参赛密码 _____

（由组委会填写）



第十二届“中关村青联杯”全国研究生数学建模竞赛

题 目 数据的多流形结构分析

摘 要：

本文对多流形数据的聚类，子空间样本点相关度函数以及特征选取问题进行了研究。

由于各题中的数据分布各有特色，并不能找到一种普适的模型能够将所有题都纳入其中，因此本文针对不同的题目数据分布的不同，取多种策略建立了不同数学模型，并在一些情况下尝试建立新的模型去解决题目中的问题。

在题 1 中，首先用PCA将高维稀疏数据映射到主成分空间，然后去掉低贡献率主成分以消除冗余性，最后在降维后的子空间中使用$k-means$聚类；

在题 2 中分别使用了空间变换(2a)，局部高斯特征$(Local\ Gaussian\ Distribution)$提取(2b)，基于$knn$距离的$normalized-cut$谱聚类 (2c)以及增强SMMC度量函数(式 20）解决几个多流形线性不可分数据点的聚类问题。

题3a中设计了度量运动点轨迹的$CDI(Correlation\ Distance\ Index)$指数用于区分不同模式的运动点，题 3b 使用主成分分析找到了到了区分照片的光照以及人脸的主成分，并将数据分别变换到这两个主成分，完成了明/暗图像以及不同人脸头像的聚类。

题 4 采用题 2b 一样的LGD特征完成空间中不同模式的点的聚类，先区分曲面和平面，然后再区分出底面与顶面。

本文的创新之处主要有以下 4 点：

1、在 2b，4a 中，提出了一种能够区分点阵局部分布模式的局部高斯特征特征$LGD(Local\ Gaussian\ Distribution)$,用多维高斯分布拟合空间中某邻域内点阵的分布，并以其参数作为点的特征，在多流形聚类中将样本点映射到 LGD 特征空间中完成聚类。

2、在$SMMC$算法的基础上,提出了改进的相关函数(式 20),引入了对称项，对 knn 距离无法包含的非临近同流形点的相关度进行了预估,用于题 2d 的解答。

3、在 3a 十字工件的聚类问题上，本文提出了一种度量两点是否共线的系数$SLI(Share-line\ Index)$。通过分析其他数据点与两点连线的位置关系，来确定这两点共同一直线的概率。实验说明该系数对直线型物件的检测具有很强的鲁棒性。

4、在题 3b 提出了基于运动点相关性的度量指数CDI $(Correlation\ Distance\ Index)$,使用$CDI$指数作为谱聚类的相关函数。

本文综合了$k-means$聚类，谱聚类，主成分分析以及空间转换等数学方法,并在此基础上改进或提出新的模型。在题 2 以及题 3 等多流形聚类问题上，能够充分利用采样自不同流形的数据之间的相关性，找出合适的相关函数加以区分，或寻得合适的变换使得数据线性可分。能够充分利用现有模型解决当前问题，在适当情况下能够创造新的工具解决问题，为后续研究提供一些思路。

关键字：多流形数据结构，聚类分析，主成分分析，相关性矩阵，局部高斯模型，$k-means$

一、问题背景

我们已经进入了一个信息爆炸的时代，海量的数据不断产生，迫切需要对这些大数据进行有效的分析。在大量的数据分析方法之中，几何结构分析是进行数据处理的重要基础,已经被广泛应用在人脸识别、手写体数字识别、图像分类、等模式识别和数据分类问题，以及图象分割、运动分割等计算机视觉问题中。更一般地，对于高维数据的相关性分析、聚类分析等基本问题，结构分析也格外重要。

一个人在不同光照下的人脸图像可以被一个低维子空间近似，由此产生大量的数据降维方法被用来挖掘数据集的低维线性子空间结构，这类方法假设数据集采样于一个线性的欧氏空间。但是，在实际问题中很多数据具备更加复杂的结构。例如运动分割($motion\ segmentation$)中的特征点数据具有多个混合子空间的结构，判断哪些特征点属于同一子空间是这个问题能否有效解决的关键。

针对单一子空间结构假设的后续讨论主要是两个方面，首先是从线性到非线性的扩展，主要的代表性工作包括流形（流形是局部具有欧氏空间性质的空间，欧氏空间就是流形最简单的实例）学习等。流形学习是基于数据均匀采样于一个高维欧氏空间中的低维流形的假设，流形学习试图学习出高维数据样本空间中嵌入的低维子流形，并求出相应的嵌入映射。流形学习很好地解决了具有非线性结构的样本集的特征提取问题。

