**深 圳 大 学 实 验 报 告**

|  |
| --- |
| **课程名称 机器学习**  **项目名称 实验二：LDA算法**  **学 院 计算机与软件学院**  **专 业 软件工程（腾班）**  **指导教师 赖志辉**  **报 告 人 黄亮铭 学号 2022155028**  **实验时间 2024年3月18日至2024年3月24日**  **实验报告提交时间 2024年3月24日** |

**教务处制**

# 一、实验目的与要求

1. 熟悉机器学习的基础算法之一LDA；

2. 学习从零实现LDA算法；

3. 了解不同监督学习算法之间的优劣性。

# 二、实验内容与方法

1. 简述LDA 原理、 算法模型与优化问题，并给出求解的全程推导细节（拉格朗日乘子法）；证明St=Sb+Sw；

2. 给出LDA的各种等价模型表示(除法的、减法的及其调换位置的等)， 在各数据集（不少于3个）比较PCA与“LDA的各种等价模型与正则模型”的人脸识别精度。

3.比较eigenface与fisherface的不同，并取3个类的图像投影在二维和三维空间中， 并用不同颜色的点表示不同的类，每个类选3个有代表性的点对应的人脸图像显示在该点的边上（用plot命令或imshow），比较PCA与LDA的结果的异不同。

4.参看前人其它论文实现一个新的线性鉴别分析方法或子空间学习方法，并与PCA,LDA做比较实验，打印出各方法训练样本数的增长其识别曲线变化图等各种情况，把简要内容写在本实验报告中

# 三、实验步骤与过程

1. 简述LDA 原理、 算法模型与优化问题，证明St=Sb+Sw。

I．简述LDA原理。

LDA（线性判别分析）利用了样本的类别（数据）标签，为有监督学习，是主流的一种**线性降维**算法。以**最小化类内方差，最大化类间方差**为目标导向，通过降维（投影），达到降维的目的更好地将样本分类。

假设我们有两类数据 分别为红色和蓝色，如下图所示，这些数据特征是二维的，我们希望将这些数据投影到一维的一条直线，让每一种类别数据的投影点尽可能的接近，而红色和蓝色数据中心之间的距离尽可能的大。

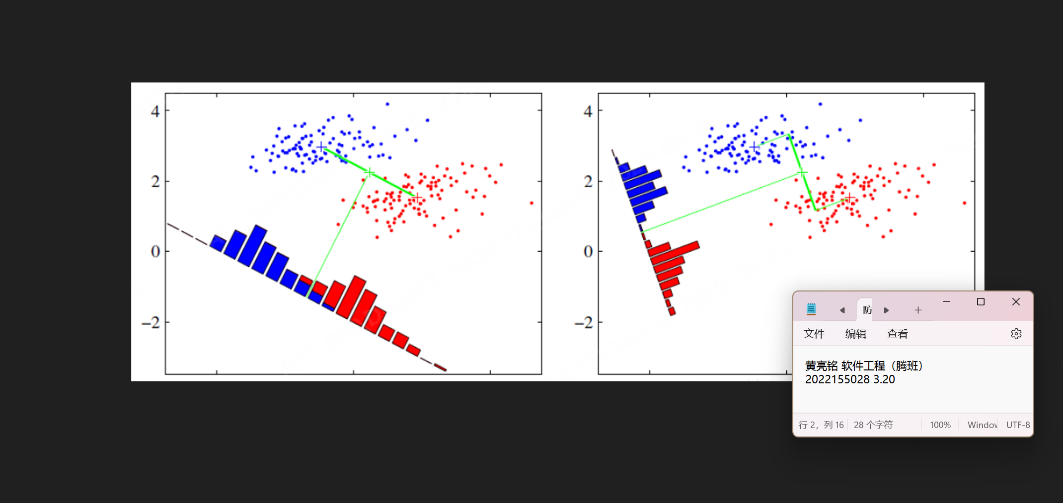


图1

从直观上可以看出，右图要比左图的投影效果好，因为右图的黑色数据和蓝色数据各个较为集中，且类别之间的距离明显。左图则在边界处数据混杂。以上就是LDA的主要思想了，当然在实际应用中，我们的数据是多个类别的，我们的原始数据一般也是超过二维的，投影后的也一般不是直线，而是一个低维的超平面。

