3) 因子载荷量： $\rho\left(Y_k, X_i\right) = \dfrac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} t_{ki}$ ，即为主成分 Y_k 与原始变量 X_i 的相关系数；

4) 累计贡献率： $\varphi_p = \sum_{i=1}^{p} \lambda_i \Big/ \sum_{k=1}^{n} \lambda_k$ ， $p = 1, 2, 3, \cdots$ 且 $p \leq n$ ，就是前 p 个特征值的和在所有特征值和里的占比。

2. 基于主成分分析的数据降维

本节从主成分的数据降维理论过程、其应用场景和最后的 R 语言实践三个方面进行阐述。

1) 用主成分进行数据降维的步骤如下：

根据已知 $m\times n$ 的数据 X^0，每一行是一个样本，每一列是一个变量。根据 X^0 进行以下步骤的操作，进行对 X^0 的降维。

上接主成分分析的主要步骤的第(4)步，n 个主成分不一定全部选用，而是根据特征值的累计贡献率来决定选取主成分的个数。后续实施过程见如下两步：

a) 计算特征值的累计贡献率：

$$\lambda_1\Big/\sum_{k=1}^{n}\lambda_k,\left(\lambda_1+\lambda_2\right)\Big/\sum_{k=1}^{n}\lambda_k,\cdots,\sum_{i=1}^{p}\lambda_i\Big/\sum_{k=1}^{n}\lambda_k$$

当累计到某个 i，使得累计贡献率 ≥ 0.85 时，就把这个 i 取出来，为后续降维步骤做准备；

b) 取部分特征值并进行数据变换：

让特征向量 T 里只取前 i 个出来得到 $T_1=\left(T_1,T_2,\cdots,T_i\right)$，用 T_1 对数据 X 做变换，得到 $Y=T_1'X'$，因为 T_1 是 $n\times i$ 的，X 是 $m\times n$ 的，所以彼此转置再乘。这里得到的 Y 是 $i\times m$ 的，再转置一下 Y，得到 $m\times i$ 的，就是变换完的数据、降维后的数据，用于代替原来的 X 去做其他数据分析的用途。

2) 基于主成分降维的应用场景

在主成分分析适用的场合，可以用较少的、互不相关的主成分来代替较多的、有相关性的原始变量。既降低了数据维数，又能保留原数据的大部分信息，从而降低了选取指标的难度，减少了计算的工作量。目前，运用主成分分析对数据进行降维广泛应用于图像压缩、模式识别、食品和医学等各个领域。

在文献[2]中，作者介绍了一种基于 PCA-SVM 的手写数字识别系统模型。运用 PCA 将 784 维的数据降至 25 维，大大减少了冗杂数据对特征提取的影响，并明显的改善了实验时效和设备的运行损耗。

在文献[3]中，作者设计了数据流降维算法 PSPCA。使用滑动窗口机制确定处理数据的范围，合并了标准化和相关系数矩阵的计算步骤。实验结果表明，PSPCA 适用于数据流降维，并能保留合理范围内的原数据信息量，同时保证后续数据挖掘的准确性。

但传统的主成分分析算法存在一定的局限性。比如处理超高维稀疏数据时耗时过长、特征子集的有效性不强等，不少学者提出了相应的改进措施。

在文献[4]中，对于降维前用信息熵做特征筛选的 PCA 改进算法，作者认为其没有考虑特征与类间的关系，导致分类准确率不高。对于用互信息矩阵代替协方差矩阵的 PCA 改进方法，认为其虽然有效提高了降维结果和分类准确率，但特征子集的有效性不好。因此，在引入绝对互信息可信度和相对互信息可信度的基础上，该文给出互信息综合可信度，用其对数据特征进行筛选，最后对筛选出的特征进行 PCA 降维。

在文献[5]中，作者提出一种基于最优 RBF 核主成分的非线性降维重构方法，解决了传统核主成分的核参数选择困难问题，实验结果证明降维结果的准确性优于线性降维方法。

3) 基于主成分降维后预测的应用场景

维度低的数据，在采集时就可以有意选择相关性低的指标。数据维度较高时，想要控制指标间完全不相关或相关性低就较为困难。所以，如果原始数据各个指标不相关，或者相关性很低，利用主成分降维后，会损失较多信息，造成预测精度降低；如果原始数据维度高且指标间有较大相关性，主成分分析可以减少冗杂信息，消除指标间的相关性，提高预测准确度的效果就较好。近年有学者将 PCA 与逐步回归法、BP 神经网络、机器学习等方法结合进行相关预测。

文献[6]中，作者提出了基于主成分分析的组合预测方法。首先对单项模型的预测结果进行 PCA 求出主成分，然后用 AIC 准则确定建模的主成分数量，最后建立沉降实测值与主成分之间的多元回归预测模型，实验结果表明，组合模型预测精度明显优于各单项模型。

文献[7]中，作者利用 PCA 对原始数据进行预处理，建立了基于主成分分析和粒子群支向量机的岩爆预测模型。实验结果表明，此模型的预测准确率高于 SVM 模型和 ANN 模型。