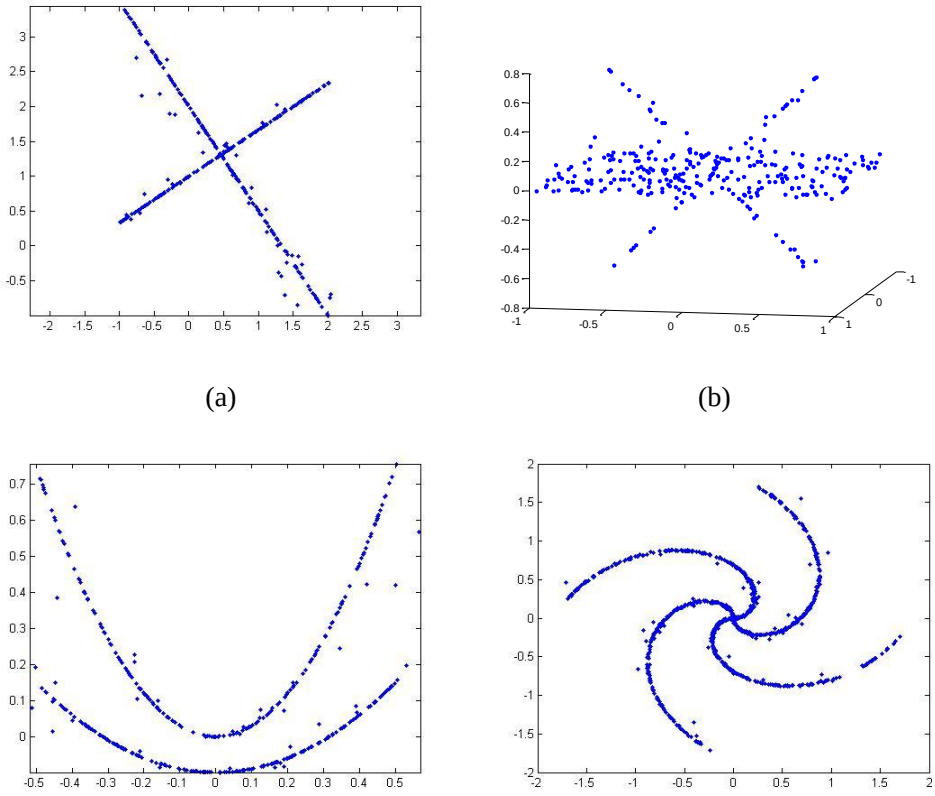
其次是流形或子空间从一个到多个的扩展，即假设数据集采样于多个欧氏空间的混合。子空间聚类（又称为子空间分割，假设数据分布于若干个低维子空间的并）是将数据按某种方式分类到其所属的子空间的过程。通过子空间聚类，可以将来自同一子空间中的数据归为一类，由同类数据又可以提取对应子空间的相关性质。子空间聚类的求解方法有代数方法、迭代方法、统计学方法和基于谱聚类的方法。其中基于谱聚类的方法在近几年较为流行，代表性的基于谱聚类的子空间分割方法包括低秩表示和稀疏表示等。

二、问题重述

本文主要考虑以下问题：

1.当子空间独立时，子空间聚类问题相对容易。附件一中 1.mat 中有一组高维数据（.mat 所存矩阵的每列为一个数据点，以下各题均如此），它采样于两个独立的子空间。请将该组数据分成两类。

2.请处理附件二中四个低维空间中的子空间聚类问题和多流形聚类问题，如图 1 所示。图 1(a)为两条交点不在原点且互相垂直的两条直线，请将其分为两类；图 1(b)为一个平面和两条直线，这是一个不满足独立子空间的关系的例子，请将其分为三类。图 1(c)为两条不相交的二次曲线，请将其分为两类。图 1(d) 为两条相交的螺旋线，请将其分为两类。



(a) (b)

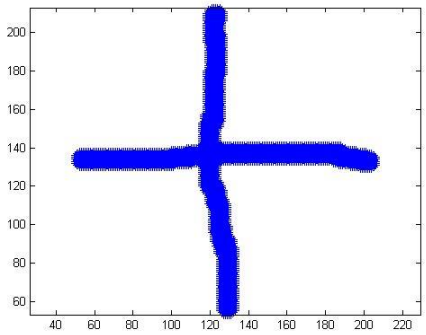
(c) (d)