LDA的数学原理：

*类内散度：SW*​=∑*i*=1*c*​*Si*

*类间散度：SB*​=∑*i*=1*c*​*Ni*​(*μi*​−*μ*)(*μi*​−*μ*)*T*

*我们的优化目标可以描述为：，这是一个广义瑞利商，令J=，对J求导我们可以得到：*

*由上式我们可以得到：*

*计算出的k个最大特征值和k个对应特征向量即可得到投影矩阵W*

II．LDA算法模型。

1.将数据集分类，计算每个类的均值；

2.计算类间散度矩阵Sb与类内散度矩阵Sw；

3.构造目标函数（多种不同的目标函数）并对其进行特征分解；

4.取出一定数量的特征向量得到投影矩阵；

5.将测试数据投影到子空间中，使用KNN进行分类（实际问题中）。

1. LDA优化问题。

缺点：① LDA不适合处理非线性问题；② LDA降维最多降到类别数k-1的维数，如果我们降维的维度大于k-1，则不能使用LDA。；③ LDA处理高维数据时，计算代价很大；④ LDA可能过度拟合数据。

优化方法：对于①，我们可以使用带有核函数的LDA升维处理数据；对于③，先用PCA预处理，再使用LDA；对于④，我们可以加入正则项，用于防止过拟合。

1. 证明。

*SW​=∑i=1c​Si*

*SB​=∑i=1c​Ni​(μi​−μ)(μi​−μ)T*

*ST​​=i=1∑c​x∈Xi​∑​(x−μ)(x−μ)T*

*=i=1∑c​x∈Xi​∑​[(x−μi​)+(μi​−μ)][(x−μi​)+(μi​−μ)]T*

*=i=1∑c​x∈Xi​∑​[(x−μi​)(x−μi​)T+(μi​−μ)(x−μi​)T+(x−μi​)(μi​−μ)T+(μi​−μ)(μi​−μ)T]*

*=i=1∑c​x∈Xi​∑​(x−μi​)(x−μi​)T+i=1∑c​x∈Xi​∑​(μi​−μ)(x−μi​)T+i=1∑c​x∈Xi​∑​(x−μi​)(μi​−μ)T+i=1∑c​x∈Xi​∑​(μi​−μ)(μi​−μ)T*

*=SW​+SB​​*

*所以，得证。*

2. 给出LDA的各种等价模型表示(除法的、减法的及其调换位置的等)， 在各数据集（不少于3个）比较PCA与“LDA的各种等价模型与正则模型”的人脸识别精度。

I．数据的导入

利用imread导入数据集，并在导入时将每个人脸拉成列向量；同时分割数据集，将前30%作为测试数据，后70%作为训练数据。（这里只给出AR数据集的导出代码截图，其他的数据集代码类似）

