基因组学-高通量测序实验与数据分析

PCA降维分析

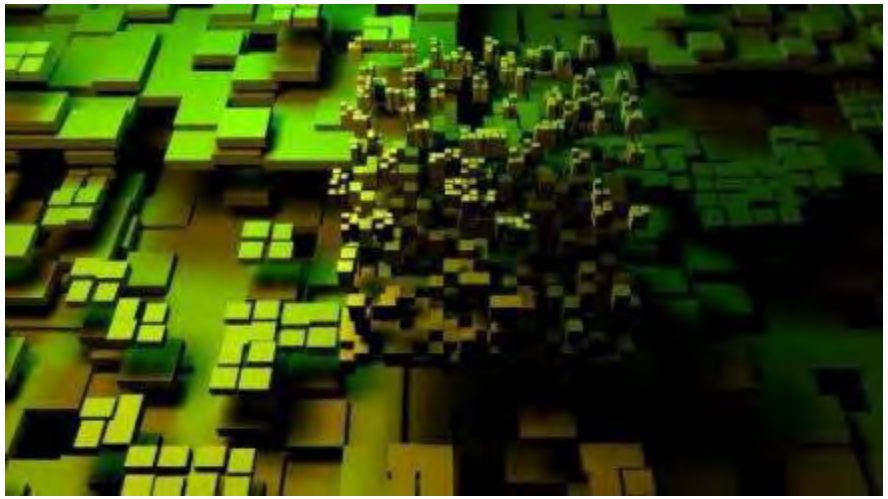
纲要

01

高维数据的降维分析

1.1 高维数据的特点

- 维度（dimensionality）：
  - 又称为维数，是数学中独立参数的数目
  - 在一定的前提下描述一个数学对象所需的参数个数
  - 完整表述应为“对象X基于前提A是n维”



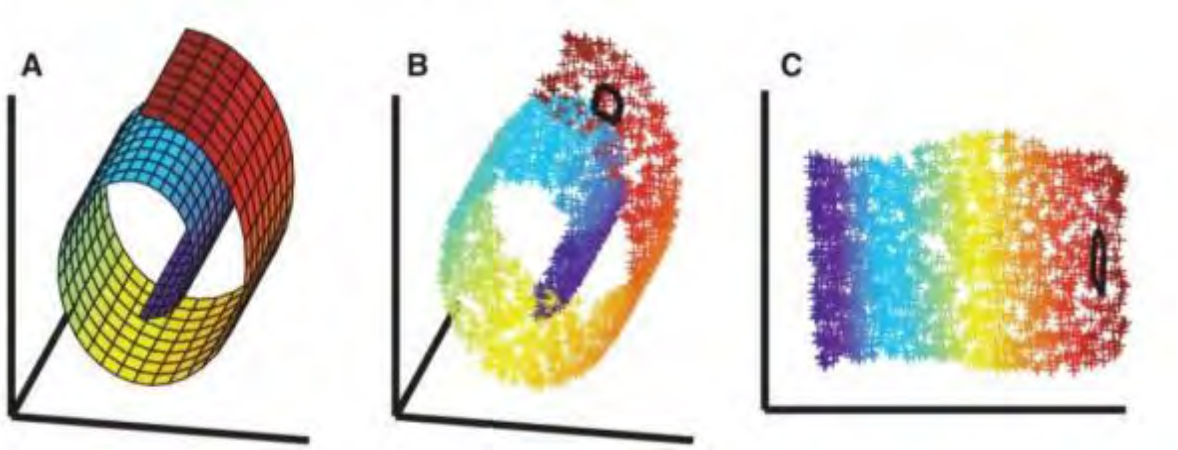
- 高维数据
  - 测序数据是一种高维数据
  - 每一个基因都是数据集的一个特征
  - 测序数据的维数考虑基因及样本数目
- 维数灾难(Curse of Dimensionality)
  - 通常是指在涉及到向量的计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象
  - 维数灾难涉及数字分析、抽样、组合、机器学习、数据挖掘和数据库等诸多领域。

1.1 高维数据的特点

- 如何理解高维空间?

➢ 什么是降维（Dimensionality reduction）

- ●指采用某种映射方法，将原高维空间中的数据点映射到低维度的空间中
- ●降维的本质是学习一个映射函数f:x→y， y是数据点映射后的低维向量表达，通常y的维度小于x的维度
- ● f 可能是显式的或隐式的、线性的或非线性的。



降维的目的

- 是为了更好地适应模型
- 避免维度灾难的问题

降维的意义：克服维数灾难， 获取本质特征， 节省存储空间， 去除无用噪声， 实现数据可视化

降维是为聚类等下游分析服务的

➢ 聚类分析 (Clustering Analysis)

- 指将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程
- 聚类分析步骤：

1、确定距离或相似性度量：这定义了数据点之间的相似性或距离程度

“距离”：欧氏距离、欧式距离的平方、曼哈顿距离、切比雪夫距离等

相似程度:“相关系数”主要是皮尔逊相关系数

2、初始化簇：选择初始簇中心或分配点到初始簇

3、迭代分配：使用距离或相似性度量，将每个数据点分配到与其最相似的簇中心

4、更新簇中心：重新计算每个簇的中心点，表示簇中数据点的平均位置

5、重复步骤 3 和 4：直到簇中心不再变化或达到预定义的条件（如迭代次数或误差阈值）

➢ 聚类分析 (Clustering Analysis)

- 聚类算法类型

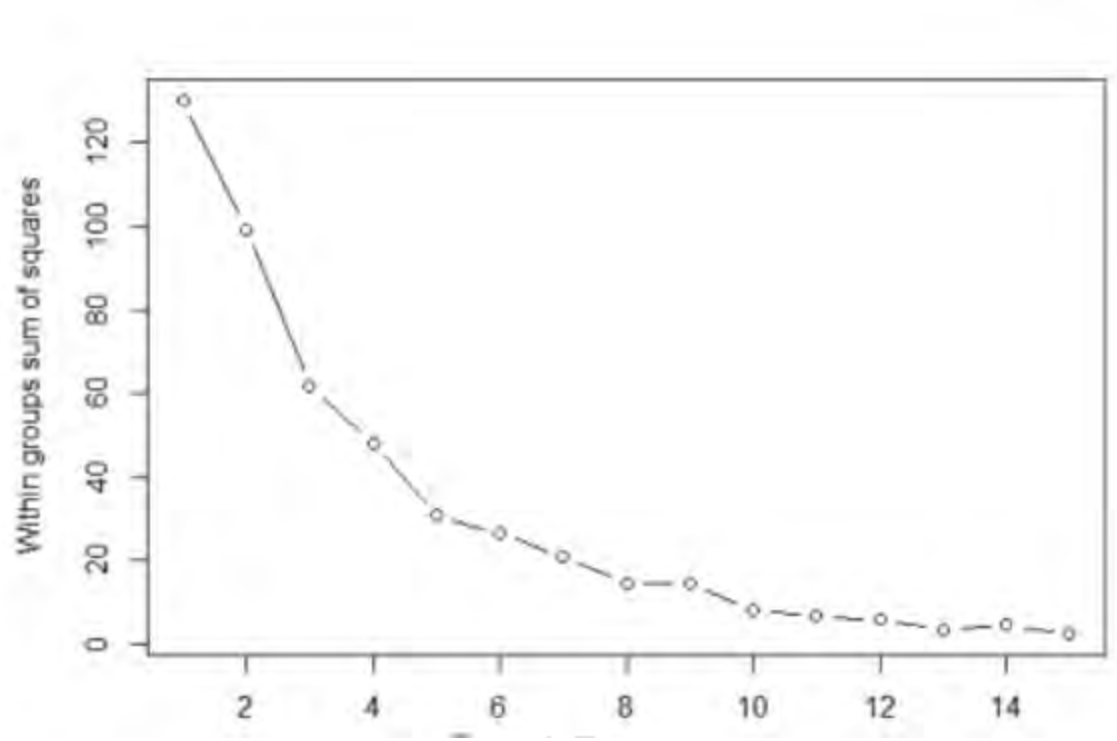
1、k 均值(k-means)聚类：将数据点分配到 k 个预定义的簇。

K-means聚类是最常用的聚类方法，其基本原理是先随机选取K个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。一旦全部对象都被分配了，每个聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是以下任何一个:

1、没有（或最小数目）对象被重新分配给不同的聚类。

2、没有（或最小数目）聚类中心再发生变化。

3、误差平方和局部最小。



从图中可以看出，聚类数在8-15之间时，簇内间距变化最小，故选择聚类数K=8

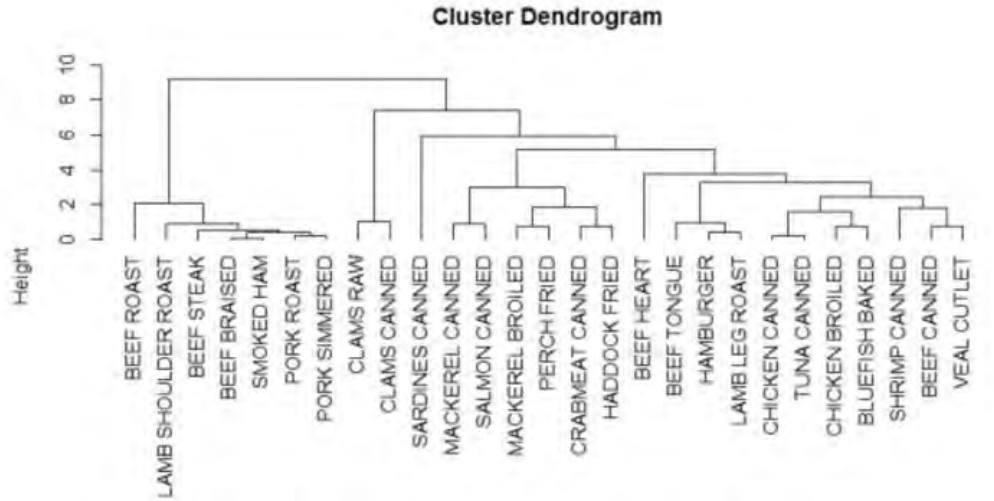
聚类中心一级被分配到的对象就代表一个聚类，或一个簇。该方法计算简单，易于实现，计算速度快，适用于连续型数值数据、大规模数据和高维数据；同时对初始值敏感，不同的初始值可能导致不同的结果，只能形成球形的簇。

