实验 5 主成分分析（PCA）

（4 学时）

一、实验背景

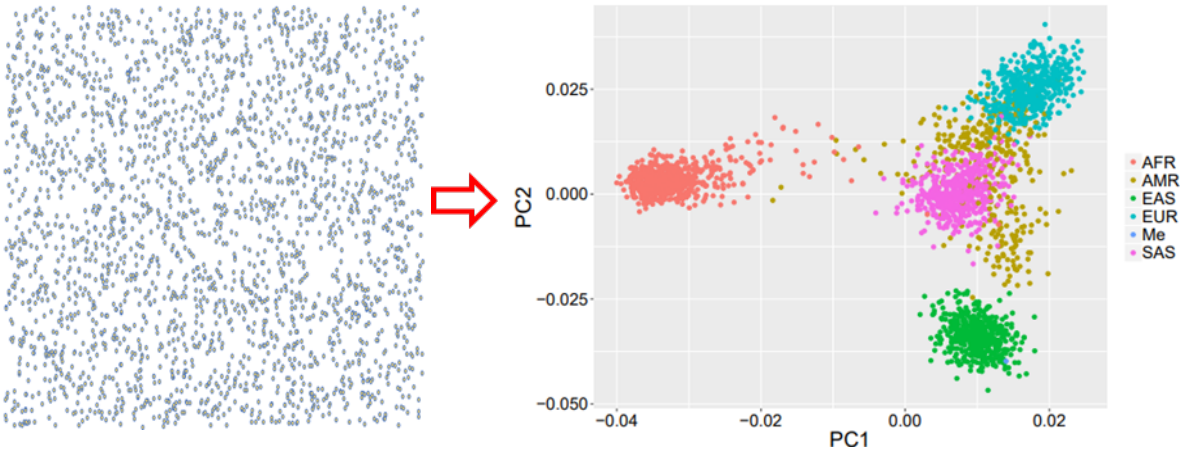
主成分分析(Principal Components Analysis，简称 PCA)，也称为主分量分析，是统计学当中的一种简化数据集的技术，通过降维的方法把多个指标转化为少数几个综合指标，是一种数据降维方法。当数据集非常庞大，包含变量数量很多时，可使用 PCA 方法提取其中数量较少但能代表数据集主要特征的变量，达到降维目的。PCA 方法非常实用，在文本挖掘、图像分析、生物数据分析、客户偏好分析等领域应用广泛。

二、教学目标

本次实验，我们将追溯个体的遗传分类。结合千人基因组数据库，将实验 2.3 获取的个体特征数据与来自全球其他地区的 2504 个样本变异位点信息进行混合，通过主成分分析方法，对数据进行降维，最终实现样本分组。

分析工具：gcta、LDAK

推荐语言：R



三、实验原理

什么是降维？

以下，我们通过一个简单的例子展示对多个变量构成的数据集进行降维的思路：

如何找到有代表性的特征，快速对不同西瓜进行区分？

	西瓜A	西瓜B	西瓜C	西瓜D	西瓜E
重量	1	2	3	4	5
颜色	1	2	2	4	6
形状	8	9	8	8	9
纹路	3	3	3	9	9
气味	1	2	1	1	1

在以上这组数据集当中，西瓜颜色、形状、纹路、气味等都已经做了数值化处理，如颜色值越小表明西瓜越接近浅黄色，颜色值越大表明西瓜越接近深绿色；形状值越接近 8 表明西瓜越接近球形等。

显然，重量和颜色在不同西瓜之间有显著差异，适合用来对不同的西瓜进行描述和区分；相比之下，形状和气味在不同西瓜之间差异不大，分析比较时，可以丢弃不考虑。

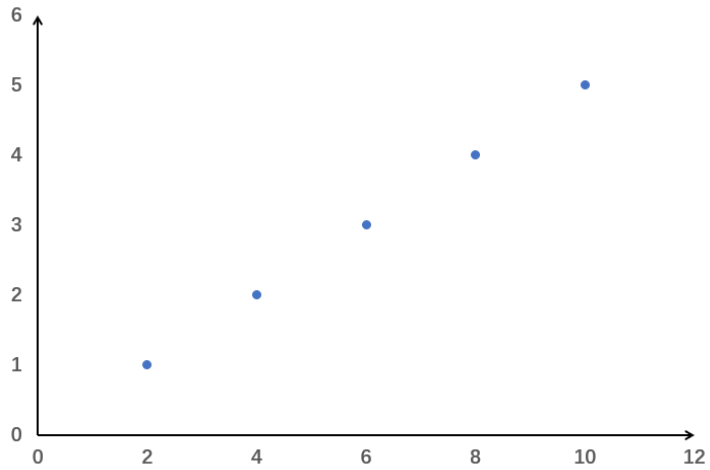
我们对这个例子进行简单总结：在对数据集进行信息提取时，不同样本间有显著差异的特征才是我们需要关注的。选择显著差异特征，舍弃无差异特征，这样的数据处理过程，就是降维的核心思路。

