PCA:

PCA（Principal Component Analysis）是一种常用的数据分析方法。PCA通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维。

PCA本质上是将方差最大的方向作为主要特征，并且在各个正交方向上将数据“离相关”，也就是让它们在不同正交方向上没有相关性。

## PCA算法

总结一下PCA的算法步骤：

设有m条n维数据。

1）将原始数据按列组成n行m列矩阵X

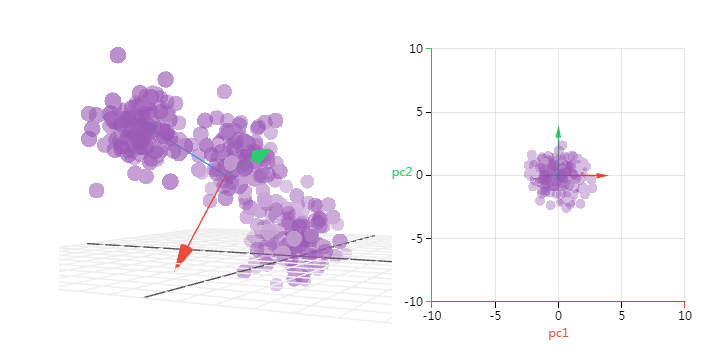
2）将X的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值

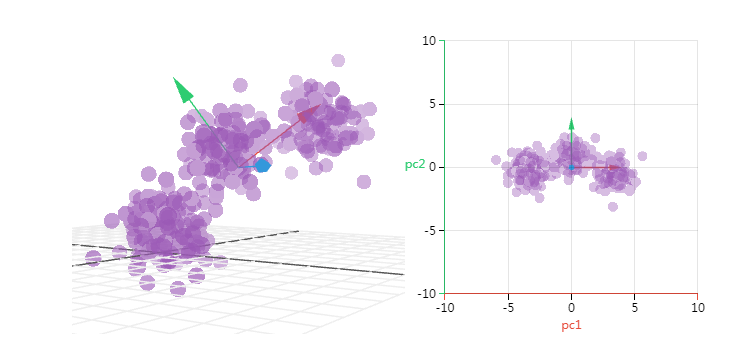
3）求出协方差矩阵

4）求出协方差矩阵的特征值及对应的特征向量

5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P

6）即为降维到k维后的数据





PCA的主要适用场景：

（1）非监督式的数据集