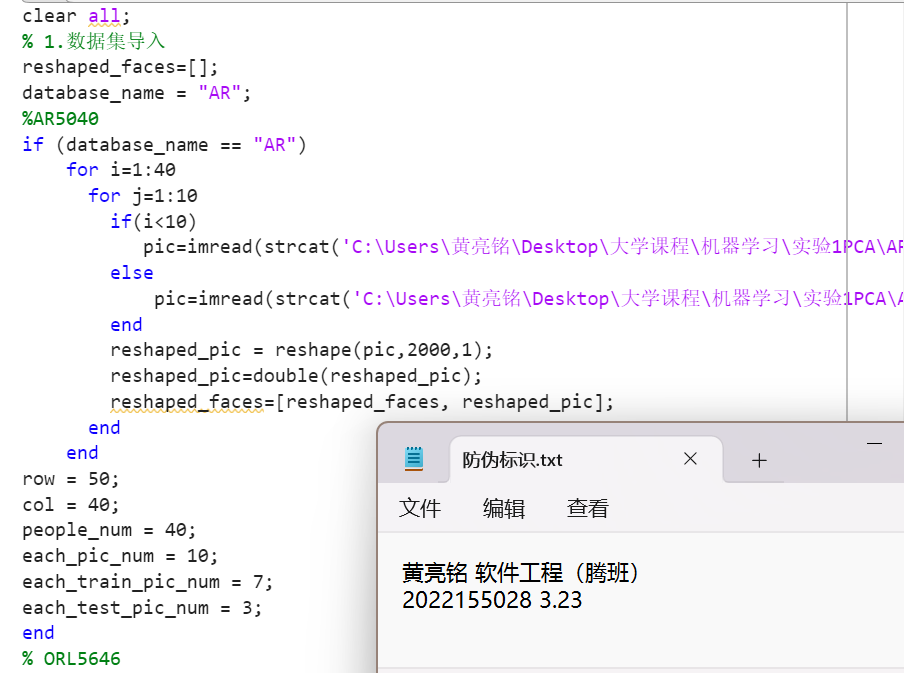


图1

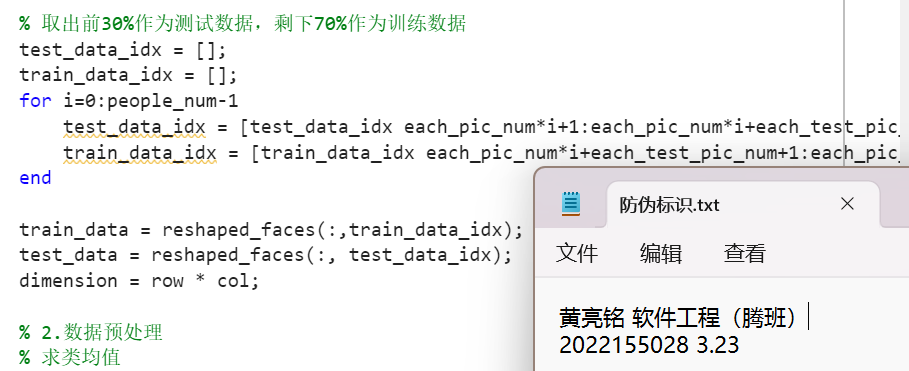


图2

II．数据预处理

计算不同类类内的均值、散度矩阵，以及类间的散度矩阵。

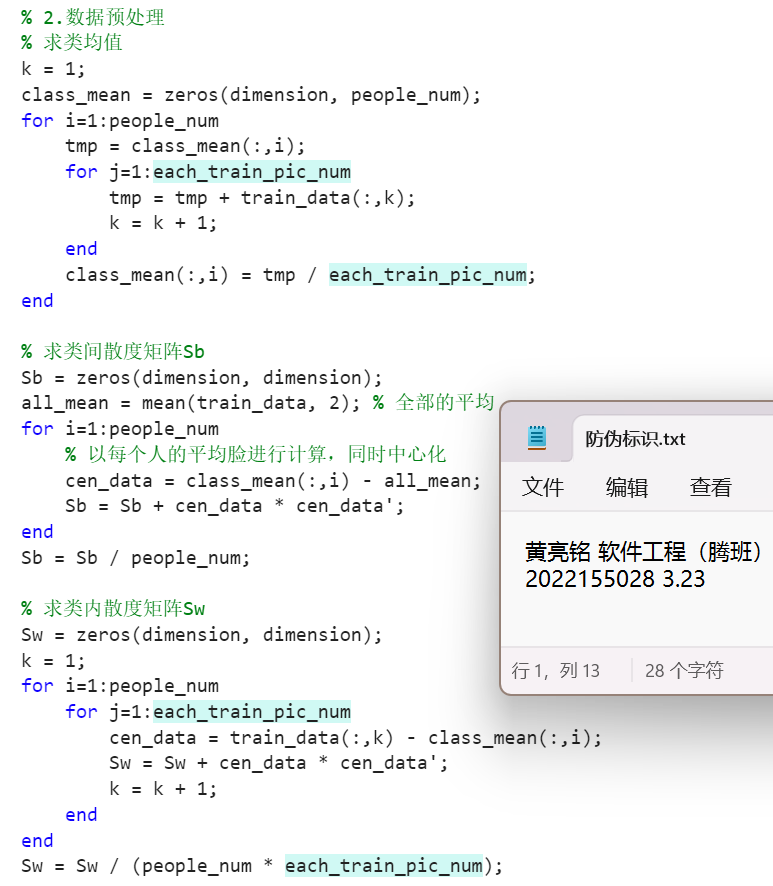


图3

III．构造目标函数，并对目标函数进行特征值分解。