4) 基于主成分降维后的数据进行预测的 R 语言实践

下面根据 UCI 上的一个数据 Waveform，见图 1，进行主成分分析的降维，并在降维前后结合 BP 网络对数据样本进行分类，考察分类的预测效果。

原始数据(波形数据库发生器数据集)一共有 5000 个样本，每个样本有 41 个属性，最后一个属性为样本类别。一共有 3 个类别(0,1,2)。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
2	−0.23	−1.21	1.2	1.23	−0.1	0.12	2.49	1.19	1.34	0.58	1.22	2.3	4.65
3	0.38	0.38	−0.31	−0.09	1.52	1.35	1.49	3.81	2.33	1.34	1.45	3.7	3.08
4	−0.69	1	1.08	1.48	2.44	3.39	3.09	4.08	5.48	3.61	0.47	1.68	2.35
5	0.4	0.68	0.27	1.39	1.03	−0.32	−1.23	−0.5	0.11	0.87	1.27	4.41	3.51
6	−0.81	1.59	−0.69	1.16	4.22	4.98	4.52	2.54	5.6	4.66	4.25	1.58	2.51
7	0.59	0.77	−0.61	1	1.8	2.08	2.16	3.59	4.08	3.63	4.27	4.43	3.45
8	−0.15	0.13	2.27	2.39	4	6.14	5.36	4.08	3.81	3.89	2.46	1.78	−1.43
9	−0.3	−0.42	0.25	−0.61	−1.39	−0.6	1.71	4.01	2.96	5.81	6.56	5.69	5.46
10	−1.45	2.71	3.04	3.21	4.26	5.01	6.24	5.09	3.95	4.84	2.15	−0.3	1.53
11	0.28	0.97	−1.01	−2.34	−1.89	0.54	0.05	2.05	2.38	3.66	3.09	5.12	4.14
12	−1.09	−0.44	1.15	0.17	2.1	3.77	2.4	5.16	5.13	3.66	2.42	2.83	1.02

Figure 1. Part screenshot of UCI dataset Waveform
图 1. UCI 数据集 Waveform 部分截图

1) 将原始数据代入神经网络进行训练，代码如下：

```
wave <- read.csv(file = 'waveform.csv')
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
wave_n <- as.data.frame(lapply(wave[,1:40], normalize))
table(wave$class)  #统计各类别样本量
library(nnet);
n=length(wave[,1]); #样本量
set.seed(1);  #设随机数种子
samp=sample(1:n,n/2);  #随机选择半数观测作为训练集
b=class.ind(wave[,41]);  #生成类别的示性函数
a=nnet(wave_n[samp,],b[samp,],size=8,rang=0.1,decay=5e-4,maxit=200);  #利用训练集中前 18 个变量作为输入变量，隐藏层有 3 个节点，初始随机权值在[-0.1,0.1]，权值是逐渐衰减的。
table(max.col(b[samp,]),max.col(predict(a,wave_n[samp,])))
table(max.col(b[-samp,]),max.col(predict(a,wave_n[-samp,])))
pred<-predict(a,wave_n[samp,])
test1<-table(max.col(b[samp,]),max.col(predict(a,wave_n[samp,])))
test2<-table(max.col(b[-samp,]),max.col(predict(a,wave_n[-samp,])))
cat("训练样本正确率",sum(diag(test1))/sum(test1))
cat("测试样本正确率",sum(diag(test2))/sum(test2))
```

得到测试样本正确率为：0.8244。

2) 对原始数据进行 PCA 降维处理

根据累计方差贡献率 ≥0.85，选择主成分个数为 25 个。代码如下：

```
#然后对 wave_n 做 pca,再然后神经网络训练和测试
library(psych)
library(tidyverse)
d<-wave_n
d.pr=princomp(d,cor=TRUE)  #主成分分析
#screeplot(d.pr,type='lines')  #碎石图
#summary(d.pr,loadings=TRUE) 结果
y=eigen(cor(d))  #特征值和特征向量
y$values
#y$vectors
sum(y$values[1:25])/sum(y$values)  #累计方差贡献率
s=d.pr$scores  #主成分得分
s=s[,1:25]
wave_pca=data.frame(s)
wave_pca$Class <- wave$class
```

3) 将 PCA 降维后的数据代入神经网络进行训练

```
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
wave_pca_n <- as.data.frame(lapply(wave_pca[,1:25], normalize))
table(wave$class)
library(nnet);
n=length(wave[,1]); #样本量
set.seed(1);  #设随机数种子
samp=sample(1:n,n/2);  #随机选择半数观测作为训练集
b=class.ind(wave_pca[,26]);  #生成类别的示性函数 #Find the maximum position for each row of a matrix, breaking ties at random.喂进去的数据必须是个矩阵
a=nnet(wave_pca_n[samp,],b[samp,],size=8,rang=0.1,decay=5e-4,maxit=200);  #利用训练集中前 18 个变量作为输入变量，隐藏层有 3 个节点，初始随机权值在[-0.1,0.1]，权值是逐渐衰减的。
table(max.col(b[samp,]),max.col(predict(a,wave_pca_n[samp,])))
table(max.col(b[-samp,]),max.col(predict(a,wave_pca_n[-samp,])))
pred<-predict(a,wave_pca_n[samp,])
test1<-table(max.col(b[samp,]),max.col(predict(a,wave_pca_n[samp,])))
test2<-table(max.col(b[-samp,]),max.col(predict(a,wave_pca_n[-samp,])))
cat("训练样本正确率",sum(diag(test1))/sum(test1))
cat("测试样本正确率",sum(diag(test2))/sum(test2))
```