为什么要应用 PCA 方法对数据进行降维？

一方面，当数据集体量过大，或其中包含的变量数太多时，数据采集、建模和分析的复杂度都将加大；另一方面，当变量数太多时，一些变量相互之间可能存在相关关系，使得我们无法一目了然地对不同的数据特征进行取舍，需要借助 PCA 这一类专业的降维方法，实现简化计算过程、节约计算成本的目的。

PCA 分析到底对数据做了什么处理？

我们从简单的二维数据集出发



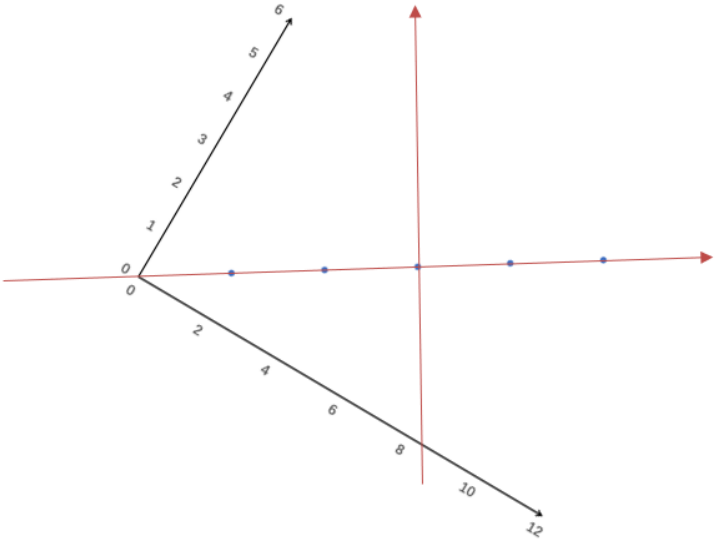
有这样一列数据点:

(x1=2,y1=1）,(x2=4,y2=2),(x3=6,y3=3),(x4=8,y4=4),(x5=10,y5=5)

在二维坐标系中，我们需要保存 x 和 y 这 2 个维度的信息。

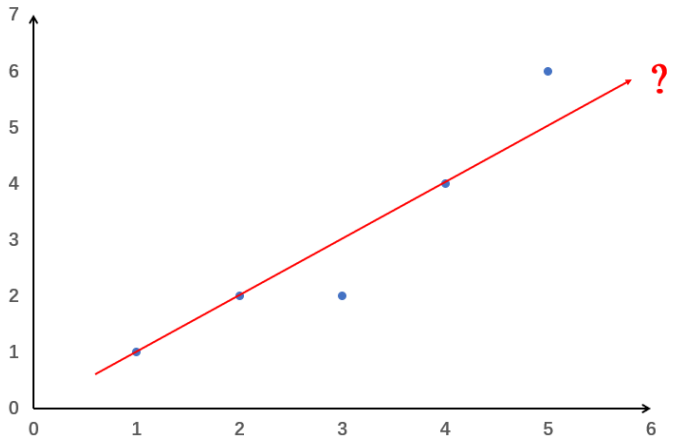
但如果我们旋转坐标轴，就会发现，只用一个维度就能把这些点的相对位置信息保存下来。

PCA 分析核心就是要找到一个新的坐标系，使得这个数据集能只保留一个维度，同时数据信息损失最小。构建新坐标系时，我们将原数据点的重心作为新坐标系原点，新的坐标轴沿着数据走向分布，这就是第一主成分的方向，穿过原点与之垂直的坐标轴就是第二主成分。只要记住这个原点的位置以及旋转的角度，就能快速地将两个坐标系对应起来。



接下来，我们还是以西瓜的数据集为例，重量和颜色这 2 个特征构成的二维数据集如下：

（1，1），（2，2），（3，2），（4，4），（5，6）



这时候，数据并不是严格落在一条直线上，换句话讲，我们不能像理想数据集那样，快速地找到一条穿过数据重心的线，直接形成新的坐标系。该如何寻找新坐标系的原点和数据分布方向呢？