每个目标函数分别对应不同LDA的等价模型以及经典PCA模型。（每一次选择一个目标函数，并将其他的目标函数注释，一共跑6次）



图4

IV．实现人脸识别

使用KNN进行分类预测，实现人脸识别，同时比较不同LDA的等价模型和PCA的人脸识别率。（投影到10、20……160维）

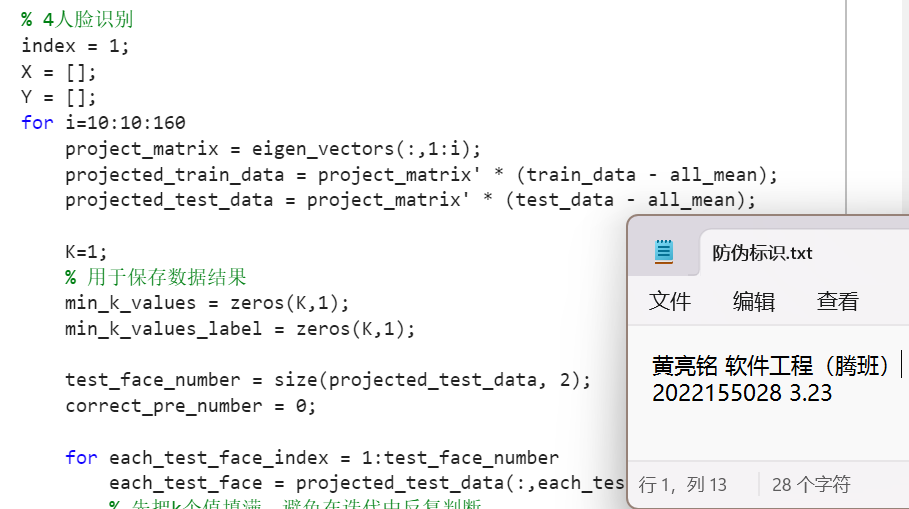


图5