得到测试样本的正确率为：0.8296。

由 0.8296 > 0.8244 可知，利用主成分分析对数据进行降维并消除相关性的处理后，预测正确率会得到一定的提升，但提升幅度不大。原始数据变量各个之间相关性较强时，利用主成分降维后，正确率会

有明显提升。

3. 基于主成分分析的综合评价

本节从主成分分析综合评价的理论过程，其使用情况和最后的 R 语言实践三个方面进行阐述。

1) 综合评价的理论过程[1]

在得到主成分分析过程的变换后的数据 Y (各个主成分)和特征值后 $\lambda=\left(\lambda_1,\lambda_2,\cdots,\lambda_n\right)$ 以后，根据如下方法得到评价函数 Z，原来的数据样本 X，每一个观测都会得到一个 Z 值，作为对每个样本的评价得分，根据这个得分看出原来 X 里每个样本的重要性。

设 $Y_1,Y_2,\cdots,Y_n$ 是所求出的 n 个主成分，它们的特征根分别是 $\lambda_1,\lambda_2,\cdots,\lambda_n$。将特征根“归一化”即有

$$w_i=\frac{\lambda_i}{\sum_{k=1}^{n}\lambda_k},i=1,2,\cdots,n$$

记为 $W=\left(w_1,w_2,\cdots,w_n\right)'$，由 $Y=T'X'$，构造综合评价函数为

$$Z=w_1Y_1+w_2Y_2+\cdots+w_nY_n=W'Y=W'T'X'=\left(TW\right)'X'$$

令 $TW=w^*_{n\times1}$，并代入综合评价函数，有：

$$Z=\left(w^*\right)'X'$$

这里我们应该注意，从本质上来说，综合评价函数是对原始指标的线性综合，从计算主成分到对之加权，经过两次线性运算后得到综合评价函数。然后，每个样本经过这个计算得到一个相应的 Z 值，根据 Z 值的大小，可以对样本进行排序和比较，从而实现对样本的优先顺序评价。

2) 基于主成分分析的综合评价使用情况

一般情况下，选择评价指标体系后要对各指标加权来进行综合。指标加权要依据指标的重要性，而对重要性的判断往往带有一定的主观性。而主成分分析能从选定的指标体系中归纳出大部分信息，根据指标间的相对重要性进行客观加权。各主成分的权数为其贡献率，它反映了该主成分包含原始数据的信息量占全部信息量的比重，这样确定权数是客观的、合理的。另外，还有些文献在主成分分析中的综合评价得分基础上，又进行了其他改进，以便更好地评价自己的数据样本。这些文献的使用场景如下：

文献[8]中，通过主成分分析将牡丹的 10 个主要性状降维为 4 个主成分，并以主成分贡献率为权重，建立了高产评价模型，然后依据综合评价得分来划分牡丹等级。

也有许多学者将主成分分析与其他方法结合起来，确定更加合理的评价指标权重。

文献[9]中，作者运用主成分分析对原始指标进行降维，简化了信用评价体系，在确定权重时借助熵权法对评价指标进行客观赋权，求得综合评价函数。

文献[10]中，作者采用层次分析法和熵权法对指标变量进行分层赋权，合理地考虑了指标的重要性差异。

3) 基于主成分分析的综合评价的 R 语言实践

下面这个例子来自文献[1]，不同于原文献的是，我们用 R 语言实现一遍其综合评价过程，代码以供其他研究者们参考和使用。

采用某市工业部门 13 个行业的 8 项重要经济指标的数据，见表 1。8 项经济指标为：

X_1：年末固定资产净值，单位：万元；

X_2：职工人数据，单位：人；

X_3：工业总产值，单位：万元；
X_4：全员劳动生产率，单位：元/人年；
X_5：百元固定资产原值实现产值，单位：元；
X_6：资金利税率，单位：%；
X_7：标准燃料消费量，单位：吨；
X_8：能源利用效果，单位：万元/吨。

Table 1. The original data for comprehensive evaluation based on PCA
表 1. 主成分综合评价原始数据

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
冶金	90,342	52,455	101,091	19,272	82.0	16.1	197,435	0.172
电力	4903	1973	2035	10,313	34.2	7.1	592,077	0.003
煤炭	6735	21,139	3767	1780	36.1	8.2	726,396	0.003
化学	49,454	36,241	81,557	22,504	98.1	25.9	348,226	0.985
机器	139,190	203,505	215,898	10,609	93.2	12.6	139,572	0.628
建材	12,215	16,219	10,351	6382	62.5	8.7	145,818	0.066
森工	2372	6572	8103	12,329	184.4	22.2	20,921	0.152
食品	11,062	23,078	54,935	23,804	370.4	41.0	65,486	0.263
纺织	17,111	23,907	52,108	21,796	221.5	21.5	63,806	0.276
缝纫	1206	3930	6126	15,586	330.4	29.5	1840	0.437
皮革	2150	5704	6200	10,870	184.2	12.0	8913	0.274
造纸	5251	6155	10,383	16,875	146.4	27.5	78,796	0.151
文教	14,341	13,203	19,396	14,691	94.6	17.8	8354	1.574