➢ 聚类分析 (Clustering Analysis)

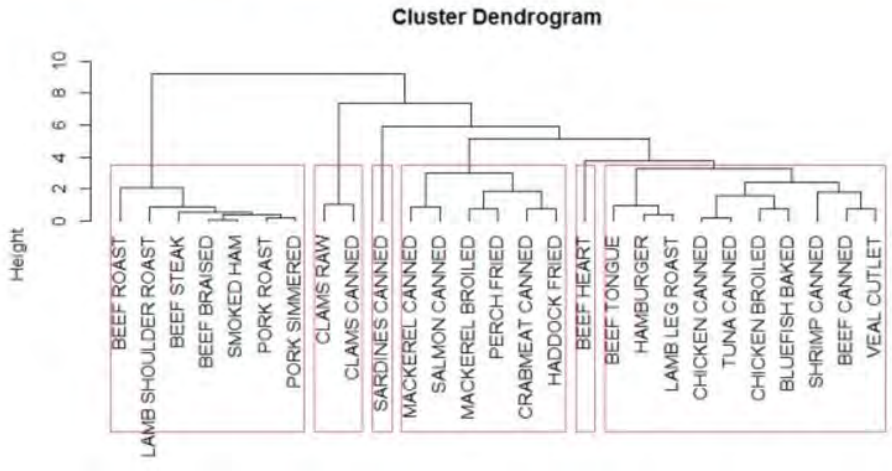
- 聚类算法类型

2、层次(Hierarchical Clustering)聚类：在层次结构中生成簇，其中子簇嵌套在更大的簇中.

层次聚类的原理是将数据集构建成一个层次结构，其中每个样本最初表示为一个单独的簇，然后通过计算样本之间的相似度或距离来逐渐将簇合并成更大的簇。整个过程可以表示为一棵树形结构，称为聚类树或者树状图。通过该树状图，我们可以选择合适的切割点来确定最终的聚类结果。层次聚类的好处是不需要指定具体类别数目，在聚类完成之后，可选择任意层次得到相应数目的簇。适用于小规模数据集和低维数据。但是该方法的计算较为复杂，异常值会对聚类结果产生很大影响，并且可能聚类成链状。



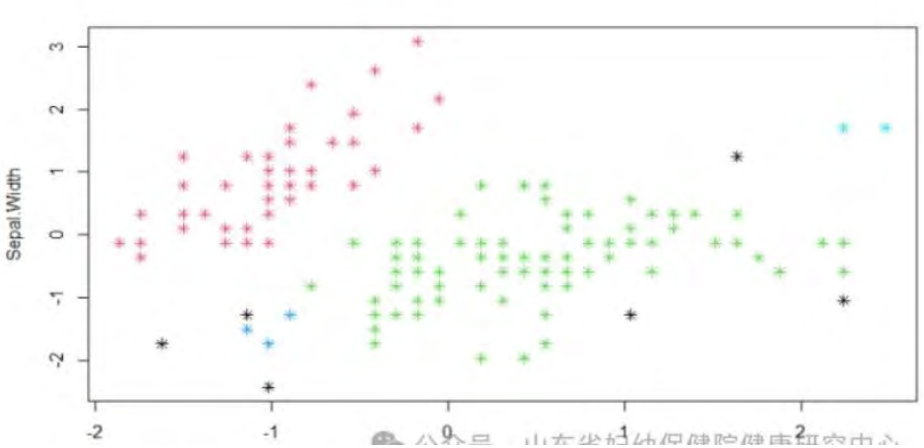

```
rect.hclust(hc, k = 6) ## 设置聚类数为6
```

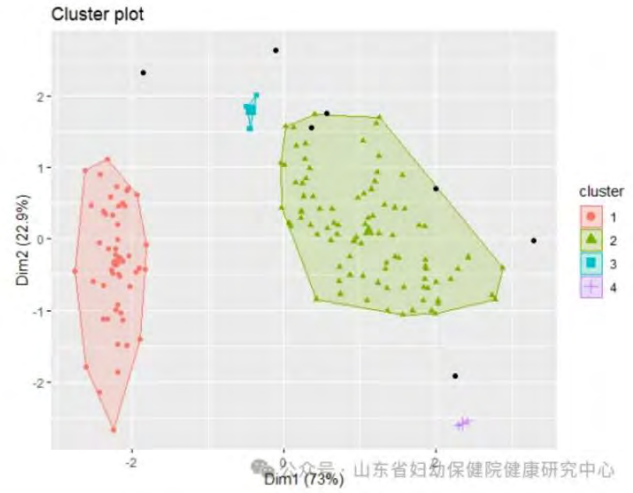


➢ 聚类分析（Clustering Analysis）

- 聚类算法类型

3、基于密度（Density）的聚类：识别数据点密度较高的区域，并将其分组为簇。在现实数据中，簇多表现为非球状簇，而基于密度的聚类方法可以发现任意形状的簇。DBSCAN算法的原理是从某个核心点出发，不断向密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，把具有足够高密度的区域划分为簇，并可在“噪声”的空间中实现任意形状的聚类。DBSCAN算法的优点是能够克服基于距离的算法只能发现“类球形”聚类的缺点，对噪声数据不敏感，适用于密度可测量的数据。但是计算较复杂，耗时较多，聚类结果受到扫描半径和最小包含点数设置的影响。。





可见，在eps = 0.6, MinPts = 2的设定下，将数据表聚为了4类。但是也需要注意DBSCAN聚类对参数eps、MinPts的设置是非常敏感的，若指定不当，会导致聚类效果降低。

1.2 高维数据的降维分析

聚类分析 (Clustering Analysis)

根据对象分类

基因聚类 (Gene clustering)

样本聚类 (Sample clustering)

双聚类 (Bi-clustering)

根据监督分类

无监督聚类

(Unsupervized Clustering)

有监督聚类

(Superviszed Clustering)

根据算法分类

层次聚类

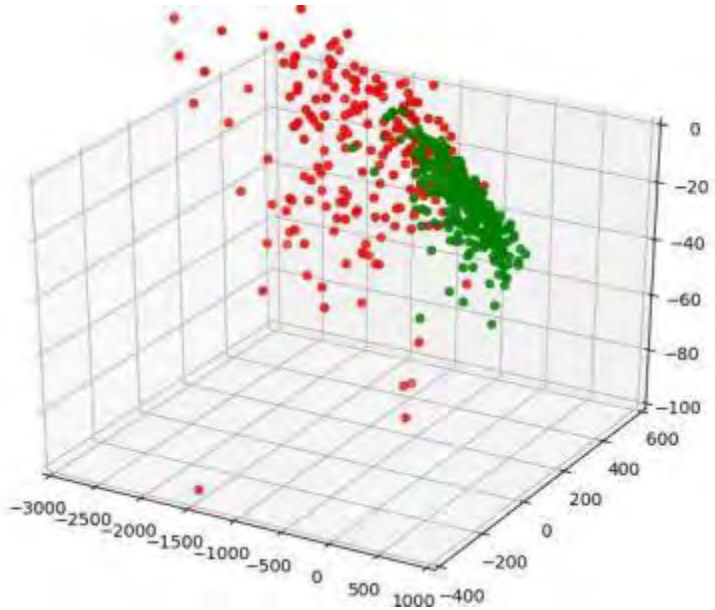
(Hierarchical Clustering)

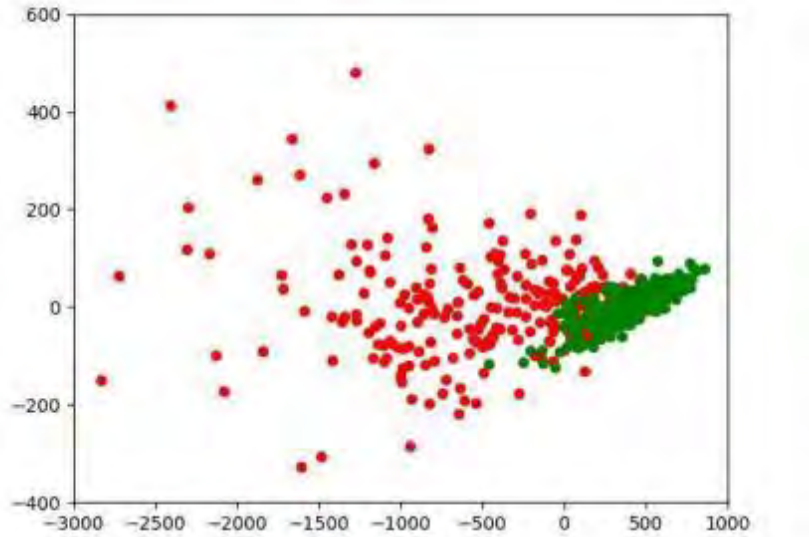
K-means聚类 (Partition clustering)