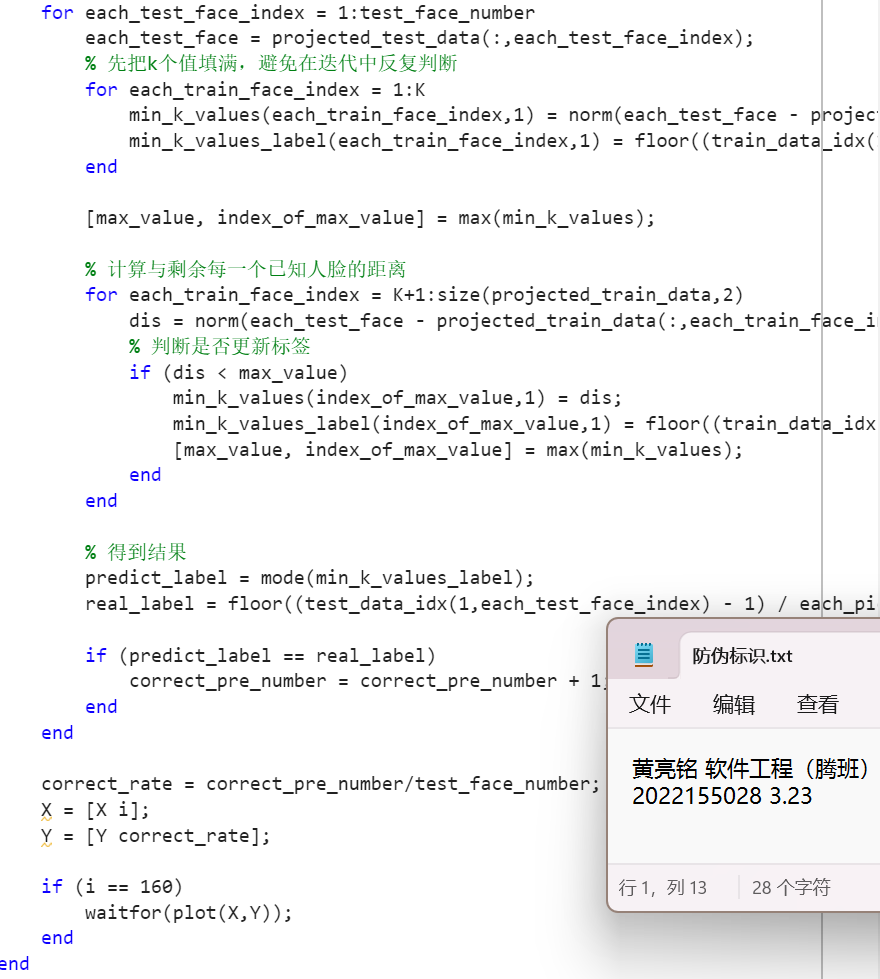


图6

不同的LDA等价模型和PCA的人脸识别率在不同维度下的对比如下图：可以看出经典LDA和正则LDA的识别率较高并且比较稳定，其他LDA等价模型和PCA在低维度识别较低，并且在高维下识别率稳定性较差（除相见形式的LDA）。其中，调换LDA和相除LDA的识别率低于PCA，其他LDA等价模型的识别率均高于PCA。