把数据读入 R 里，记为 d。对 d 作主成分分析，求出特征值和主成分得分。再用特征值求得综合评价指标即主成分的权重，乘以相应的主成分得分，加和得到综合评价得分。代码如下：

```
d<-read.csv('工业指标.csv',sep=',',row.names = 1)
d<-data.frame(d)
d<-scale(x=d)
d.pr=princomp(d,cor=TRUE)  #主成分分析
#summary(d.pr,loadings=TRUE) 结果
y=eigen(cor(d))  #特征值和特征向量
y$values
y$vectors
sum(y$values[1:3])/sum(y$values)  #累计方差贡献率
s=d.pr$scores  #主成分得分
options(digits = 4)
scores=0.0
for(i in 1:8)
  scores = (y$values[i]*s[,i])/(sum(y$values[1:8]))+scores  #计算综合得分
```

```
score=cbind(s,scores)   #综合得分信息
score
rank(-score[,9])
```

得到主成分得分和综合评价得分，见表 2：

Table 2. Principal component scores and comprehensive evaluation scores
表 2. 主成分得分和综合评价得分

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	scores
冶金	1.5348	0.79099	0.55866	0.51158	1.10161	−0.002995	0.411064	0.0046339	1.03168
电力	0.5202	−2.69727	0.23431	0.88774	0.16732	−0.303073	−0.132410	0.0696096	−0.67399
煤炭	1.1017	−3.35680	0.42333	0.60898	−0.96813	0.061749	0.085551	−0.0249946	−0.72599
化学	0.4781	1.23183	−1.04718	1.65945	0.01407	0.078155	−0.009141	−0.0542182	0.64460
机器	4.7117	2..35812	0.49002	−0.78693	−0.51722	0.019927	−0.126042	0.0235048	2.65667
建材	0.3443	−1.84540	0.03855	−0.97712	0.38395	0.214681	−0.028396	−0.0695329	−0.59196
森工	−1.1477	−0.33100	0.29701	−0.71930	0.09491	0.315707	−0.005289	−0.0364443	−0.58156
食品	−2.2863	2.33510	1.13904	0.58521	−0.59622	0.011482	−0.041479	−0.0545409	0.11501
纺织	−0.8764	0.93223	0.36648	0.13476	0.54799	−0.487952	−0.299943	−0.0009293	0.06528
缝纫	−2.1156	0.85811	0.24140	−0.53419	−0.67422	−0.185767	0.290747	0.0756481	−0.55160
皮革	−0.7423	−0.78640	−0.12157	−1.15797	0.24382	−0.397621	0.018487	−0.0307510	−0.67499
造纸	−1.2508	0.03134	0.29873	0.08574	0.38570	0.668550	−0.176206	0.0818775	−0.41128
文教	−0.2718	0.47915	−2.91877	−0.29795	−0.18359	0.007157	0.013057	0.0161372	−0.30189

对上面的综合评价得分进行排序：

```
> rank(-score[,9])
冶金 电力 煤炭 化学 机器 建材 森工 食品 纺织 缝纫 皮革 造纸 文教
   2   11   13    3    1   10    9    4    5    8   12    7    6
```

从综合评价得分可知，机器行业在该地区的综合评价排在第一。另外，从原始数据和前两个主成分得分上也能看出，该行业存在明显的优势。而该地区综合评价排在最后一位的则是煤炭行业，其第二主成分的得分为负数，说明低于平均水平。

4. 基于主成分分析的关键特征确定

本节从利用主成分分析确定关键特征的理论依据，其应用场景和相应的 R 语言实践三个方面进行阐述。

1) 关键特征确定的理论依据

关键特征的确定，主要和主成分得分和因子载荷有关。

原始数据矩阵 X 是 $m\times n$ 的，载荷矩阵 T 是 $n\times n$ 的。X 的转置就是 $n\times m$ 的。

$$X'=\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix},\quad T'=\left(T_1,T_2,\cdots,T_n\right)'=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{12} & t_{12} & \cdots & t_{12} \end{bmatrix}$$

X' 的每一列是一个样本，每一行是一个变量。T' 的每一行是一个特征向量的转置。作主成分变换 $Y=T'X'$ 可得：

$$Y=T'X'=\begin{bmatrix} t_{11}x_{11}+t_{12}x_{12}+\cdots+t_{1n}x_{1n} & t_{11}x_{21}+t_{12}x_{22}+\cdots+t_{1n}x_{2n} & \cdots & t_{11}x_{m1}+t_{12}x_{m2}+\cdots+t_{1n}x_{mn} \\ t_{21}x_{11}+t_{22}x_{12}+\cdots+t_{2n}x_{1n} & t_{21}x_{21}+t_{22}x_{22}+\cdots+t_{2n}x_{2n} & \cdots & t_{21}x_{m1}+t_{22}x_{m2}+\cdots+t_{2n}x_{mn} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1}x_{11}+t_{n2}x_{12}+\cdots+t_{nn}x_{1n} & t_{n1}x_{21}+t_{n2}x_{22}+\cdots+t_{nn}x_{2n} & \cdots & t_{n1}x_{m1}+t_{n2}x_{m2}+\cdots+t_{nn}x_{mn} \end{bmatrix}$$

先看主成分得分：

上式右边的矩阵中，第 i 行、第 j 列代表样本 j 在第 i 个主成分上的得分。比如第 2 行第 1 列中的 $t_{21}x_{11}+t_{22}x_{12}+\cdots+t_{2n}x_{1n}$ 的值比较大，代表的就是样本 1 在 PC2 上的得分较高。