它是一种非监督式的降维方法，因此适用于不带有标签的数据集，对于带有标签的可以采用LDA

（2）根据方差自主控制特征数量

最大的主成分的数量会小于或等于特征的数量，即，PCA可以输出全部的特征，具体取决于选择特征中解释的方差比例

（3）更少的正则化处理

选择较多的主成分将导致更少的平滑，因为能保留很多特征，减少正则化

（4）数据量较大的数据集

数据量大指数据记录多和维度多两种情况，PCA对大型数据集的处理效率高

（5）数据分布是位于相同平面上，数据中存在线性结构

加快算法的运算速度

压缩数据,节省内存

算法运行太慢或者占用内存太多了就用PCA

电脑跑不动则PCA

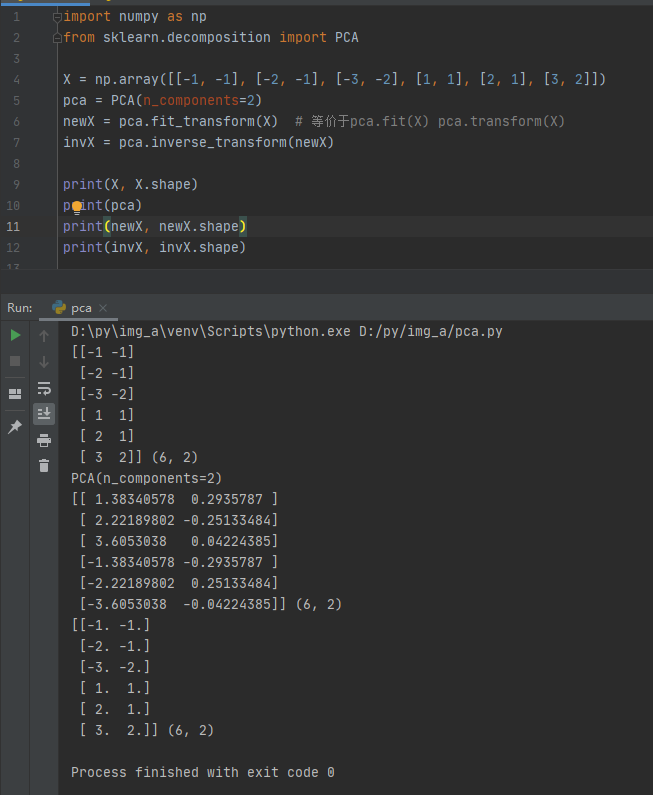
## **数据降维**

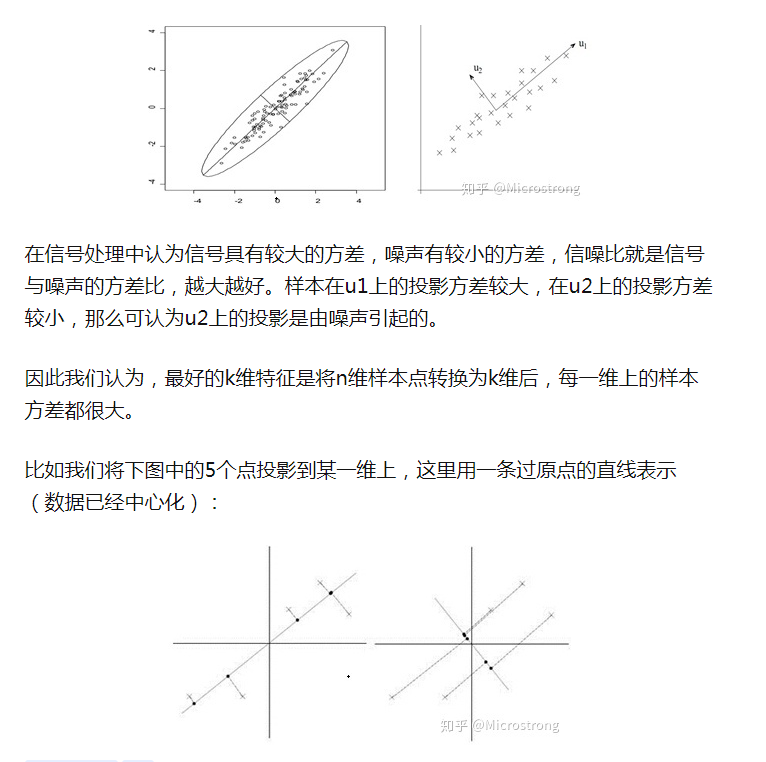
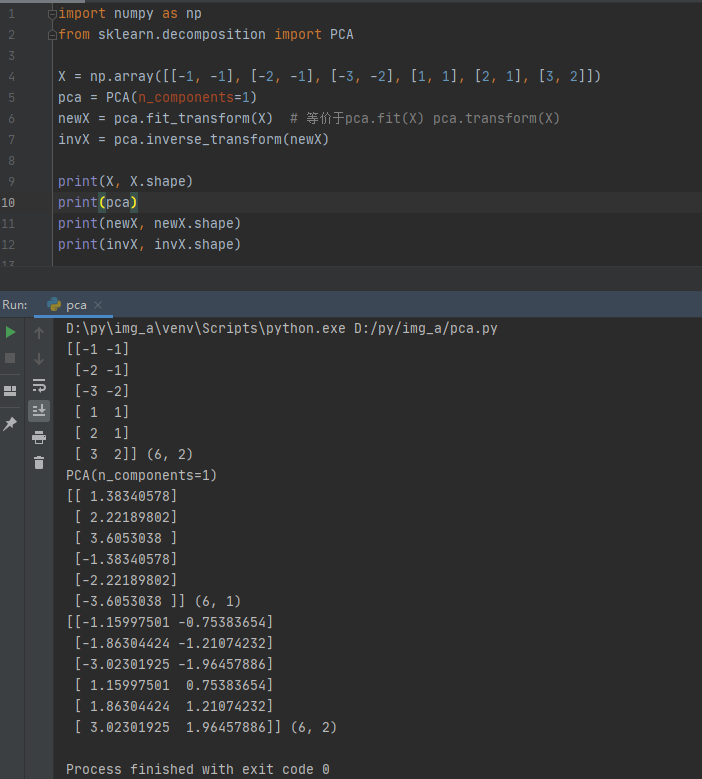
降维就是一种对高维度特征数据预处理方法。降维是将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的信息损失范围内，可以为我们节省大量的时间和成本。降维也成为应用非常广泛的数据预处理方法。