基于密度的聚类 (Density based clustering)

DBSCAN密度聚类法

- 例：三维→二维


0
−20
−40
−60
−80
−100
600
400
200
0
−200
−400
−3000
−2500
−2000
−1500
−1000
−500
0
500
1000


600
400
200
0
−200
−400
−3000
−2500
−2000
−1500
−1000
−500
0
500
1000


三体
THREE-BODY

1.3 降维的类型

特征选择（feature selection）

从原始数据特征中选取出最有价值特征的特征子集，以提高模型的性能，也是分离出最一致的、非冗余的和相关的特征以用于模型构建的过程。

目的是减少数据的维度，提高模型的训练效率，降低过拟合的风险，并可能提高模型的泛化能力。

应用场景：

1、**高**维数据处理：当数据集具有大量特征时（如文本分类、基因数据分析），特征选择有助于降低计算复杂度。

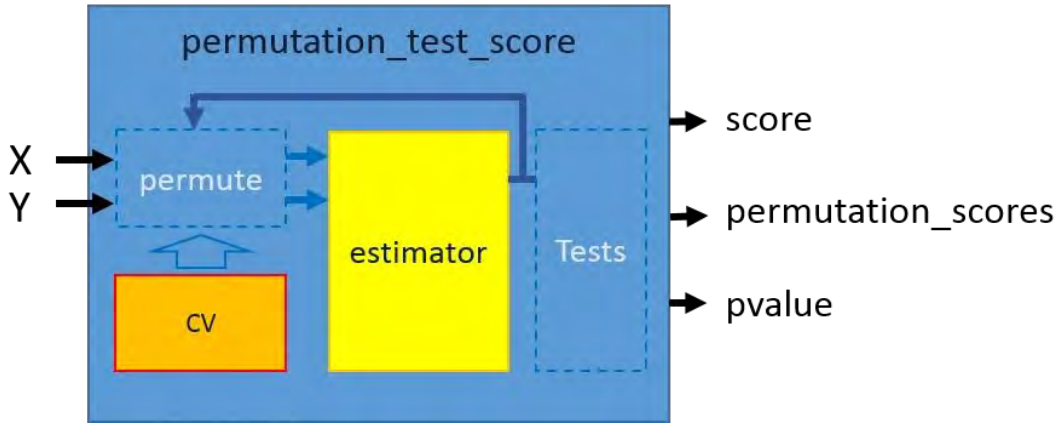
2、**提升模型性能**：**通**过去除无关或冗余特征，提高模型的预测准确率。

3、**数据可解**释性：选出关键特征后，模型的决策过程更容易被理解

特征选择（feature selection）

实现方式主要分为三类：

过滤法（Filter Methods）
包装法（Wrapper Methods）
嵌入法（Embedded Methods）。



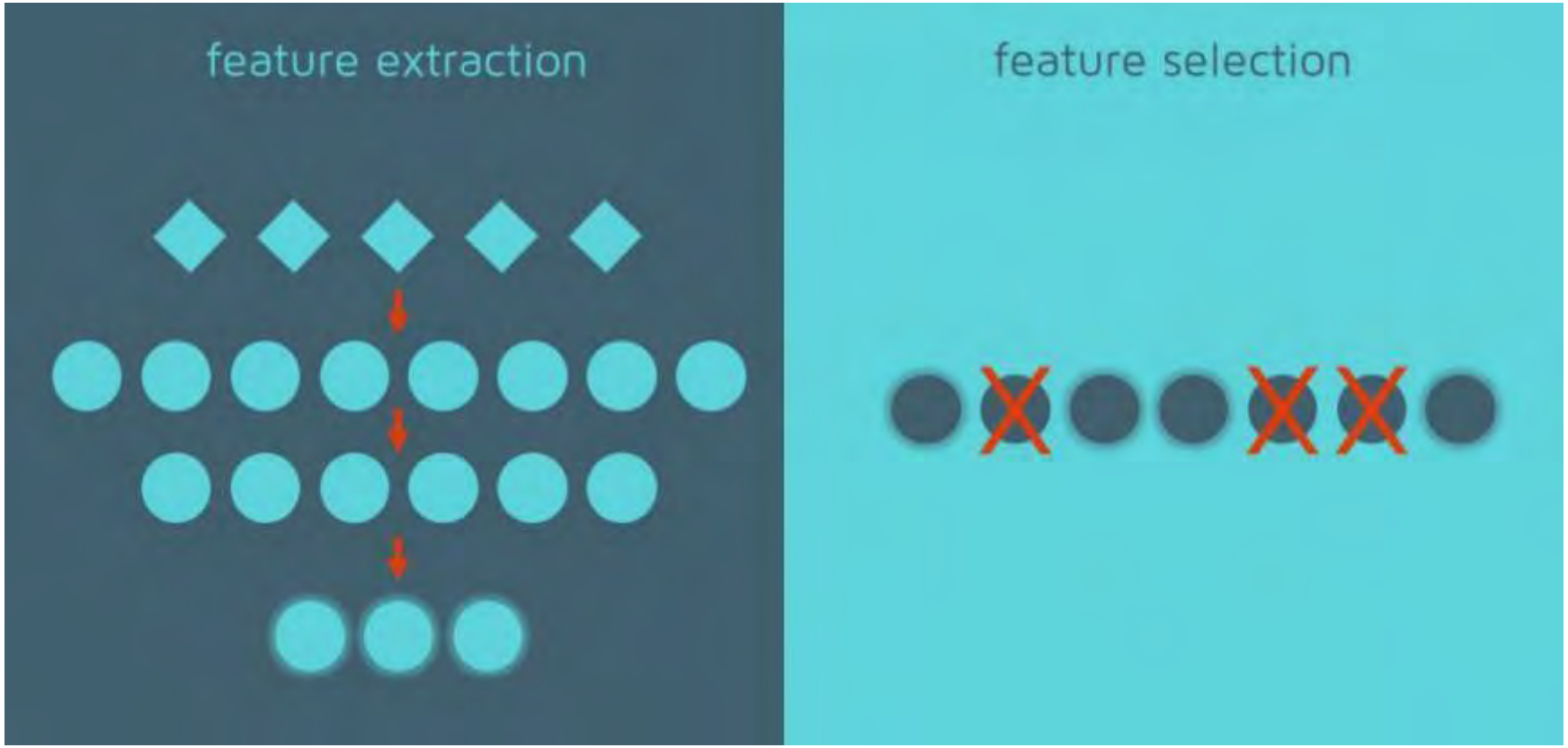
过滤方法（Filter Methods）：这类方法先对数据集进行特征选择，**然后再**训练学习器。它们根据特征的统计性质来选择特征，如相关**系数、卡方**检验等。(**差异性基因表达**fold change & t-test)
常见的过滤法包括：
方差选择法：选择方差大于某阈值的特征。
相关系数法：计算特征与目标变量之间的相关系数，选择相关系数**大于某**阈值的特征。
卡方检验：用于分类任务，检验非负特征和输出之间的关系。
互信息法：计算特征与目标变量之间的互信息，选择互信息大于某阈值的特征。

包装方法（Wrapper Methods）：这类方法将特征选择过程和模型训练过程结合起来。它们通过选择一组特征并训练模型，根据模型**的性能来**评价这组特征的好坏。

嵌入方法（Embedded Methods）：这类方法在模型训练过程中进**行特征**选择，如Lasso**回**归。

1.3 降维的类型

- 特征选择 V.S.特征提取
  - 提取：从现有数据中获取有用的特征
  - 选择：从原始特征库中选择一个子集



1.3 降维的类型

- 特征提取（feature extraction）

将数据从高维空间转换到维数较少的空间。数据变换可能是线性的， 如主成分分析(PCA)，但也存在许多非线性降维技术，例如UMAP,t-SNE。

PCA基本思想：构造原变量的一系列线性组合形成几个综合指标，以去除数据的相关性，并使低维数据最大程度保持原始高维数据的方差信息。

主成分个数的确定：

贡献率：第i个主成分的方差在全部方差中所占比重，反映第i个主成分所提取的总信息的份额。

累计贡献率：前k个主成分在全部方差中所占比重

主成分个数的确定：累计贡献率>0.85

02

PCA主成分分析原理

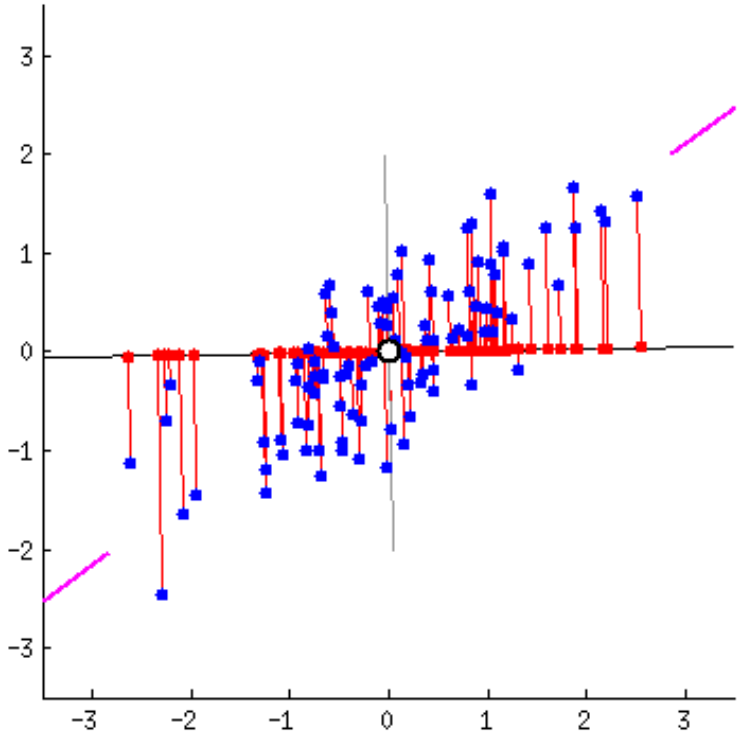
2.1 PCA主成分分析概述

Principal component analysis (PCA) 主成分分析

- PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第1,2个轴正交的平面中方差最大的。依次类推，可以得到n个这样的坐标轴。
- Principal Component Analysis or PCA is a linear feature extraction technique. It performs a linear mapping of the data to a lower-dimensional space in such away that the variance of the data in the low-dimensional representation is maximized. It does so by calculating the eigenvectors from the covariance matrix. The eigenvectors that correspond to the largest eigenvalues (the principal components) are used to reconstruct a significant fraction of the variance of the original data.