这就是处理复杂数据集时 PCA 方法需要处理的核心问题：我们需要找到一个新的坐标轴，使得我们既能提取出数据集的主要特征，又不至于损失掉太多的数据信息。

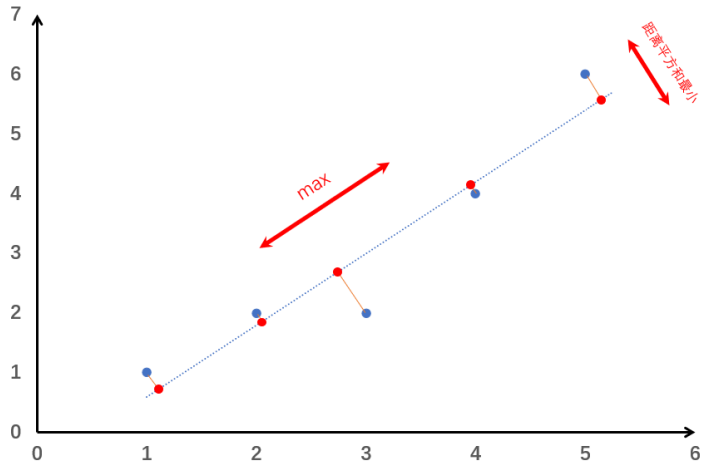
在 PCA 方法中，最佳坐标系有两个核心评价指标：

1） 原数据点在新坐标轴上的投影点分布的距离最大

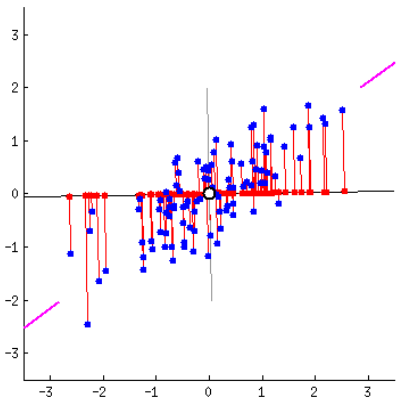
此时，原数据点在新坐标轴上的投影点最分散，即数据差异度最大的方向被保留在第一主成分中；考虑一种相反的极端情况：当原数据点在新坐标轴上的投影点聚集为一个点的时候，保留的数据信息最少。

2） 投影点与原数据点之间距离的平方和最小

此时，第一主成分相比于原数据集信息丢失最小。



记住这两个核心原则，我们来看一个复杂数据集的主成分坐标轴构建过程：



总结 PCA 方法的核心，一是寻找数据重心，作为构建新坐标系的原点；二

是寻找新的坐标轴方向，保证数据点在此方向上的投影点分布范围最大，其次投影点离原数据点距离平方和最小。

PCA 算法与工具

接下来，我们将会讨论 PCA 分析的核心数据计算过程。

在以上分析思路的整理过程当中，我们反复强调，求解主成分的 2 大要点：

1） 将数据重心作为新的坐标系原点

2） 计算新坐标系相对于原坐标系的旋转角度

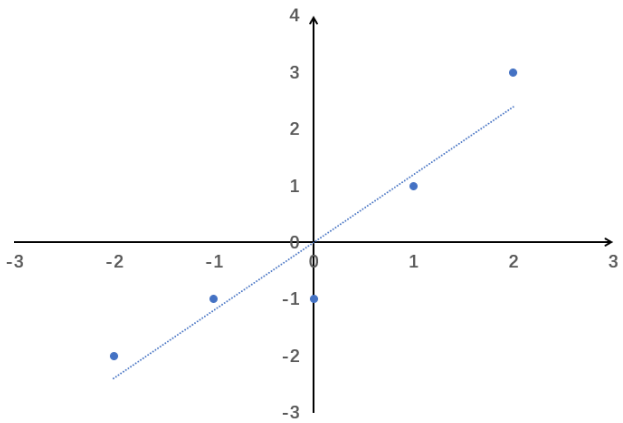
计算数据重心比较简单，通常采用归一化减均值方法。以二维数据点集 p(x,y) 为例，先计算其平均值 p (x,y),然后各数据点坐标值与均值作差，得到新的数据分布点阵就是以数据重心为原点分布的。