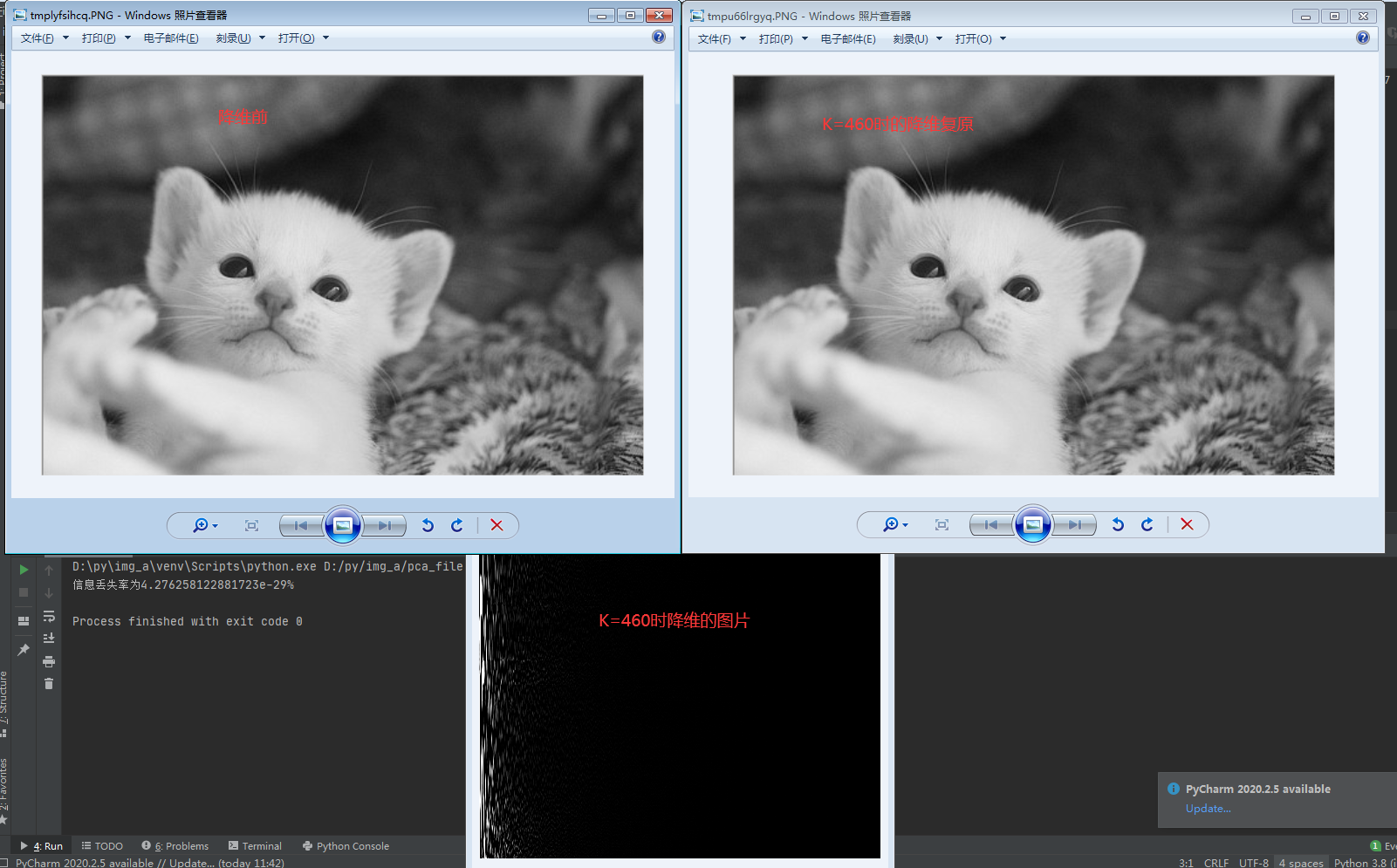
降维具有如下一些优点：