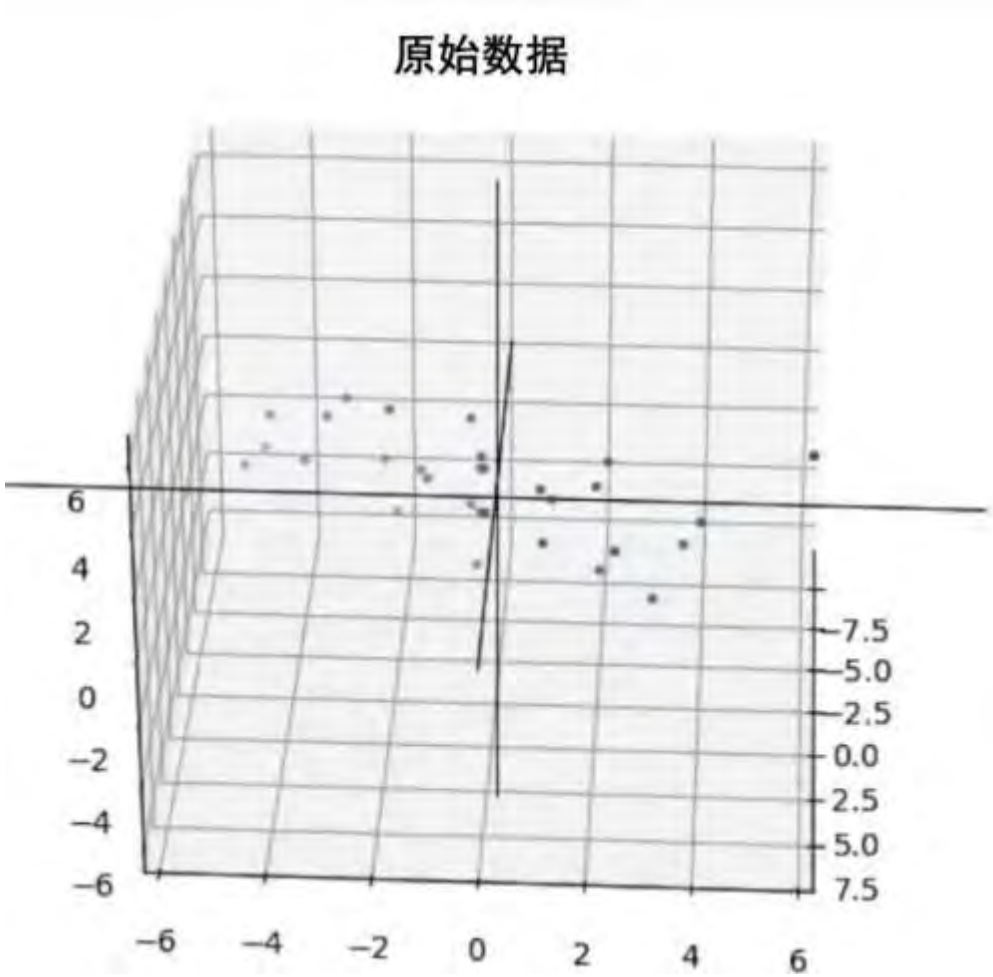
2.1 PCA主成分分析概述

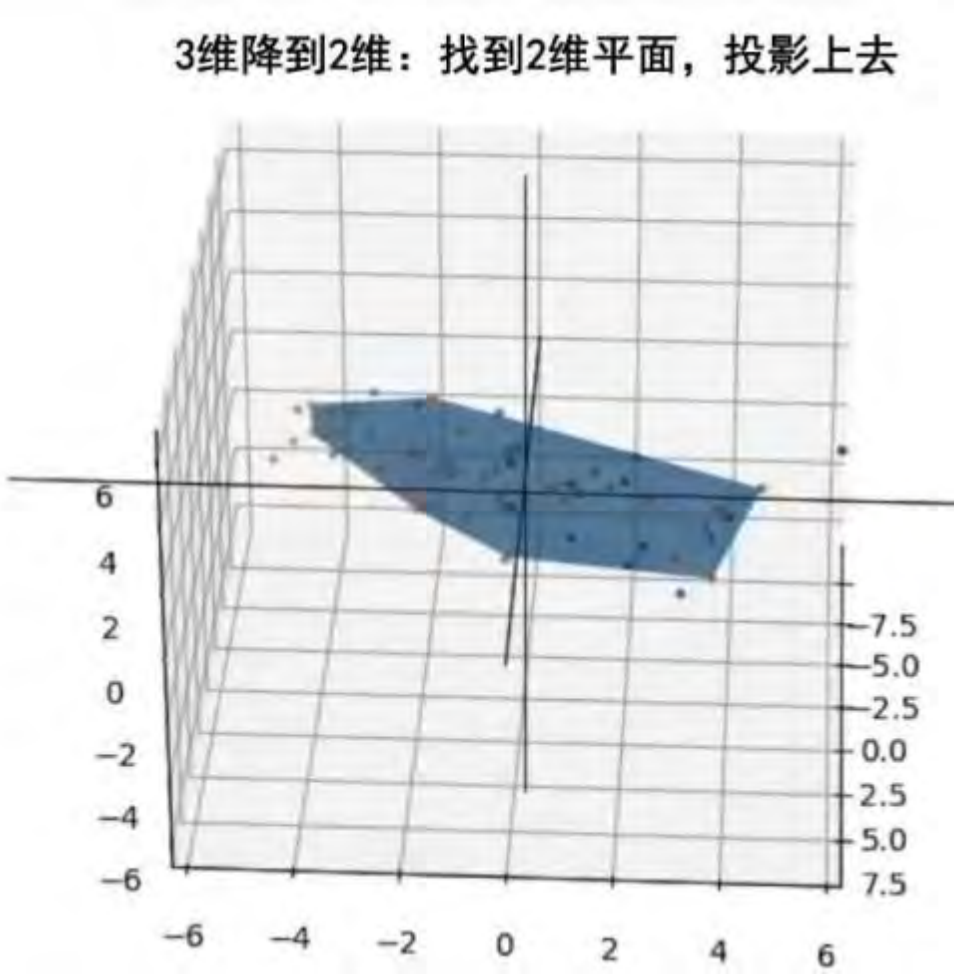
- PCA：二维数据集-寻找一维坐标轴



2.1 PCA主成分分析概述

- PCA：三维数据集-寻找二维平面



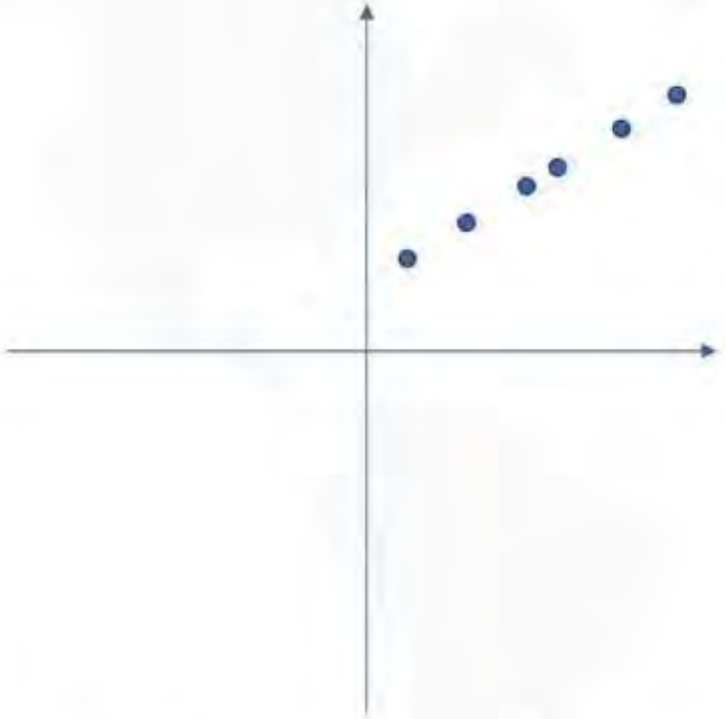


2.2 PCA寻找坐标系

- PCA怎么求解

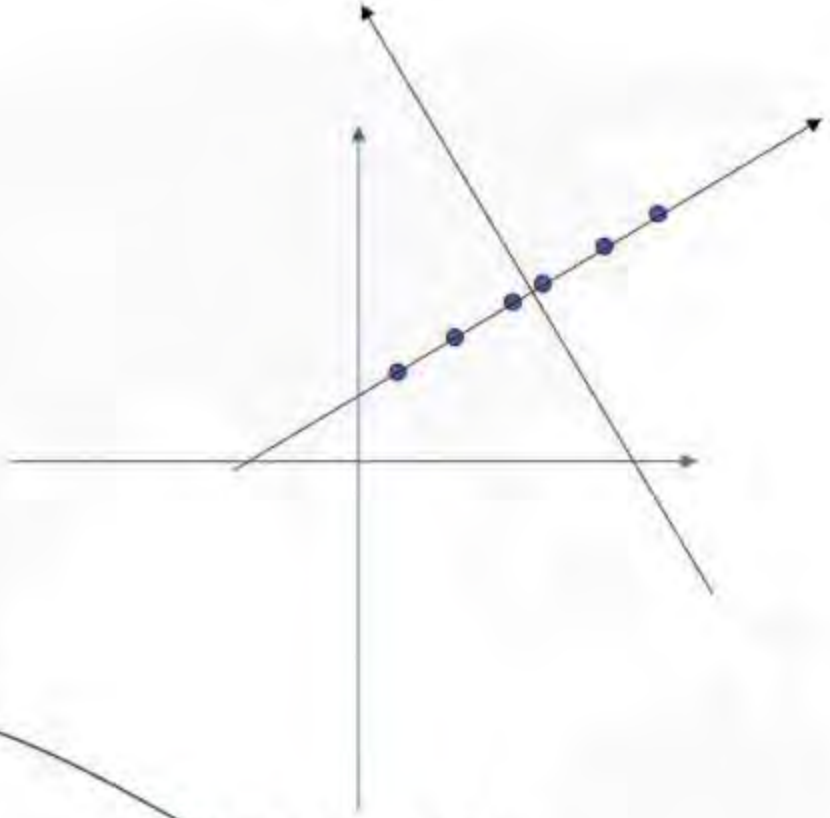
2.2 PCA寻找坐标系

- PCA是什么?
- 答：找坐标系



需要保存的信息 $D = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{bmatrix}$

降维



需要保存的信息

新坐标系的原点

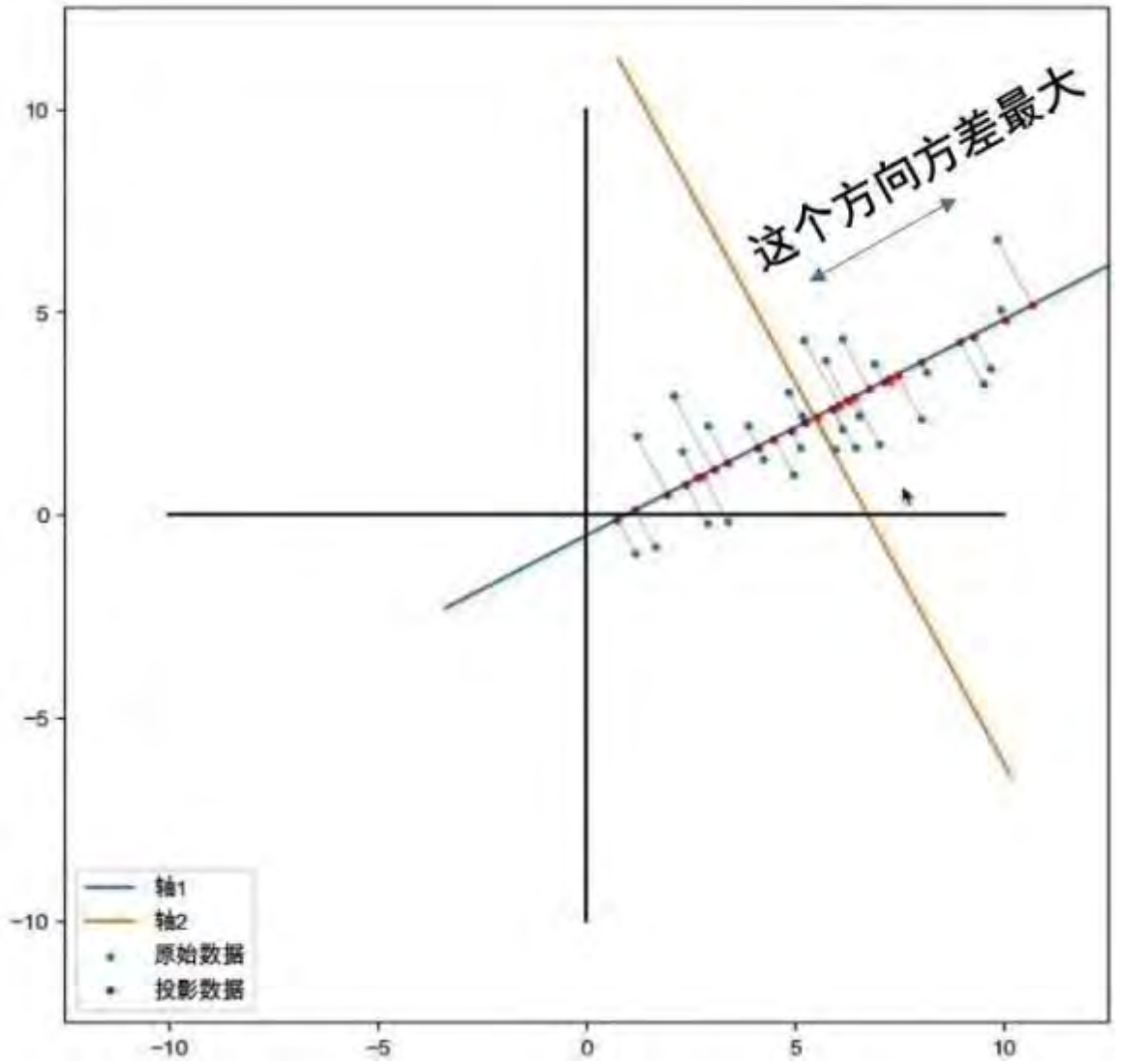
新坐标系的角度

新坐标点 $D' = \begin{bmatrix} x_1' & x_2' & x_3' & x_4' \\ 0 & 0 & 0 & 0 \end{bmatrix}$

2.2 PCA寻找坐标系

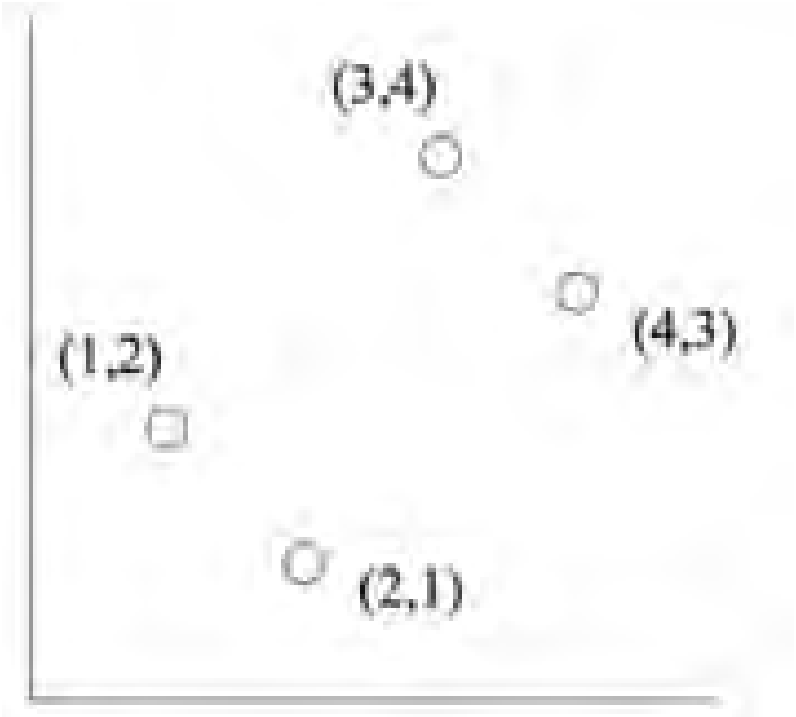
- 目标：只保留一个轴的时候（二维降到一维），信息保留最多

- 怎么样最好：找到数据分布最分散的方向（方差最大），作为主成分（坐标轴）



找坐标系

假设二维空间中有四个点：(1,2),(2,1),(3,4),(4,3)，如下图：



1）用矩阵表示这四个点：

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$$

2）**求出每一列的均**值，然后用 M **的每一行减去**对应列的均值：

$$M=\begin{bmatrix} -1.5 & -0.5 \\ -0.5 & -1.5 \\ 0.5 & 1.5 \\ 1.5 & 0.5 \end{bmatrix}$$

3）计算样本的协方差矩阵：

$$M^TM=\begin{bmatrix} -1.5 & -0.5 & 0.5 & 1.5 \\ -0.5 & -1.5 & 1.5 & 0.5 \end{bmatrix}\begin{bmatrix} -1.5 & -0.5 \\ -0.5 & -1.5 \\ 0.5 & 1.5 \\ 1.5 & 0.5 \end{bmatrix}=\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$$

