1 PCA(3 人)

在科研过程中，我们经常碰到需要处理大量数据的情况，举个例子，一个体系有 N 个可分辨粒子，每个粒子的动量大小作为描述其特征的物理量，这样我们在完全描述这样一个体系时需要用到 N 维的信息，而实际上，我们有时候要关注的仅仅是这个体系的突出特征，N 维特征对我们来说显然是过大了，因此我们需要合适的算法进行降维处理。在本题中，我们即考虑主成分分析 (Principal Component Analysis,PCA).

假设我们关注的体系有 N 个特征，在我们对体系进行 M 次采样后我们得到了 M 个 N 维向量，因此我们可以想象我们的体系是分布在这样一个 N 维空间的 M 个点，而 PCA 的思路就是对这个空间寻求新的基矢，使得我们的数据在这些基矢上有最大的方差. 更数学化的表述如下：

- 我们得到了具有 N 维特征的 M 组采样数据 $A(M \times N$ 维矩阵).
- 我们对得到的采样数据进行去中心化，即将数据挪到中心为原点；再计算 N 维特征的协方差矩阵 $\frac{1}{M}A^TA$.
- 对得到的协方差矩阵求取特征值与特征向量，选择较大的 k 个特征向量，将数据投影到这 k 个向量构成的空间内.

试回答以下问题:

1. 根据我们给出的数据，画出数据在不同的维度上的分布图 (可以是一维也可以是二维)，思考按照前面所说的 PCA 的原理应该得到怎样的结果.
2. 在题目中我们计算协方差时考虑的系数是 $\frac{1}{M}$，而对于真实的协方差应当是 $\frac{1}{M-1}$，试考虑这一系数的存在与否是否有影响.
3. 计算协方差矩阵并进行矩阵分解，将数据投影到新的 k 维空间内，观察在不同维度上的分布图，思考是否得到了你想要的特征.

除此之外，对于非方阵的矩阵分解也经常使用奇异值分解 (SVD)，即我们有一个 $m \times n$ 维的矩阵 A，我们可以将其分解为 $A = U\Sigma V^T$，其中 U、V 分别为 $m \times m$ 维和 $n \times n$ 维的矩阵，Σ 为除对角线均为 0 的 $m \times n$ 维矩阵. 其具体的操作方式如下：

- 获得 AA^T 的特征值和特征向量，用单位化的向量构成 U.
- 获得 A^TA 的特征值和特征向量，用单位化的向量构成 V.
- 将 AA^T 或 A^TA 的特征值求平方根，构成 Σ.

4. 试思考如何利用 SVD 分解的方法对数据做 PCA 并比较两种方法的结果.

提示：

- 给出的数据在 data.txt 文件中，共有 10000 组数据，每组数据有 6 维特征，排列格式按照每行为一组数据的六个特征.