西瓜的重量、颜色二维数据集：

（1，1），（2，2），（3，2），（4，4），（5，6）

均值点为（3，3），各数据点坐标值减去均值后得到新的坐标点为：

（-2，-2），（-1，-1），（0，-1），（1，1），（2，3）



接下来，我们需要计算获得能代表第一主成分的新坐标轴要旋转的角度。

这里，我们采用矩阵来计算图上的点阵变换过程：

假定任意一点 p 经过线性变换（拉伸或旋转）得到的新数据点为 q，我们记 q=Ap，这里 A 为变换矩阵。

为了计算得到变换矩阵，我们引入两个新的矩阵：

1）尺度矩阵：S=$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$

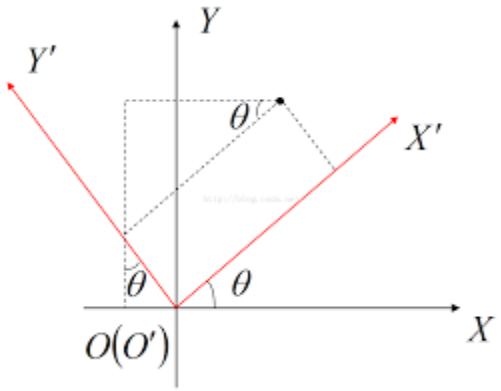
我们如果记原数据集矩阵为 D，则 D = $\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ y_1 & y_2 & y_3 & y_4 & y_5 \end{bmatrix}$

$$SD = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ y_1 & y_2 & y_3 & y_4 & y_5 \end{bmatrix} = \begin{bmatrix} ax_1 & ax_2 & ax_3 & ax_4 & ax_5 \\ by_1 & by_2 & by_3 & by_4 & by_5 \end{bmatrix}$$

原坐标系在 x 和 y 方向上分别经过了 a 倍和 b 倍的拉伸（压缩）。

我们只需要记住，这里 S 矩阵的作用，是对坐标系进行拉伸或压缩变换。

2）旋转矩阵：$R=\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$，其中 θ 是新坐标系相对于原坐标系逆时针旋转的角度。以下是 R 矩阵的数值解释：



经过 θ 角度逆时针旋转后，原坐标系 x 轴上点（1，0）的新坐标为（$\cos(\theta)$，$\sin(\theta)$）；原坐标系 y 轴上点（0，1）新坐标为（$-\sin(\theta)$，$\cos(\theta)$）。

介绍完这两个矩阵，我们回到待求解的变换矩阵 A，不难得出：

$$\text{变换矩阵A} = RS = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

换句话讲，只要我们能求得 R、S 矩阵，就能找到第一主成分的坐标方向。

怎么求 R？

协方差矩阵的特征向量就是 R，协方差用于衡量变量两两之间的相关性，其计算公式如下：

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

第一主成分的协方差矩阵由代表不同数据点的向量两两之间的协方差组成，求解该矩阵的特征向量，即是坐标轴旋转矩阵 R；求解其特征值，该协方差矩阵特征值 λ 与所有主成分协方差矩阵特征值总和的比值可用于衡量第一主成分与原数据集的一致程度。

以下是我们简单回顾 PCA 分析的关键步骤：

3） 归一化减均值；

西瓜的例子

4） 求协方差矩阵的特征向量；

几何处理

5） 特征值排序，从大到小，取大舍小；

某个主成分的方差占全部方差的比重也称为贡献率，在对特征值进行取舍时，一般要求累计贡献率达到 80%以上

〖**教学重点**〗在本实验环节中，学生需掌握 PCA 降维原理和实际的分析操作。

四、数据文件解释

1)ALL.218snpFrom1000genomes.vcf/ALL.218snpFrom1000genomes-id-maf0.05.vcf：千人基因 218 个位点信息/过滤后的千人基因 218 个位点信息

2) merged.vcf：个人位点信息

3) mergy.py：将个人位点信息与 218 个位点信息进行合并

4) integrated_call_samples_v3.20130502.ALL.panel：各个 sample 对应的性别、地区信息

5) add_race.py：将性别、地区信息添加到分析文件中

6) pca_vitualize.R：将分析得出的 eigenvector 在 R 中用 ggplot 画出 pca 分析图

7) pca.sh：综合以上所有步骤，在 linux/unix 上统一运行

mkdir ./pca

参数解释：mkdir：make a directory

/Users/kirelin/Desktop/plink_mac_20210606/plink --vcf ${sample_name}.merged.vcf --make-bed --allow-extra-chr --threads 4 --vcf-half-call haploid --id-delim . --double-id --out ./pca/${sample_name}.merged.vcf.plink

参数解释：--make-bed： creates a new PLINK 1 binary fileset, after applying sample/variant filters and other operations below