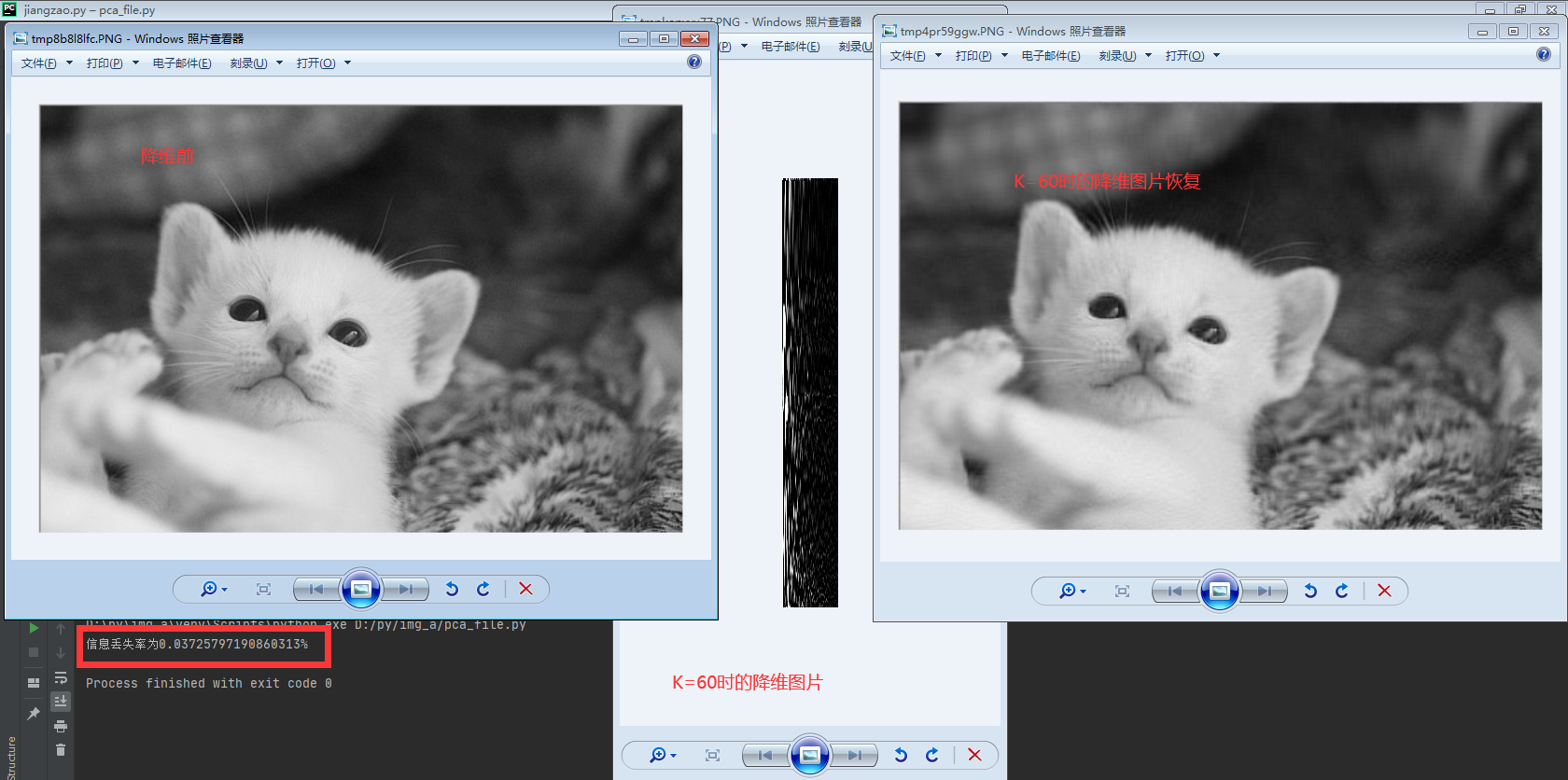
* 使得数据集更易使用。
* 降低算法的计算开销。
* 去除噪声。
* 使得结果容易理解。