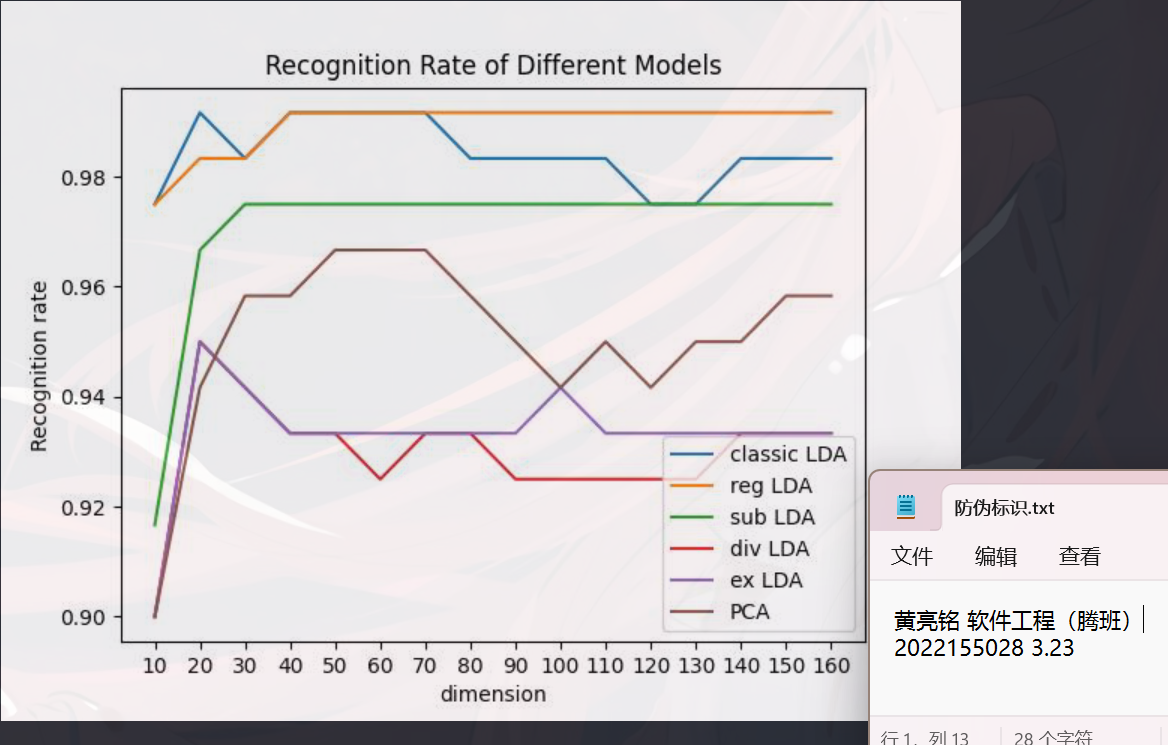


图7

该图使用了python的matplotlib库，代码如下：



图8

3. 比较eigenface与fisherface的不同，并取3个类的图像投影在二维和三维空间中， 并用不同颜色的点表示不同的类，每个类选3个有代表性的点对应的人脸图像显示在该点的边上（用plot命令或imshow），比较PCA与LDA的结果的异不同。

I．LDA与PCA图像降维与可视化对比（二维）

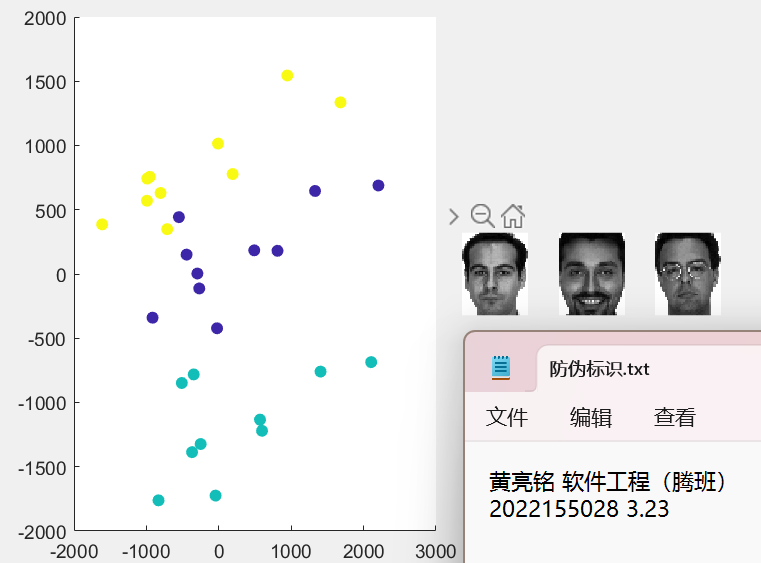
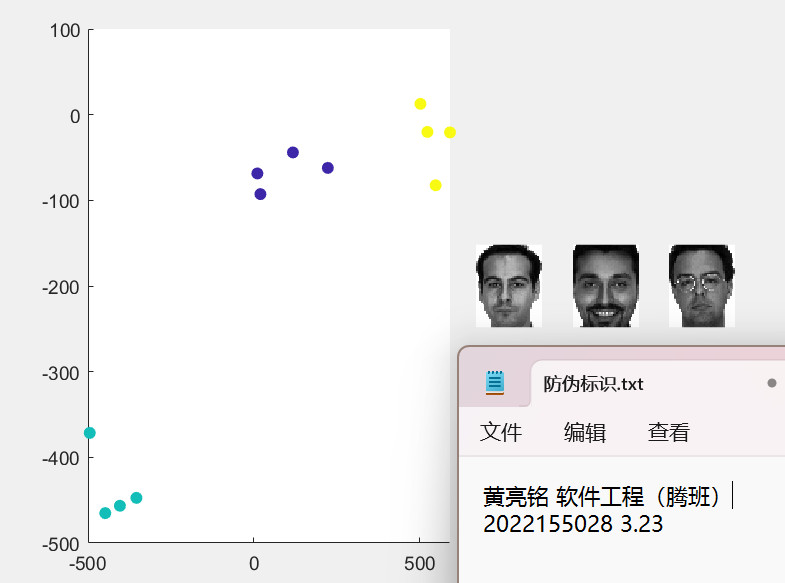