再看因子载荷量：

由因子载荷量公式：$\rho(Y_k,X_i)=\frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}t_{ki}$ 可推出：$t_{ki}=\rho(Y_k,X_i)\frac{\sqrt{\sigma_{ii}}}{\sqrt{\lambda_k}}$。所以：

$$t_{21}x_{11}+t_{22}x_{12}+\cdots+t_{2n}x_{1n}=\rho(Y_2,X_1)\frac{\sqrt{\sigma_{11}}}{\sqrt{\lambda_2}}x_{11}+\rho(Y_2,X_2)\frac{\sqrt{\sigma_{22}}}{\sqrt{\lambda_2}}x_{12}+\cdots+\rho(Y_2,X_n)\frac{\sqrt{\sigma_{nn}}}{\sqrt{\lambda_2}}x_{1n}$$

因为主成分分析前所做的数据标准化，会使数据的每一个变量均值变为 0，方差即 σ_{ii} 变为 1，所以有：

$$t_{21}x_{11}+t_{22}x_{12}+\cdots+t_{2n}x_{1n}=\rho(Y_2,X_1)\frac{1}{\sqrt{\lambda_2}}x_{11}+\rho(Y_2,X_2)\frac{1}{\sqrt{\lambda_2}}x_{12}+\cdots+\rho(Y_2,X_n)\frac{1}{\sqrt{\lambda_2}}x_{1n}$$

其中，x_{11} 是原始变量 X_1 的分量，…，x_{1n} 是原始变量 X_n 的分量。

那么，$\rho(Y_2,X_1)$ 即第二主成分和第一原始变量的因子载荷大，就代表原始变量 X_1 为样本 1 的关键特征。

总的来说，做完主成分变换后，就可以用主成分来解释样品差异。所以可以观察不同样本在第几主成分上的得分高。然后去看这个主成分上因子载荷量大的原始变量有哪些。这些原始变量，就对应样本的关键特征。找到关键特征后，可以根据关键特征对事物进行进一步的研究和改变。

从载荷和得分双重图上，也能直观地看出不同类别的关键特征。

2) 基于主成分分析关键特征确定的应用场景

利用主成分分析对数据进行降维并保留大部分原始信息，易于突出主要特征。因子载荷量(主成分与原始变量的相关系数)较大值对应的原始变量，可以认为是关键特征。下面是一些文献中的应用场景：

文献[11]中，作者对不同香型白酒的 29 种主要风味物质进行主成分分析，提取出 2 个主成分，并画出了主成分载荷和白酒酒样得分双重图，得到了乙酸乙酯、乙酸是清香型白酒的主体风味物质，苯乙醇是芝麻香型白酒的关键风味物质等结论。

文献[12]中，采用 KMO 和 Bartlett 球形度对茎瘤芥的 12 中营养元素间的相关性进行检验，结果表明适合对原始变量进行主成分分析。然后采用 PCA 提取出 2 个主成分，用最大方差法对因子载荷矩阵旋转，构建出主成分得分模型。模型显示出的具有较大载荷值的变量为茎瘤芥的特征元素。

3) 基于主成分分析确定关键特征的 R 语言实践

下面使用文献[11]的数据，见图 2，进行 R 语言实践，从主成分得分及因子载荷量数据、得分和载荷双重图两方面进行实验，并给出一个样本的关键特征结果，以展现本节对于关键特征确定的 R 语言实验

代码效果。

采用 16 种不同种类白酒的 29 种风味物质质量浓度(mg/L)数据。

16 种不同种类白酒为：

浓香型白酒：N_1 ~ N_9；清香型白酒：Q_1, Q_2；酱香型白酒：J_1, J_2, J_3；

芝麻香型白酒：Z_1；特香型白酒：T_1

29 种风味物质为：

酯类物质 9 种：己酸乙酯(F_1)、甲酸异戊酯(F_2)、乳酸乙酯(F_3)、丁酸乙酯(F_4)、乙酸乙酯(F_5)、戊酸乙酯(F_6)、庚酸乙酯(F_9)、辛酸乙酯(F_{11})、棕榈酸乙酯(F_{14})；

醇类物质 9 种：正丙醇(F_{18})、活性戊醇(F_{19})、异丁醇(F_{21})、 2,3-丁二醇(F_{22})、甲醇(F_{23})、正己醇(F_{24})、苯乙醇(F_{27})、（2R,3R）-（-）-2,3-丁二醇(F_{29})、3-呋喃甲醇(F_{31})；

酸类物质 6 种：己酸(F_{33})、丁酸(F_{34})、乙酸(F_{35})、正戊酸(F_{37})、3-甲基戊酸(F_{38})、辛酸(F_{39})；

酮类物质 1 种：3-羟基-2-丁酮(F_{41})；

醛类物质 2 种：乙醛(F_{43})、3-糠醛(F_{45})；

其它类物质 2 种：1 ,1-二乙氧基乙烷(F_{50})、1 ,1-二乙氧基-3-甲基丁烷(F_{51})。

	A	B	C	D	E	F	G	H	I
1	物质和酒型	F1	F2	F3	F4	F5	F6	F9	F11
2	N1	1910.5	448.3	438.1	268.9	214.6	133	68	62.8
3	N2	1647.3	89.8	308.1	326	422	86.7	32.1	38.1
4	N3	1651.5	165.2	510.9	291.2	384.5	56.9	51.6	77.9
5	N4	1904.3	144.7	431.4	277	547.6	85.2	22	10.3
6	N5	2296.5	371.7	535.5	265.4	362.6	200	170.5	113.7
7	N6	1783	289.6	473.8	384.8	440	125.3	82.7	91.5
8	N7	1681.4	142.2	373.6	209	464	59.9	13.4	5.8
9	N8	1422.2	162.6	294	226.9	379.9	45.5	16.6	9.2
10	N9	1261.5	60.3	290.8	156.7	326.6	10.2	8.7	4