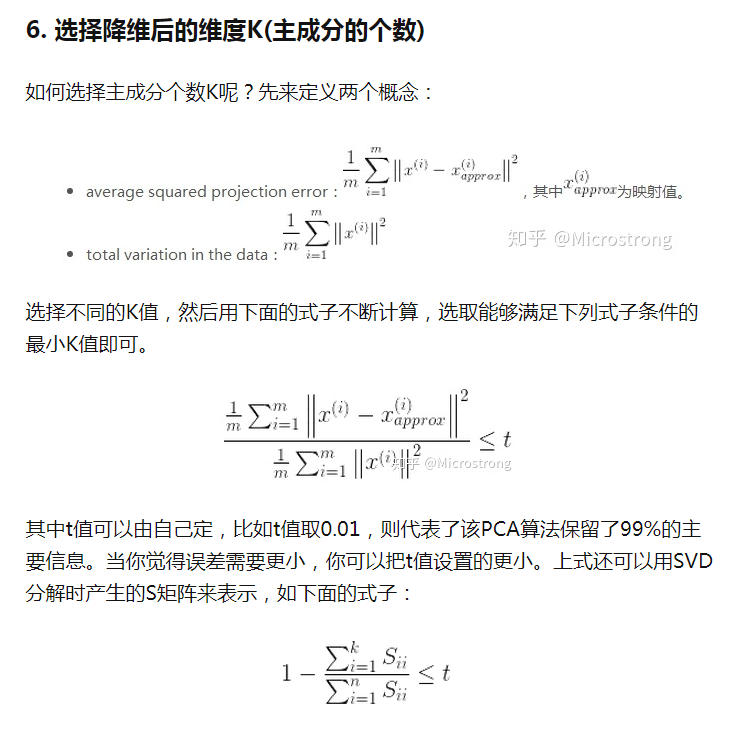
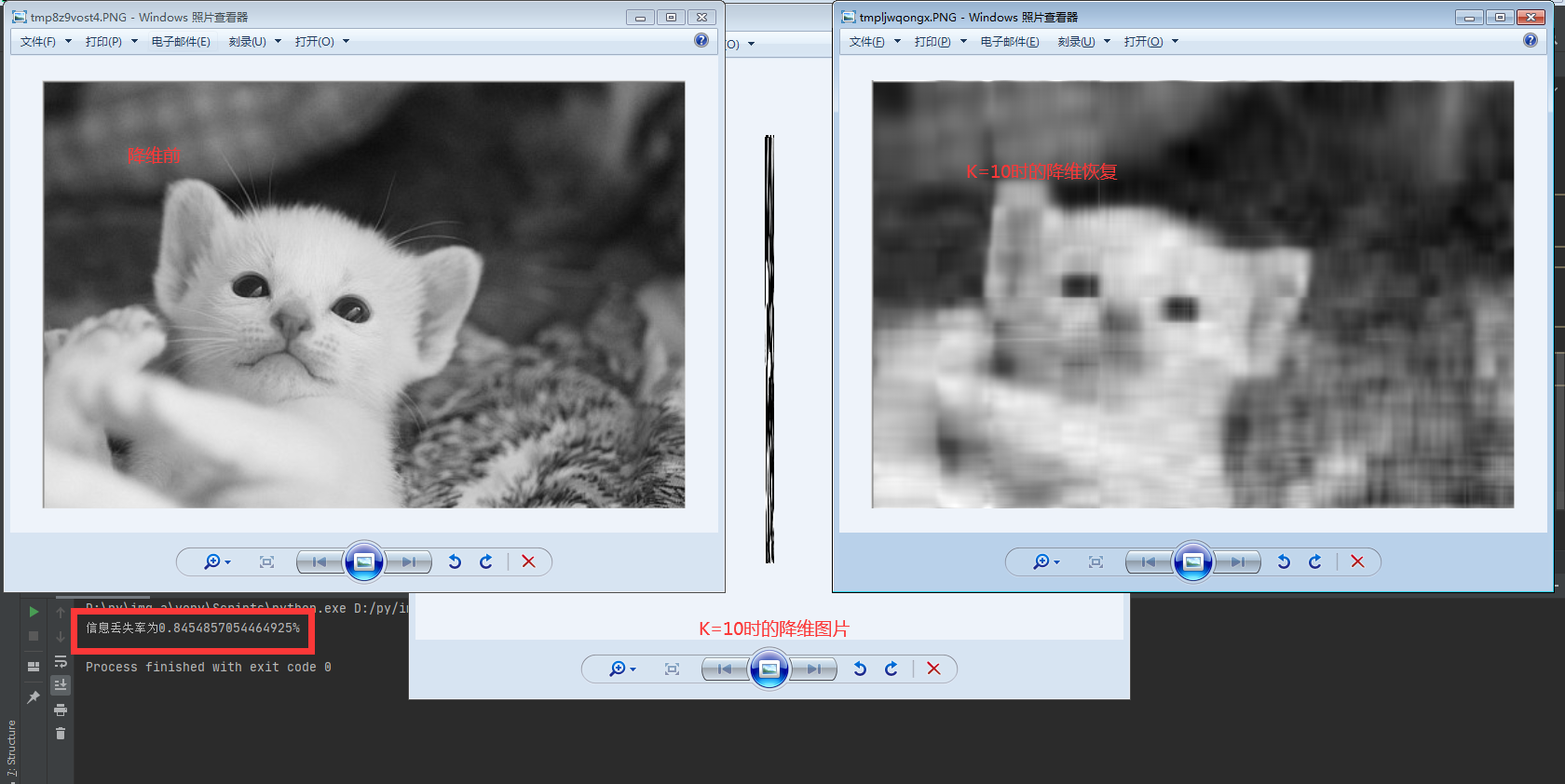
降维的算法有很多，比如[奇异值分解(SVD)](https://link.zhihu.com/?target=https://mp.weixin.qq.com/s/Dv51K8JETakIKe5dPBAPVg" \t "https://zhuanlan.zhihu.com/p/_blank)、主成分分析(PCA)、因子分析(FA)、独立成分分析(ICA)。











应用:

高维数据可视化