图9：左边LDA，右边PCA

II．LDA与PCA图像降维与可视化对比（三维）

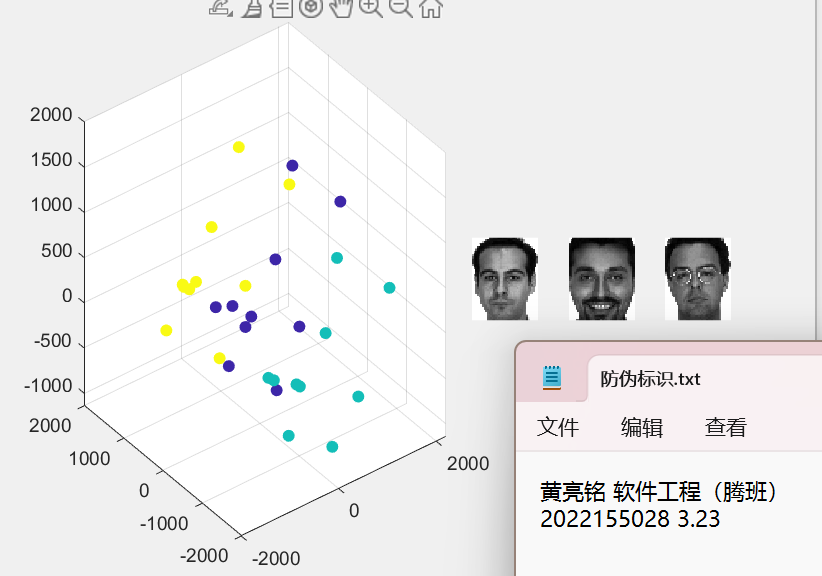
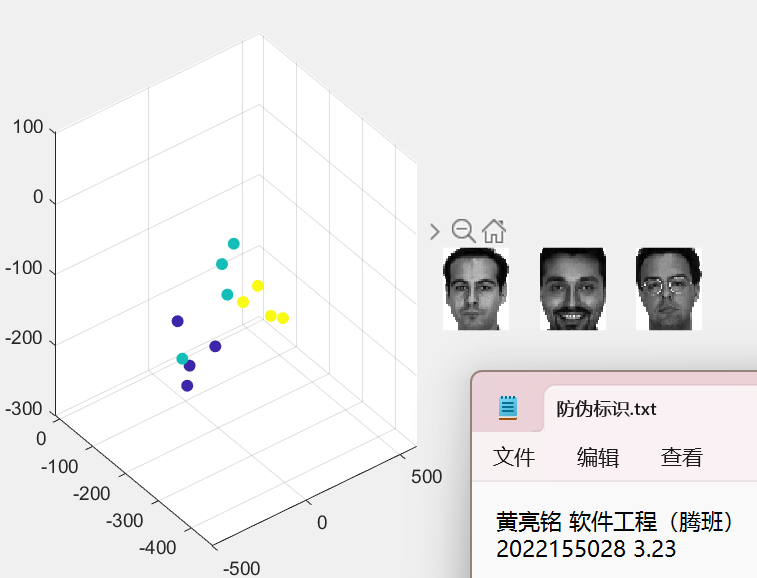


图10：左边LDA，右边PCA

III．比较LDA与PCA结果的异同。

① 无论是在二维还是在三维，都可以明显看出LDA同类图像较为聚集，PCA相对混乱，没有明显规律。由LDA和PCA的算法目的及原理，可以比较出eigenface和fisherface的不同。

② LDA和PCA都是降维技术，并且都是线性变换；

③ LDA的目标是最小类内散度，最大类间散度，而PCA的目标是最小化重构误差；

④ LDA是一种监督学习方法，可以利用类别标签的信息来优化降维过程，而PCA是一种无监督学习方法；

# 四、实验结论或体会

1. 在小范围数据集中，PCA的效果与LDA相近，甚至超过LDA；但是，在大范围数据集中，LDA明显优于PCA，特别是正则化LDA和经典LDA。

2. 通过本次实验，我了解了不同LDA的等价模型及它们在不同数据集下的表现。

3. 通过本次实现，我了解了同为线性降维算法的LDA和PCA之间的相同之处和不同之处。

4. 通过本次实验，我了解了LDA存在的问题，例如：有限投影轴问题、高维数据计算代价大问题等问题。

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。