Figure 2. Screenshot of the flavour substance sample data section
图 2. 风味物质样本数据部分截图

1) 从主成分得分和因子载荷数据可以进行分析：

```
library(psych)
library(tidyverse)
library(factoextra)
data<-read.csv('白酒.csv',sep=',',row.names=1)
data<-data.frame(data)
data<-scale(x=data)
fa.parallel(data,fa='pc')
p<-principal(data,nfactors=2,rotate="none")
p$values
sum(p$values[1:2])/sum(p$values)
p$loadings    #p$weights
p$scores
```

例如，想要找浓香型白酒 N_1 的关键特征。先看实验得到的主成分得分，见表 3。

Table 3. Principal component scores
表 3. 主成分得分

	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8
PC1	2.1305	−5.5877	−7.0099	−11.0278	7.7043	3.8682	−10.9088	−11.3459
PC2	9.6535	4.7250	5.9332	0.2416	17.0173	16.2913	−3.4110	−1.9647
	N_9	Q_1	Q_2	J_1	J_2	J_3	Z_1	T_1
PC1	−14.0962	−8.9137	−7.6846	18.7133	6.6823	28.5692	9.6397	−0.7327
PC2	−2.9394	−12.1828	−12.4465	−5.4143	−0.7327	−10.0213	−5.8014	1.0521

可以看出，浓香型白酒 N_1 的 PC2 得分较高，所以去看 PC2 上的因子载荷量，见表 4。

Table 4. Factor loadings
表 4. 因子载荷量

	F_1	F_2	F_3	F_4	F_5	F_6	F_9	F_{11}	F_{14}	F_{18}
PC1	−0.333	0.701	0.832	−	0.523	0.315	0.148	0.176	0. 612	0.904
PC2	0.847	0.118	0.114	0.836	−0.674	0.780	0.871	0.883	0.413	−0.269
	F_{19}	F_{21}	F_{22}	F_{23}	F_{24}	F_{27}	F_{29}	F_{31}	F_{33}	F_{34}
PC1	0.730	0.651	0.670	−0.726	0.215	0.628	0.899	0.908	−0.344	−
PC2	−0.257	0.158	−0.425	−0.447	0.867	−0.370	−0.339	−0.100	0.837	0.754
	F_{35}	F_{37}	F_{38}	F_{39}	F_{41}	F_{43}	F_{45}	F_{50}	F_{51}	
PC1	0.627	0.354	0.906	0.268	0.828	0.937	0.849	0.910	0.681	
PC2	−0.627	0.741	0.174	0.770	−0.166	0.177	−0.103	0.277	0.429	

由数据可知，$F_1, F_4, F_9, F_{11}, F_{24}, F_{33}$ 等物质在 PC2 上的因子载荷量较大，是浓香型白酒 N_1 的关键特征。

2) 对原始数据作主成分分析，从因子载荷量和主成分得分图可以直观看出关键特征。

```
data<-read.csv('白酒.csv',sep=',',row.names=1)
res.pca=prcomp(data,scale=TRUE)
fviz_pca_var(res.pca, col.var = "black")   #因子载荷图
fviz_pca_biplot(res.pca,
                palette="joo",
                mean.point=F,
                gradient.cols = "RdYlBu",
                ggtheme = theme_minimal())   #双重图
```