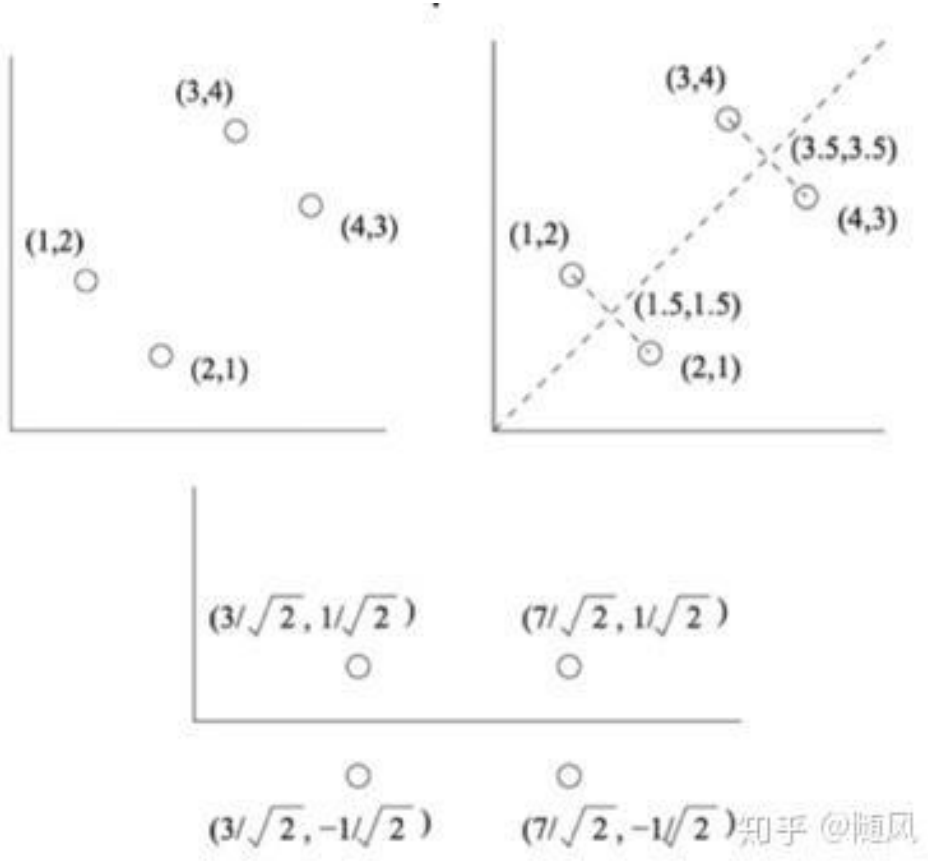
4）求解协方差矩阵的特征值与特征向量：

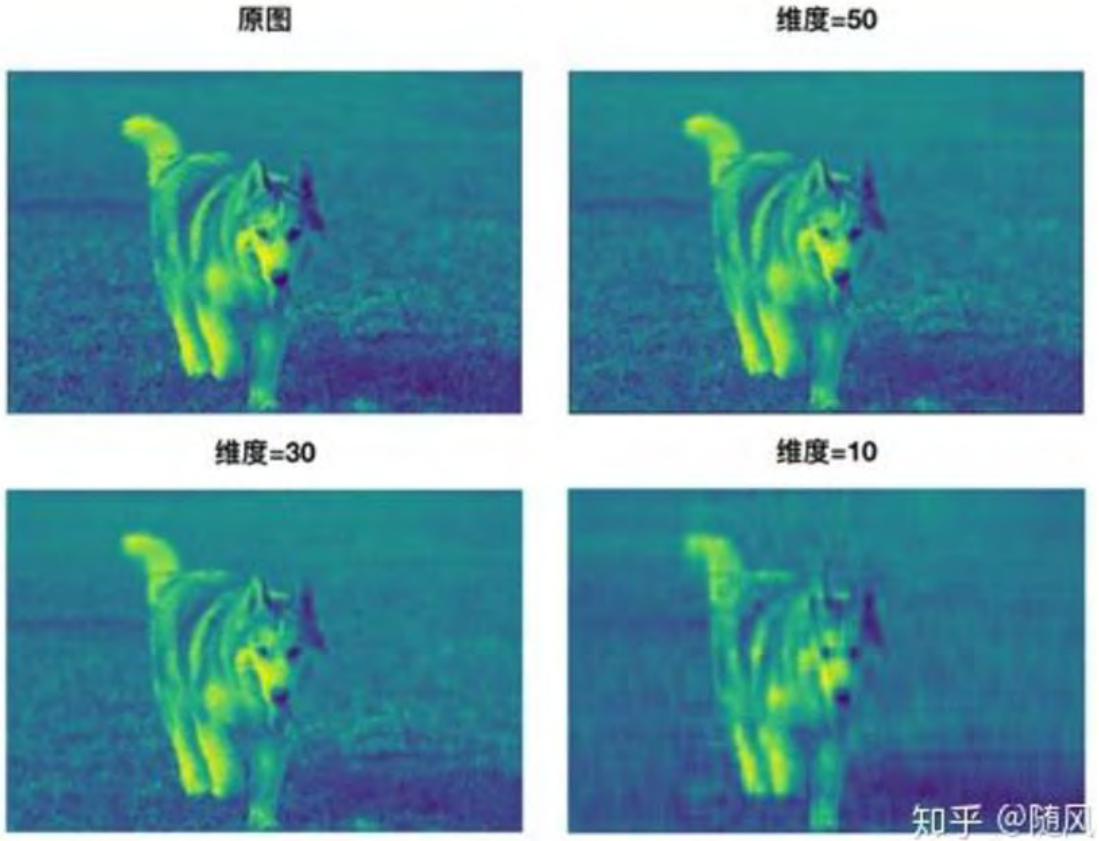
$$\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\lambda\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\rightarrow \lambda=(8,2), E=\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

5）用原矩阵 M 乘以 特征向量矩阵 E，得到新的矩阵：

$$ME=\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}=\begin{bmatrix} \frac{3}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

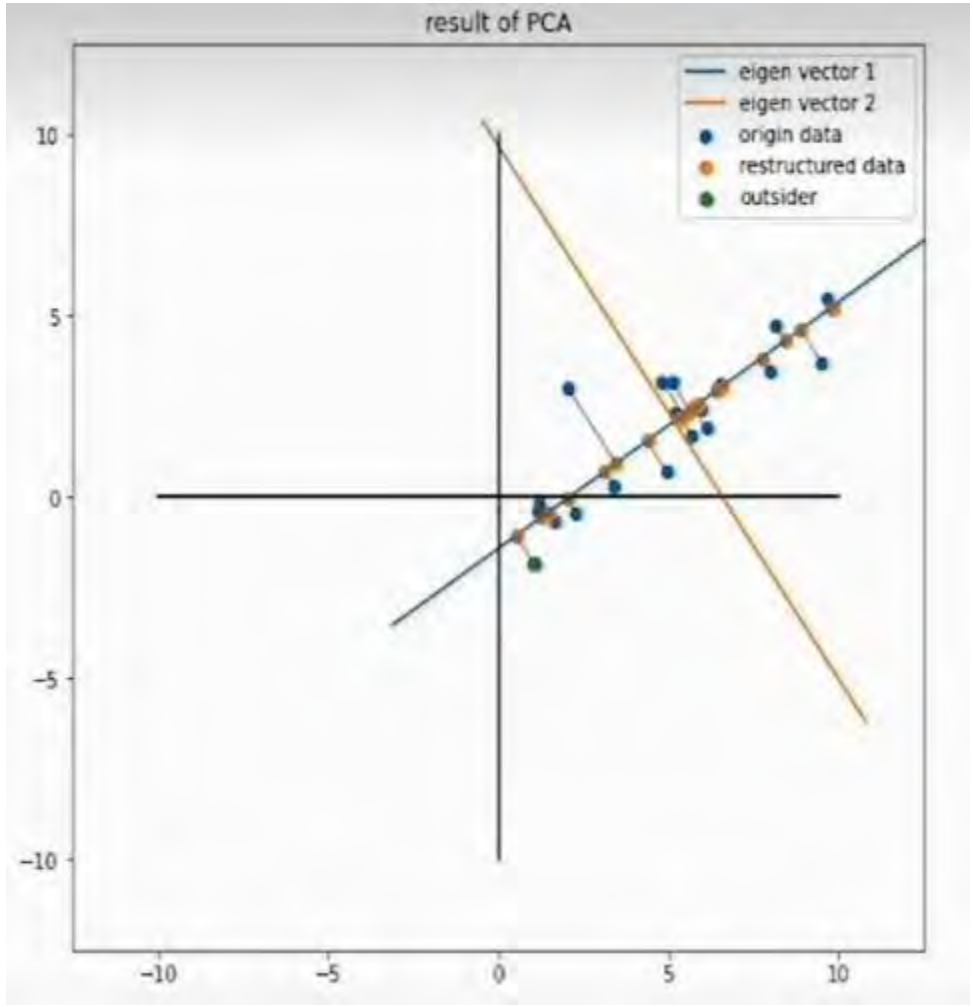
从原矩阵到新矩阵的转换就是 PCA 的处理过程，用图表示为：





2.4 离群点对PCA的影响

- PCA的缺点：离群点影响大



PCA 寻找的是数据之间差别最大的方向，假如我们有两张图片，分别是猫在白天和狗在晚上拍摄的，那么经过 PCA 处理后的向量很可能将白天和晚上作为差别最大的特征，**而不是**寻找区分猫和狗最大的差别作为坐标轴。

1.4 常见的降维方法

- PCA：主成分分析
  (Principal component analysis)

- t-SNE：t-分布式随机邻接嵌入
  (t-Distributed Stochastic Neighbor Embedding)