图 1

3. 请解决以下三个实际应用中的子空间聚类问题，数据见附件三

（a）受实际条件的制约，在工业测量中往往需要非接触测量的方式，视觉重建是一类重要的非接触测量方法。特征提取是视觉重建的一个关键环节，如图 2（a）所示，其中十字便是特征提取环节中处理得到的，十字上的点的位置信息已经提取出来，为了确定十字的中心位置，一个可行的方法是先将十字中的点按照 “横”和“竖”分两类。请使用适当的方法将图 2（a）中十字上的点分成两类。

（b）运动分割是将视频中有着不同运动的物体分开，是动态场景的理解和重构中是不可缺少的一步。基于特征点轨迹的方法是重要的一类运动分割方法，该方法首先利用标准的追踪方法提取视频中不同运动物体的特征点轨迹，之后把场景中不同运动对应的不同特征点轨迹分割出来。已经有文献指出同一运动的特征点轨迹在同一个线性流形上。图 2（b）显示了视频中的一帧，有三个不同运动的特征点轨迹被提取出来保存在了 3b.mat 文件中，请使用适当方法将这些特征点轨迹分成三类。



(a) (b)

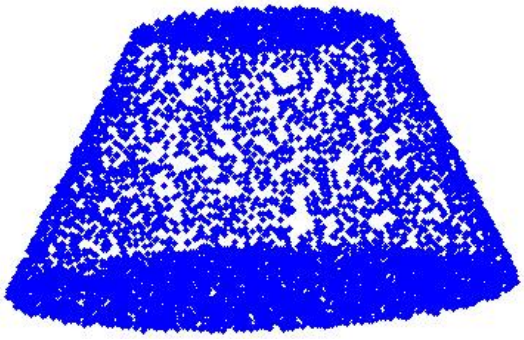
图 2

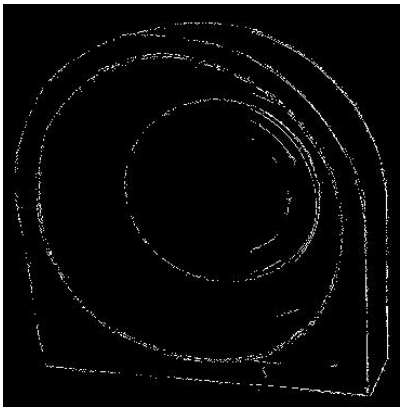
（c）3c.mat 中的数据为两个人在不同光照下的人脸图像共 20 幅（X 变量的每一列为拉成向量的一幅人脸图像），请将这 20 幅图像分成两类。

4. 请作答如下两个实际应用中的多流形聚类问题

图 3(a)分别显示了圆台的点云，请将点按照其所在的面分开(即圆台按照圆台的顶、底、侧面分成三类)。

图 3(b)是机器工件外部边缘轮廓的图像，请将轮廓线中不同的直线和圆弧分类，类数自定。





(a) (b)

图 3

三、模型假设

1、本几何结构分析问题中假设数据分布在多个维数不等的流形上。

2、其特殊情况是数据分布在多个线性子空间上。

四、符号说明

数学符号	说明
$cov(Y_i，Y_j)$	协方差
C	协方差矩阵
X	点数据集
c_k	$K-Means$聚类划分
$J(\mathrm{c_k})$	欧氏距离
LGD	局部高斯度量
$sli(p_i,p_j)$	共线指数
$W[i,j]$	谱聚类相关度矩阵
$CDI(i,j)$	相关距离指数
$G=(V,E)$	无向加权图
$Ncut(A,B)$	子图A,B间的边切集

五、模型建立与求解

5.1(题 1)

解题思路：

采样自稀疏子空间的数据存在冗余，如果能将数据降维消除这种冗余性，然后就能在子空间使用线性方法将数据分类。

解题方法：

主成分分析法$(principle\ component\ analysis, PCA)$是一种常用的线性降维[1-3]方法，它将原始数据投影到一个新的变换空间。主成分分析法[4]是通

过对原始变量的相关矩阵或协方差矩阵内部结构的研究，将多个变量转换为少数几个综合变量即主成分，从而达到降维目的一种线性降维方法。这些主成分能够反映原始变量的绝大部分信息，它们通常表示为原始变量的线性组合。

设 $X=(X_1,X_2,\dots,X_n)^T$ 是一个 n 维随机变量，$C=1/(n-1)\sum_{i=1}^{n}(X_i-X)(Xi-X)$T 为样本协方差矩阵。假设存在如下线性变换：

$$\begin{cases} Y_1=a_{11}X_1+a_{21}X_2+\cdots+a_{N1}X_N=a_1^{\mathrm{T}}X \\ Y_2=a_{12}X_1+a_{22}X_2+\cdots+a_{N2}X_N=a_2^{\mathrm{T}}X \\ \cdots \\ Y_N=a_{1N}X_1+a_{2N}X_2+\cdots+a_{NN}X_N=a_N^{\mathrm{T}}X \end{cases} \tag{1}$$