得到因子载荷图及因子载荷和酒样得分双重图，见图 3、图 4。

从图 3 中可以看到各原始变量对 PC1 和 PC2 的影响是正还是负。例如，处在第四象限，表明它对 PC1 有很大的正向影响，对 PC2 有较大的负向影响。

图 4 表明：己酸乙酯和己酸等物质对 PC2 有较大的正向影响。这类物质与的距离较近，是浓香型白酒的主体风味物质。这与主成分得分和因子载荷数据分析出的结果相同。

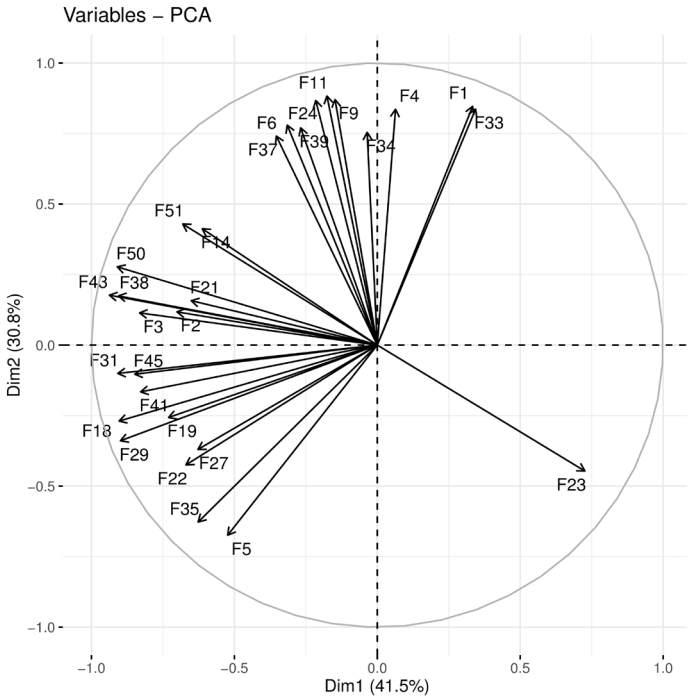

Figure 3. The diagram of factor loadings
图 3. 因子载荷图

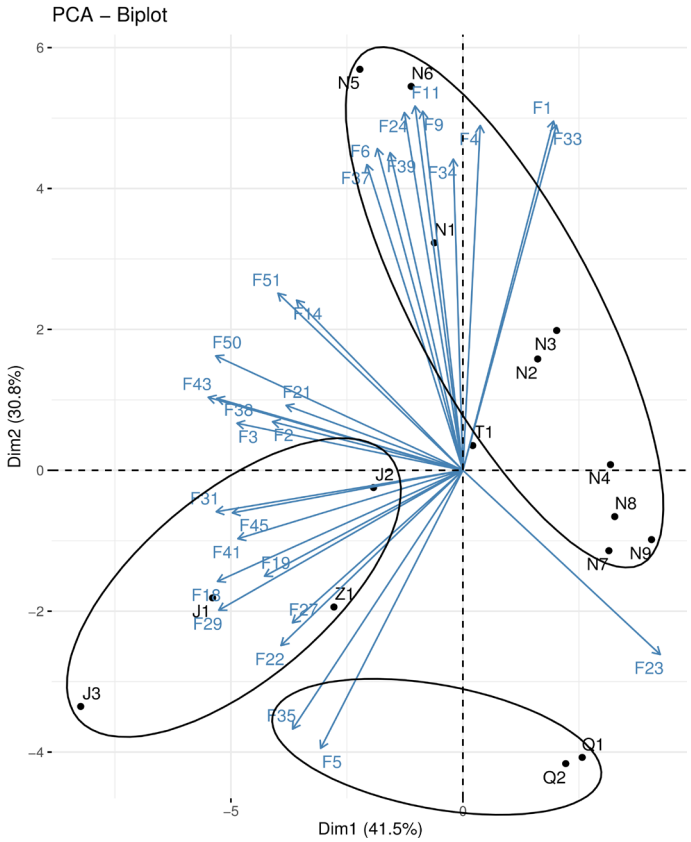

Figure 4. The overlay graph of factor loadings and principal component scores
图 4. 因子载荷和主成分得分双重图

5. 基于主成分分析的样本聚类

本节从主成分分析的样本聚类原理，其使用场景和利用主成分分析对酒样进行聚类的 R 语言实践三个方面进行阐述。

1) 基于主成分分析对样本进行聚类的原理

由于主成分包含原始数据的大部分信息，所以可以用主成分来解释样本之间的差异。样本在主成分得分图上的坐标就是它的主成分得分，同类样本的各主成分得分相似，在图上的距离就会相近，同类样本自然可以聚在一起。

2) 利用主成分分析进行样本聚类的使用场景

对多变量数据，可以选取前两个主成分，根据主成分得分画出样品在二维平面上的分布情况，从而直观地看出样品的分类情况。也可以根据关键特征有效地区分不同类别。但其分类效果不如聚类分析等方法，所以一般情况下会结合其他方法进行类别区分。

文献[11]中，作者结合聚类分析和主成分分析对白酒的风味物质进行研究。实验结果表明，主成分分析对白酒样品香型的分类效果不如聚类分析，但可以反映白酒香型与风味物质之间的关系。

文献[12]中，由 PCA 构建的主成分得分模型，可以有效区分 3 个产地的茎瘤芥。由聚类分析可根据茎瘤芥中营养元素的含量差异进行产地溯源。

文献[13]中，作者对主成分分析中的正交旋转和斜交旋转做了对照，结果证明，斜交旋转比正交旋转划分的天气形势更具代表性。

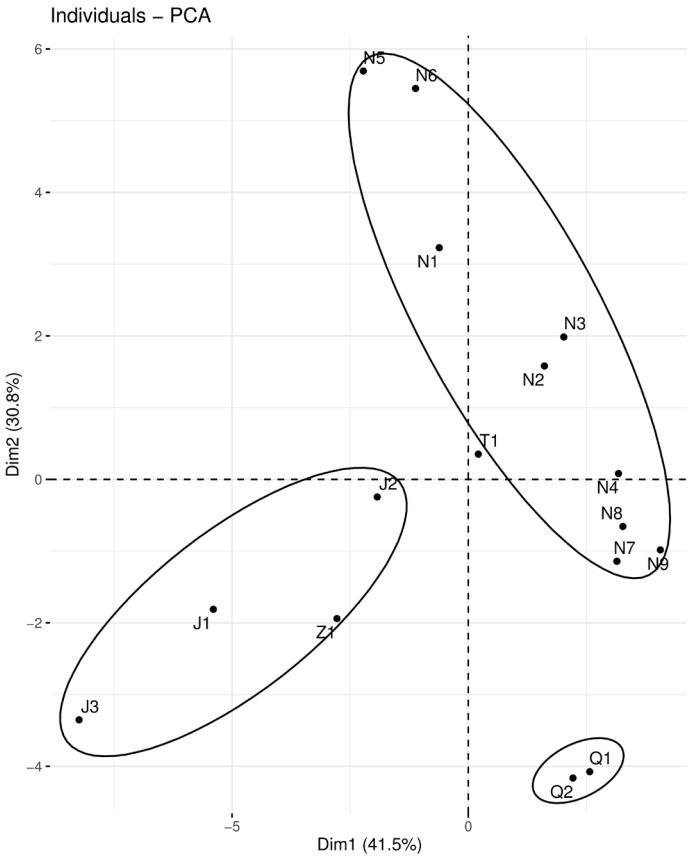

Figure 5. The diagram of the principal component scores of liquor samples
图 5. 各白酒样本的主成分得分图