- UMAP：统一流形逼近与投影
  (Uniform Manifold Approximation and Projection)

1.4 常见的降维方法

常见的降维方法性能解读

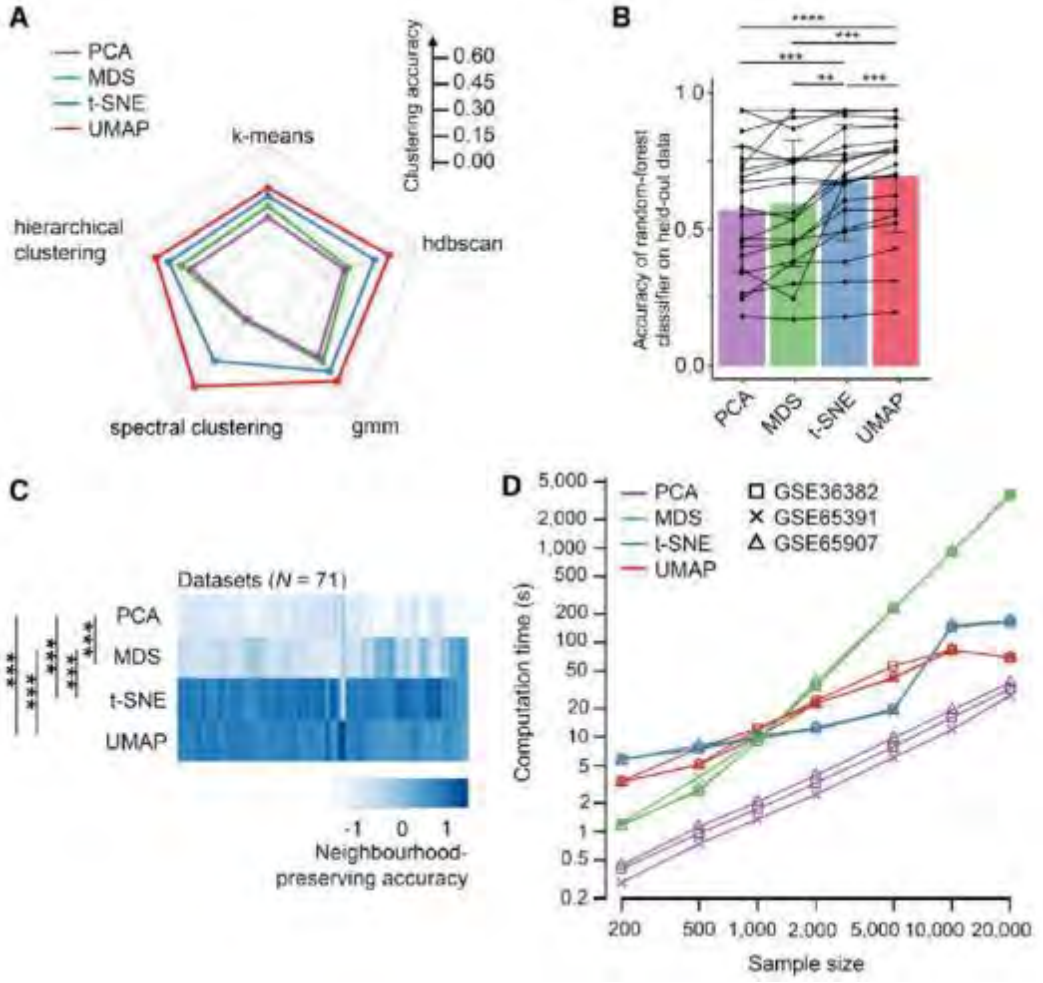
●特征提取之主成分分析PCA
（Principal component analysis ）

●特征提取之t-SNE

是一种非线性的降维技术，特别适合于高维数据集的可视化。它被广泛地应用于图像处理、 NLP、基因组数据和语音处理。

●特征提取之UMAP

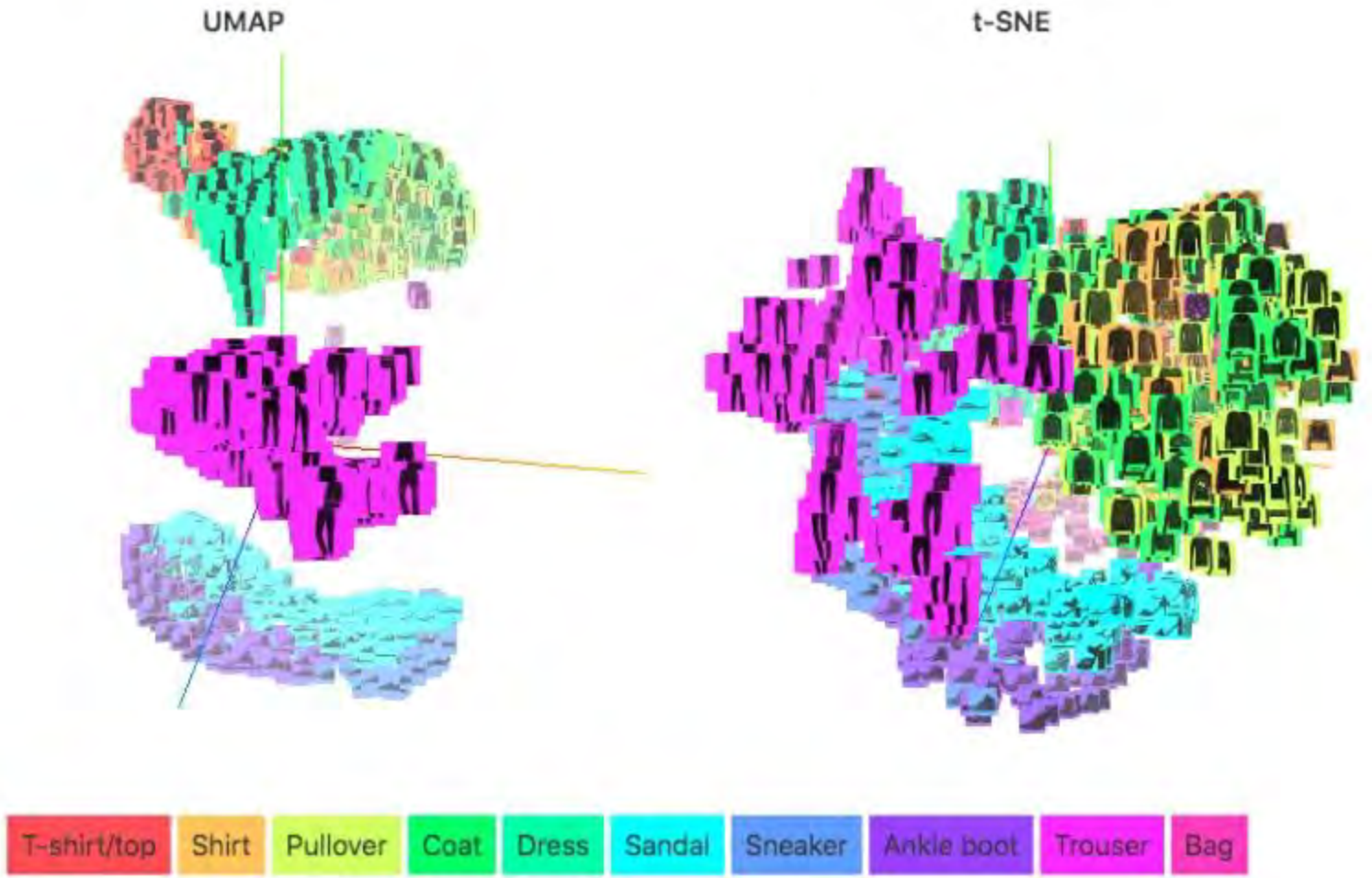
利用局部流形逼近和局部模糊单纯形集表示来构造高维数据的拓扑表示。给定数据的一些低维表示，可以使用类似的过程来构造等价的低维拓扑表示——高维和低维换种表达方式。



Yang Y, et al. *CellRep.* **2021**, 36(4):109442.

1.4 t-SNE和UMAP比较

下图是UMAP和t-SNE对一套784维Fashion MNIST高维数据集降维到3维的效果的比较。高清3D图参见：https://pair-code.github.io/understanding-umap/



1.4 常见的降维方法比较

特性	PCA	UMAP	t-SNE
类型	线性	非线性，基于流形学习	非线性，基于概率嵌入
全局/局部	全局	局部和全局平衡	局部为主
复杂度	低（快，适合大数据集）	中等（更快，更适合大数据集）	高（慢，计算量大）
可解释性	高（主成分可解释）	中等	低（结果不易解释）
保留结构	线性结构	全局和局部	局部（全局失真）
调参难度	低	中（需调整邻居数和距离参数）	高（perplexity 和学习率较难调）
适用场景	特征提取、数据预处理	大规模数据的降维和可视化	小规模数据的精细局部结构探索