若用Y_1代替原来的n个变量，则要求Y_1尽可能多地反映原来n个变量的信息。而方差$\mathrm{var}(Y_1)$越大则表示Y_1包含的信息越多，因此要求最大化$\mathrm{var}(Y_1)$，同时限定$a_1^{\mathrm{T}}a_1=1$以消除方差最大值的不确定性。根据上述条件易求得$\mathrm{var}(Y_1)=a_1^{\mathrm{T}}Ca_1$，因此，求解方差$\mathrm{var}(Y_1)$最大问题可转换为在约束$a_1^{\mathrm{T}}a_1=1$下求以下最优问题：

$$\begin{cases} \max a_1^{\mathrm{T}}Ca_1 \\ \text{s.t. } a_1^{\mathrm{T}}a_1=1 \end{cases} \tag{2}$$

通过拉格朗日乘子法求解，有$Ca_1=\lambda a_1$ 。设$\lambda=\lambda_1$为C的最大特征值，则相应的特征向量a_1即为所求。如果Y_1不能代表n个变量的绝大部分信息，则可以用同样的方法求得Y_2甚至Y_3 、Y_4等。一般地，求X的第i个主成分可通过求C的第i大特征值对应的特征向量得到。为了使它们所含信息互不重叠，通常要求它们相互独立，即

$$cov(Y_i,\ Y_j)=a_i^{\mathrm{T}}Ca_j=0(i\neq j) \tag{3}$$

通过上述方法就可以找到线性变换(1)的一组线性基，从而找到原始变量的一组综合变量（主成分）来代替原始变量。在实际应用中通常不会使用所有n个主成分，而选取$m\ll n$个主成分。m的选取根据前m个主成分的累计贡献率$\sum_{i=1}^{m}\lambda_i/\sum_{j=1}^{m}\lambda_j$来选取。

我们将原始数据做PCA变换，发现部分主成分的共享值如下：

1.5715	0.1756	0.1432	0.1366	0.1229	0.1142	0.1088	...

从以上数据可以看出，原始数据只在少数的主成分上有贡献值，不妨只保留前面两个主成分，我们将原始数据投影到前两个主成分上，结果如图 4：

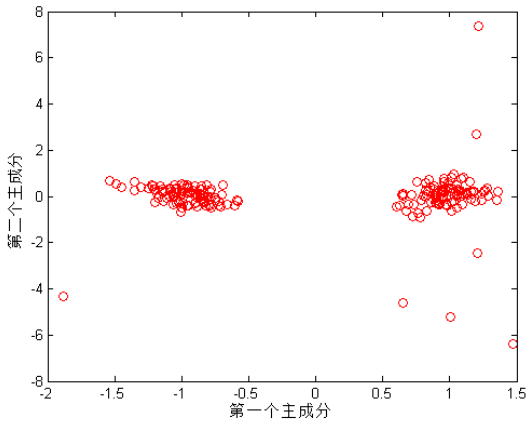


图 4

从图 4 中可以看出投影到最大两个主成分的原始数据已经是线性可分的了，使用$k-means$即可达到分为两类的目的，最终分类结果如图 5 所示，类标签如表 1 所示：

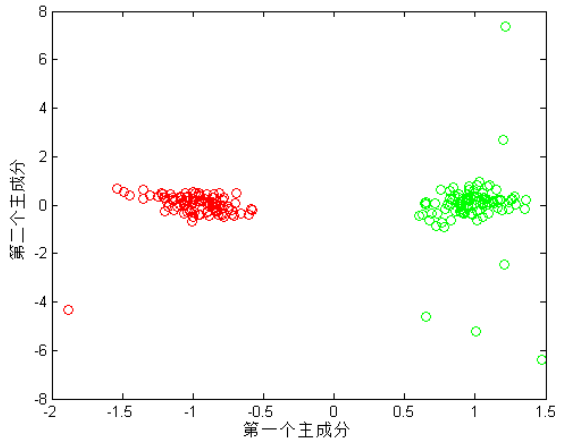


图 5

2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

表 1

$k-means$聚类算法[5]是由$Steinhaus$，$Lloyd, Ball \& Hall$，$McQueen$分别在各自的不同的科学研究领域独立的提出。$k-means$聚类算法被提出来后，在不同的学科领域被广泛研究和应用，并发展出大量不同的改进算法。