3) 基于主成分得分对样本进行聚类的 R 语言实验

与“基于主成分分析确定关键特征的 R 语言实验”数据相同，对其作主成分分析，可以画出主成分得分图。

```
data<-read.csv('白酒.csv',sep=',',row.names=1)
res.pca=prcomp(data,scale=TRUE)
#res.pca$x[,1:2] 看各样本在第一、第二主成分上的得分
fviz_pca_ind(res.pca, repel=T)     #个体图
```

主成分得分图上，各样本的坐标就是其在第一、第二主成分上的得分。

得到主成分得分图，见图 5。

从图中可以看出，清香型白酒 Q_1, Q_2 分布集中；特香型白酒 T_1 靠近浓香型白酒 $N_1 \sim N_9$；酱香型白酒 J_1, J_2, J_3 分布较为分散，但都分布在第 4 象限；芝麻香型白酒 Z_1 与酱香型白酒距离较近且位于第 4 象限，说明它们的风味组分构成相似。

6. 总结

本文对主成分分析的主要过程和相关概念进行了概述，对主成分分析在降维、综合评价、关键特征确定和样本聚类方面的应用进行了理论推导，还给出了在 R 语言上实现这些应用的实例。这些工作把主成分分析的理论和实际应用联系起来，展示了主成分分析的优点：可以减少指标选择的工作量、简化计算、消除数据的相关性和冗杂信息、突出关键特征等。我们对这些过程进行详细论述并给出相关的 R 语言实践代码及相应分析等，有助于主成分分析的广大学习者和使用者更好地理解和应用这些方法。

参考文献

[1] 朱建平. 应用多元统计分析[M]. 北京: 科学出版社, 2016.

[2] 杨济萍. 基于主成分降维模型的手写数字识别研究[J]. 网络安全技术与应用, 2021(3): 31-32. https://doi.org/10.3969/j.issn.1009-6833.2021.03.017

[3] 单燕, 李玲娟, 孙杜靖. 基于主成分分析的并行化数据流降维算法研究[J]. 南京邮电大学学报(自然科学版), 2015, 35(5): 99-104. https://doi.org/10.14132/j.cnki.1673-5439.2015.05.014

[4] 张素智, 陈小妮. 基于互信息可信度的主成分分析数据降维[J]. 湖北民族学院学报(自然科学版), 2019, 37(4): 425-430. https://doi.org/10.13501/j.cnki.42-1569/n.2019.12.014

[5] 李丹, 杨保华, 张远航, 缪书唯, 王奇. 基于最优 RBF 核主成分的空间多维风电功率降维及重构[J]. 电网技术, 2020, 44(12): 4539-4546. https://doi.org/10.13335/j.1000-3673.pst.2019.2626

[6] 于永堂, 郑建国, 黄鑫. 基于主成分分析的黄土高填方工后沉降组合预测方法[J]. 水利与建筑工程学报, 2021, 19(3): 117-123. https://doi.org/10.3969/j.issn.1672-1144.2021.03.019

[7] 刘晓悦, 张雪梅, 杨伟. 基于 PCA-SVM 的岩爆预测[J]. 中国矿业, 2021, 30(7): 176-180. https://doi.org/10.12075/j.issn.1004-4051.2021.07.005

[8] 张伟龙, 杨静慧, 宋科, 蒋鑫, 郜伟, 张超, 冯国华. 基于主成分分析油用牡丹“凤丹”主要性状评价及选优[J]. 天津农学院学报, 2021, 28(2): 12-17. https://doi.org/10.19640/j.cnki.jtau.2021.02.003

[9] 江晓欣. 基于主成分分析的建筑施工企业信用评价研究[J]. 山西建筑, 2021, 47(15): 189-192. https://doi.org/10.13719/j.cnki.1009-6825.2021.15.072

[10] 李炳军, 张一帆, 张淑华. 主成分分析法的改进及其在河南区域经济发展评价中的应用[J/OL]. 河南农业大学学报: 1-14[2021-08-15]. https://doi.org/10.16445/j.cnki.1000-2340.20210618.001

[11] 钱冲, 廖永红, 刘明艳, 徐瑾, 刘丽, 于莉. 不同香型白酒的聚类分析和主成分分析[J]. 中国食品学报, 2017, 17(2): 243-255. https://doi.org/10.16429/j.1009-7848.2017.02.032

[12] 江波, 黄建华. 基于营养元素的茎瘤芥主成分分析和产地溯源[J/OL]. 食品与发酵工业: 1-8[2021-08-15]. https://doi.org/10.13995/j.cnki.11-1802/ts.027925

[13] 侯雪伟, 吕鑫, 魏蕾. 正交与斜交旋转主成分分析法在气象因子影响细颗粒物研究中的应用[J]. 环境科学学报, 2021, 41(7): 2598-2606. https://doi.org/10.13671/j.hjkxxb.2020.0552