1.4 常见的降维方法应用场景

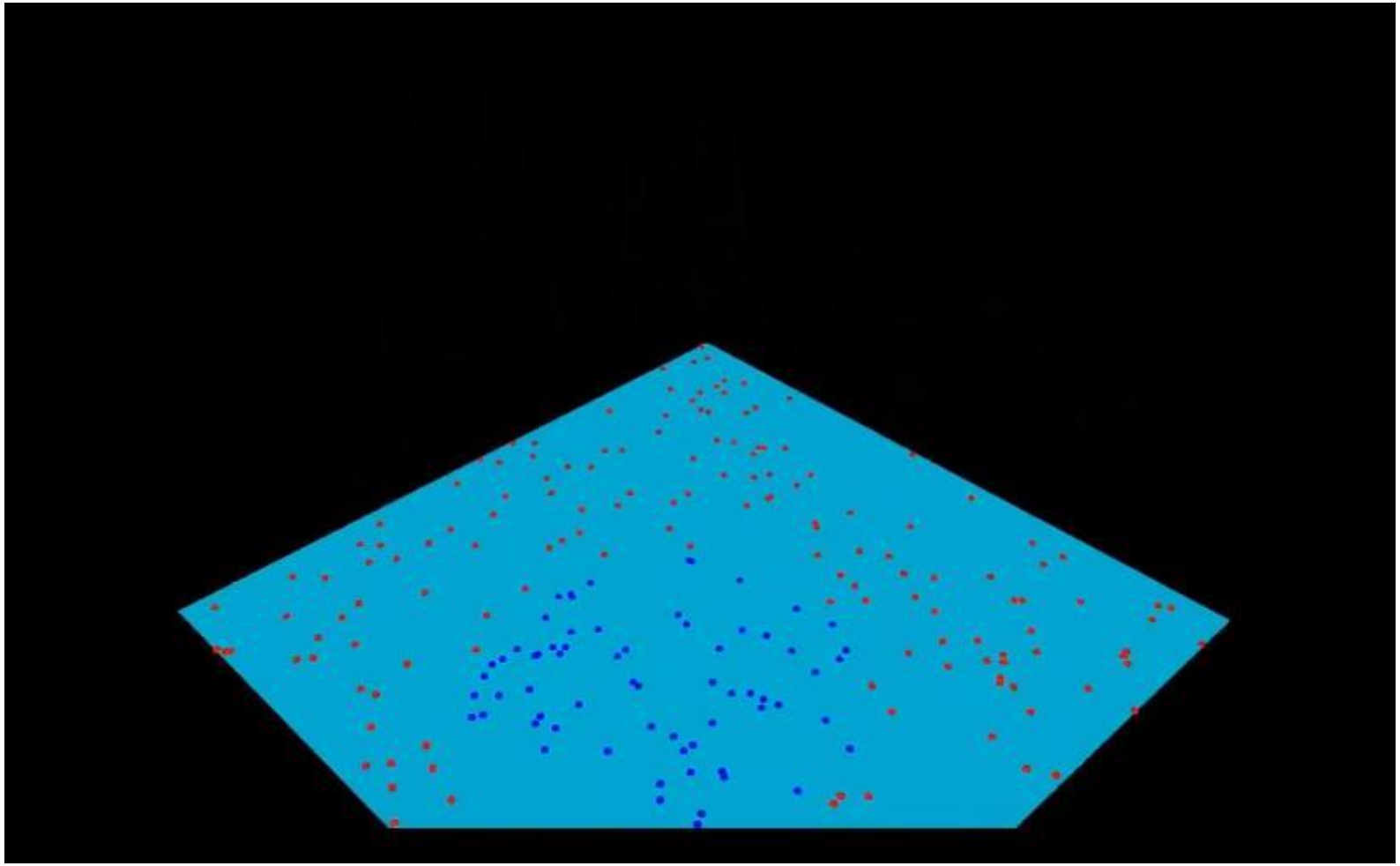
PCA：**用于**线性数据的降维、去噪和特征提取。适用于初步分析、处理大规模数据以及作为其**他复**杂分析的前置步骤。

UMAP：**适合**处理高维、大规模数据，特别是基因表达分析、单细胞 RNA-seq、图像处理和嵌**入式学**习等领域。UMAP **能很好地平衡局部和全局**结构。

t-SNE：**通常用于探索数据中的局部复**杂结构，广泛用于数据可视化，例如基因表达数据、神经网络特征嵌入等。但它计算速度慢，适合小规模数据集。

1.5 升维降维解决分类预测问题

- 例如：支持向量机SVM



03

PCA分析实践

3.1 前期准备

准备所需文件

```
# 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
```

3.2 准备VCF文件

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
```

3.3 文件合并

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

$ /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

```
LY@ecoli-ThinkCentre-M930q-N000:~/05_PCA$ ll ./pca
total 252
drwxr-xr-x 2 LY student   4096 11月 15 20:34 ./
drwxr-xr-x 3 LY student   4096 11月 15 20:33 ../
-rw-r--r-- 1 LY student 127284 11月 15 20:34 merged.vcf.plink.bed
-rw-r--r-- 1 LY student   5573 11月 15 20:34 merged.vcf.plink.bim
-rw-r--r-- 1 LY student  62619 11月 15 20:34 merged.vcf.plink.fam
-rw-r--r-- 1 LY student   1132 11月 15 20:34 merged.vcf.plink.log
-rw-r--r-- 1 LY student  40074 11月 15 20:34 merged.vcf.plink.nosex
```

3.3 生成PCA文件

- 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

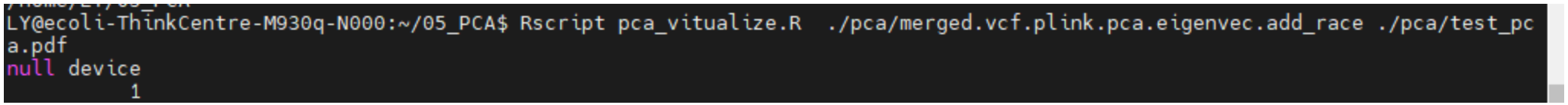
- 结果可视化

#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

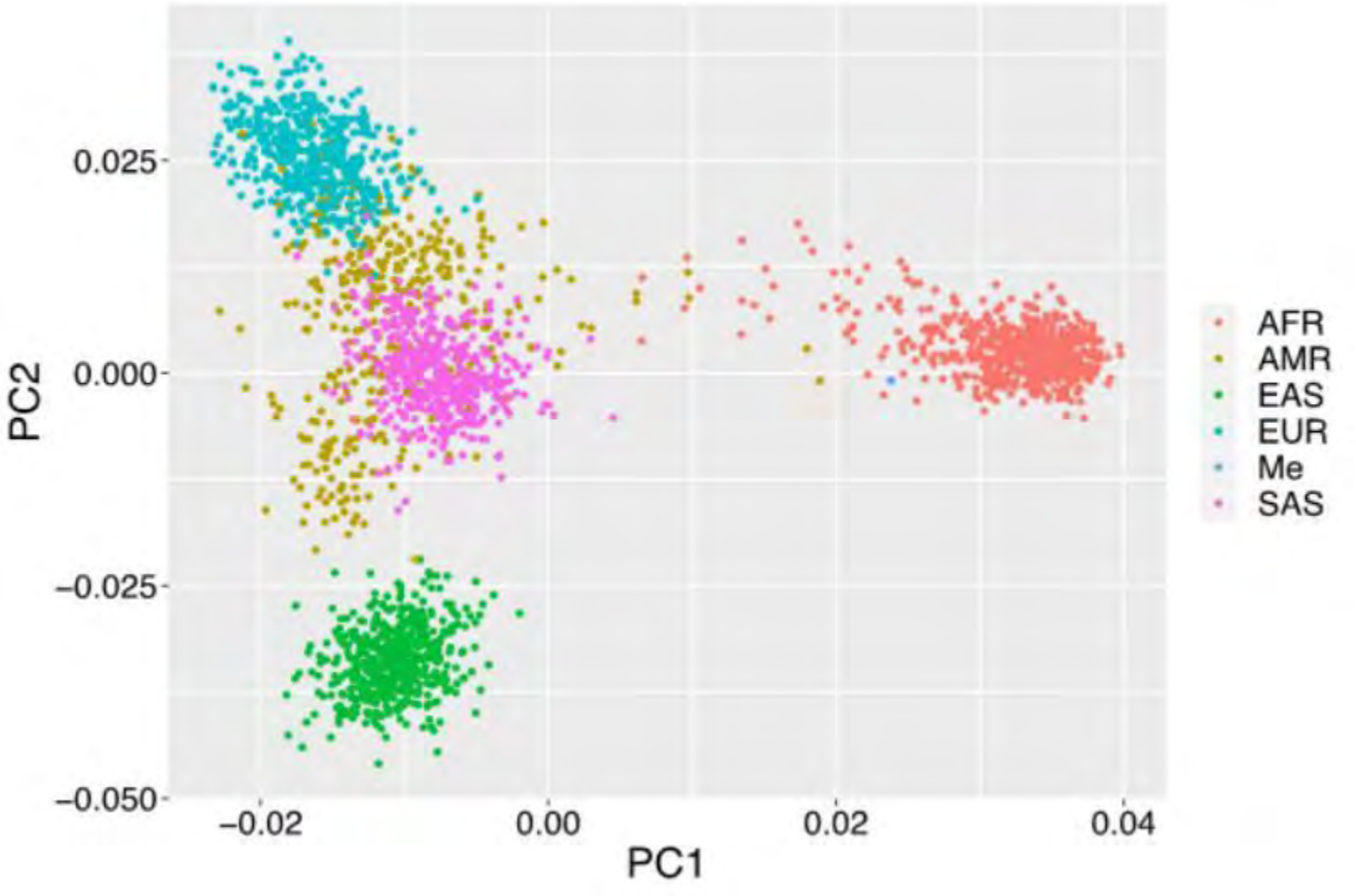
#使用R进行可视化

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



3.5 PCA分析结果

- PCA分析结果图（以人种分类）



3.5 PCA分析结果

- PCA分析结果图（以国家分类）