对于给定的一个包含n个d维数据点的数据集$X=\{x_1,x_2,\dots,x_n\}$,其中$x_i\epsilon R^d$，以及要生成的数据子集的数目K，$K-Means$ 聚类算法将数据对象组织为K个划分$C=\{c_k，i=1，2，\dots K\}$。每个划分代表一个类c_k，每个类c_k有一个类别中心μ_i。选取欧氏距离作为相似性和距离判断准则，计算该类内各点到聚类中心μ_i的距离平方和：

$$J(\mathrm{c_k})=\sum_{x_i\in c_k}\|x_i-\mu_k\|^2 \tag{4}$$

聚类目标是使各类总的距离平方和$J(C)=\sum_{k=1}^{K}J(c_k)$最小。

$$J(C)=\sum_{k=1}^{K}J(c_k)=\sum_{k=1}^{K}\sum_{x_i\in c_k}\|x_i-\mu_k\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}\|x_i-\mu_k\|^2 \tag{5}$$

其中，$d_{ki}=\begin{cases}1，若x_i\in c_i\\0，若x_i\notin c_i\end{cases}$，显然，根据最小二乘法和拉格朗日原理，聚类中心$\mu_k$应该取为类别$c_k$类各数据点的平均值。

$K-means$聚类算法从一个初始的K类别划分开始，然后将各数据点指派到各个类别中，以减小总的距离平方和。因$K-means$聚类算法中总的距离平方和随着类别个数K的增加而趋向于减小（当$K=n$时，$J(C)=0$）。因此，总的距离平方和只能在某个确定的类别个数K下，取得最小值。

5.2.1（题 2a）

解题思路：

本题的数据为二维欧氏空间中的点，呈交叉的直线状，数据本身线性不可分。

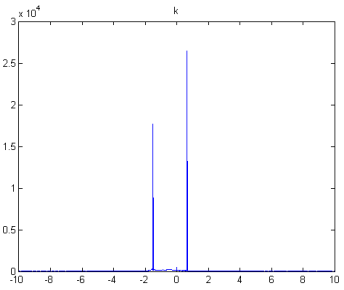
考虑到霍夫变换[6-7]在图像处理中常用于直线检测，经变换后的直线点在新的 k-b 空间中表现为穿过同一点的曲线，过原点的直线通过霍夫变换到 k-b 空间表现为一个点，通过变换到 k-b 空间达到线性可分。

解题方法：

本题数据中存在大量共线点，而多维空间中直线方程为：

$$y=kx+b \tag{6}$$

如果过任意两点求过两点的直线方程并得到方程参数$(k，b)$，n个数据点一共能求出 C_n^2 组参数，然后统计参数组出现的频次直方图，由于共线点之间形成的参数相似度大，而不共线点所形成的直线参数分布则很散，因此找出频次最高的参数组，即为共线数据点所处直线。



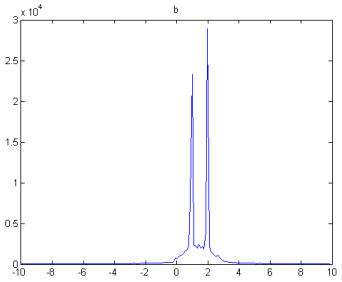


图 6　　图 7

图 6、图 7 分别为k, b值的频次直方图，可以见到有两个峰值，也就是题目中数据所处的两条直线的参数，它们分别是：$k_1 = 0.660, b_1 = 1; k_2 = -1.5, b_2 = 1.99$。

通过上面两组数据可以确定直线交点坐标为$0 = (x，y)$

$$x = (b_1 - b_2)/(k_2 - k_1); y = (b_1 - b_2)/(k_2 - k_1) * k_1 + b_1 \quad (7)$$

最后任意数据点都可以形成一条与交点O的连线，连线方程为$y = kx + b$，每一个点与交点形成的直线方程参数为$k - b$，该数据点在最终的映射空间坐标即为$k - b$。这样在同一直线上的点与交点形成的直线$k - b$参数十分一致，将在$k - b$空间聚拢成簇。$k - b$空间点的分布情况实验结果如图 8 所示：