--allow-extra-chr：allow extra characters

五、实验步骤

第一部分 前期准备：给标记加上 ID

进入个人文件夹

$ cd ~

新建 PCA 所需文件夹

$ mkdir 05_PCA

进入 05_PCA 文件夹并从/home/复制所需脚本压缩包

$ cd 05_PCA

$ cp /home/pca_file.zip ./

解压缩压缩文件

$ unzip pca_file.zip

拷贝文件到 PCA 目录下

拷贝 vcf 文件到 pca 文件夹

$ cp ../03_SNPCalling/test.clean.SNP.vcf ./

查看脚本文件及 vcf 文件

$ ls

第二部分 合并文件及格式转换

#合并 vcf 文件

python ./merge.py ./test.clean.SNP.vcf test >merged.vcf

#创建文件夹以存放 PLINK 数据格式转化的数据

$ mkdir pca

#将 vcf 文件转换为 bed 格式文件

$ /home/ecoli/2024fuda/05_PCA/plink --vcf ./merged.vcf --make-bed --allow-extra-chr --threads 4 --vcf-half-call haploid --id-delim . --double-id --out ./pca/merged.vcf.plink

#查看 plink 的数据转换结果

ll ./pca

#结果可视化

#生成 PCA 文件

/home/ecoli/2024fuda/05_PCA/plink --bfile ./pca/merged.vcf.plink --pca 5 --out ./pca/merged.vcf.plink.pca --threads 4

#将性别、地区信息添加到分析文件中

python
add_race.py ./pca/merged.vcf.plink.pca.eigenvec > ./pca/merged.vcf.plink.pca.eigenvec.add_race

#使用 R 进行可视化

Rscript pca_vitualize.R ./pca/merged.vcf.plink.pca.eigenvec.add_race ./pca/test_pca.pdf

#将性别、国家信息添加到分析文件中

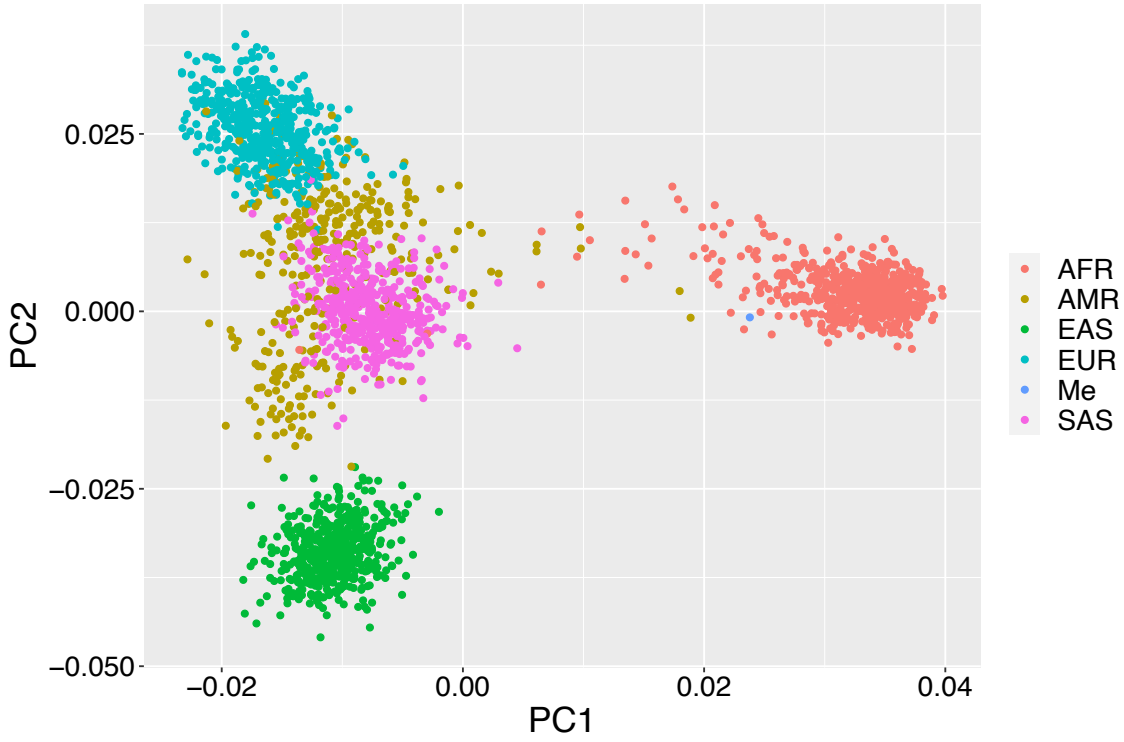
python
add_country.py ./pca/merged.vcf.plink.pca.eigenvec > ./pca/merged.vcf.plink.pca.eigenvec.add_country

#使用 R 进行可视化

Rscript pca_vitualize.R ./pca/merged.vcf.plink.pca.eigenvec.add_race ./pca/country-1.pdf

六、预期实验结果

最终可得到 pca 分析图：6.pca2.pdf



图中右边蓝色点 Me 则为个人在千人基因信息中的分布位置