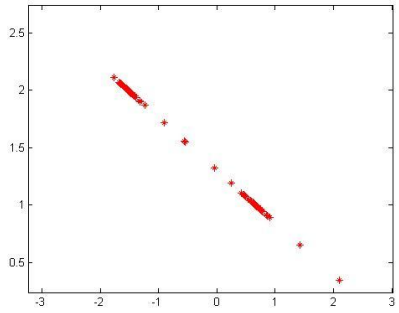


图8

上图$k - b$空间中的点已经线性可分，而且呈直线分布，用线性方法$k - means$即可将其分为两类，最终实验结果如图 9 所示：

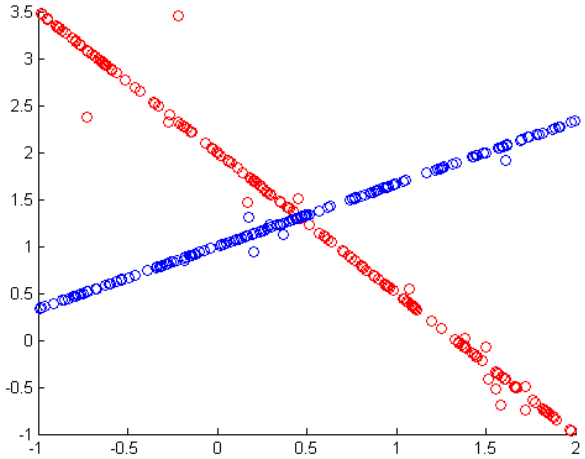


图 9

本题所用方法适用于三维或者更高维采样自相交/非相交直线的样本点的聚类问题，具有较好的泛化能力。

5.2.2(题 2b)

解题思路：

本题的数据为三维欧氏空间中的点，形成两条直线与一个平面。假设两种

直线与一个平面这三种数据点采样自三个不同的流形。由于属于不同流形的点在局部空间中的临近点[8]分布具有显著的差异性。只要设计一种合理的特征能够表现这种显著的差异性，将原始数据映射到特征空间，就能在特征空间进行聚类。

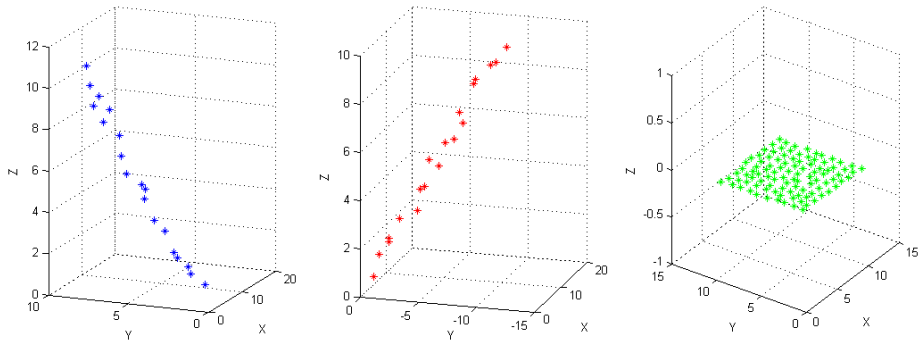


图 10

我们假设某点在一个很小的正方体邻域内的临近点由一个高斯分布产生，然后估得该高斯分布的参数，然后取高斯分布的协方差矩阵为该点的特征进行聚类，把数据点分为三类，如图 10 所示。此方法并未在我们所查阅的参考文献之中，是本文的创新点之一。这种度量局部点分布的特征本文称作 $LGD(Locally\ Gaussian\ Distribution,$局部高斯度量$)$，其灵感来源于混合高斯模型。

解题方法：

在数据点的周围产生一个边长为a的正方体，然后找出所有这个正方体内的数据点n个，加上该数据点本身，一共$n+1$个数据点，在解题思路中我们假设这些点由一个LGD产生，因此要得到这个高斯分布的参数---均值向量和协方差矩阵。

由于均值向量只包含这些点的位置信息，而并没有点之间分布状态的信息，因此被舍弃。而协方差矩阵包含着这些点的x,y,z坐标之间的相关性，是一种重要的特征。具体的，由于协方差矩阵本身具有对称性有冗余，因此只取不重复的 6 个数据作为数据点的特征。

最后我们计算每个数据点的特征[9]，假设有N个数据点，那么就会得到一个$6*N$的矩阵，矩阵的每一列表示一个数据点的特征，这样完成了样本空间到特征空间的映射，然后在特征空间完成聚类，达到分割数据的目的。实验结果如图 11 所示。

由于本题所提出的 LGD 特征仅假设来自不同流形的点所处局部空间的点分布具有差异性，而无其它限制条件，因此具有较强泛化能力。不失一般性，我们在题 4a 中也部分使用了 LGD 特征。

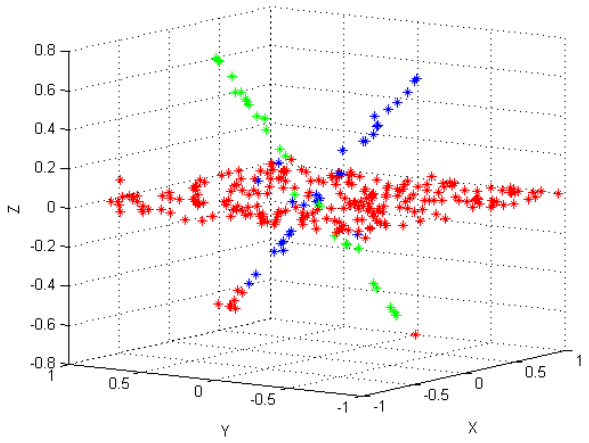


图 11

5.2.3(题 2c)

解题思路：

本题的数据点是一些采样自两条二次曲线的样本点。由于数据之间之间并没有交叉，也就是说不存在这样的一对点：它们采样自不同的流形，但是欧氏距离却比某对采样自同一流形的点小。因此本文采用谱聚类($spectral\ cluster$)进行聚类，并使用knn距离[10]作为相关性度量。

解题方法：

聚类算法的一般原则是类内样本间的相似度大，类间样本间的相似度小。假定将每个数据样本看作图中的顶点y，根据样本问的相似度将顶点间的边赋权重值，就得到一个基于样本相似度的无向加权图：$G=(V,E)$。那么在图G中，我们可将聚类问题转变为如何在图G上的图划分问题。划分的原则是：子图内的连权重最大化和各子图间的边权重最小化。

针对这个问题，Shi和$Malik$[11]提出了基于将图划分为两个子图的$2-way$目标函数$Ncut$:

$$\min Ncut(A,B)=\frac{cut(A,B)}{vol(A)}+\frac{cut(A,B)}{vol(B)}, \tag{8}$$

$$vol(A)=\sum_{i\in A}\sum_{i\sim j}w_{ij} \tag{9}$$

$$cut(A,B)=\sum_{i\in A,j\in B}w_{ij} \tag{10}$$

其中$Ncut(A,B)$是子图A,B间的边，又叫“边切集”。 从式(8)，我们可以看出改进后目标函数不仅满足类间样本间的相似度小，也满足类内样本间的相似度大。

$$asso(A)=\sum_{i\in A,j\in A}\sum_{i\sim j}w_{ij}=vol(A)-cut(A,B) \tag{11}$$

$$\min Ncut(A,B) = \min\left(2 - \frac{asso(A)}{vol(A)} + \frac{asso(B)}{vol(B)}\right) \tag{12}$$

如果考虑同时划分几个子图的话，则基于$k-way$的$Normalized\ cut$目标函数为:

$$Ncut(V_1,\dots,V_k) = \frac{cut(V_1,V_1^c)}{\sum_{i\in V_1}\sum_j W_{ij}} + \frac{cut(V_k,V_k^c)}{\sum_{i\in V_k}\sum_j W_{ij}} \tag{13}$$

假设一无向加权图$G=(V,E)$，其表示形式为一对称矩阵：$W=[W_{ij}]_{n\times n}$ ，其中W_{ij}表示连接顶点i与j的权值。那么该图的$Laplacian$矩阵表示为：

$$L = D - W \tag{14}$$

其中，D为对角阵，$D_{ij}=\sum_{i\sim j} w_{ij}$。

$Laplacian$矩阵是对称半正定矩阵，因此它的所有特征值是实数且是非负的：

$$0=\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$$

如果G是c个连接部件，那么L有c个等于0的特征向量。如果G是连通的，$\lambda_2 \neq 0$，λ_2是G的连接代数值$(Fiedler\ value)$。其对应的特征向量为$Fiedler$向量。

当我们考虑$2-way$划分时，令p是A的划分指示向量：

$$p_j = \begin{cases} -1, j \in A^c \\ 1, j \in A \end{cases} \tag{15}$$

那么：$Cut(A,A^c)=f(p)=cut(A,A^c)=f(p)=\frac{1}{4}\sum_{i,j\in V} w_{ij}\,(p_i-p_j)^2=\frac{1}{2}p^TLp$

考虑约束$x^TW_e = x^TD_e = 0$, 则

$$\min Ncut(A,A^c) = \min \frac{x^{\mathrm{T}}(D-W)x}{x^{\mathrm{T}}Dx} \tag{16}$$

将x放松到连续域$[-1,1]$, 获得minNCut的问题就是：

$$\arg \min_{x} \min \frac{x^{\mathrm{T}}(D-W)x}{x^{\mathrm{T}}Dx} \quad x^{\mathrm{T}}D_e = 0 \tag{17}$$

根据瑞利商原理，式(17)的优化问题等于下列等式的第二最小特征值的求解问题:

$$D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}x = \lambda x \tag{18}$$

对应于第二最小特征值对应的特征向量x_2。则包含了图的划分信息。人们可以根据启发式规则在x_2寻找划分点i，使得值大于等于x_{2i} 的划为一类，而小于x_{2i}的划为一类。

同理，我们可以推理得到 $k-way\ Ncut$目标函数式(13)的最优解在式(18)的是个最小特征值对应的特征向量所组成的子空间上。

本题中我们使用knn距离作为$Ncut$图中定点间的权重值。knn距离在谱聚类中是一种常用的相关度度量方法，对于每一个数据点，knn将离数据点最近的k个其他数据点与它的权重置为 1，其他的点与改点的权重为 0。

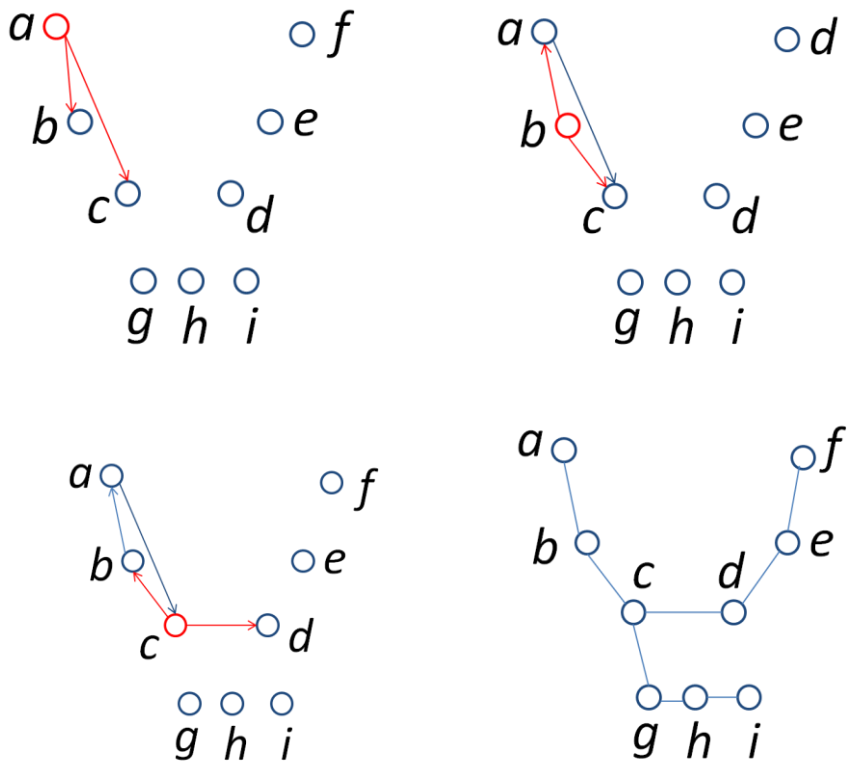


图 12

如图 12 所示，有$\{a,b,c,d,e,f\}$六个节点，取$K=2$。从a点开始

1、a的最近两点是b和c，因此$a-b$，$a-c$的相关度为 1(连线表示相关度)。

2、b的最近两点是a和c，因此$b-a$，$b-c$的相关度为 1。

遍历所有节点，形成图$(graph)$。

本文使用knn距离作为谱聚类中两节点的相关度，使用$normalized-cut$将图分为最优子图，实现聚类目的，实验结果如图 13 所示。

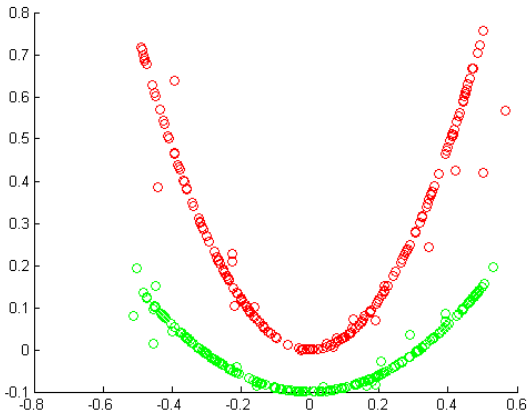


图 13

5.2.4(题 2d)

解题思路：

本题的数据点采样自两个交叉($intersected$)的螺旋线。相比 5.2.2 而言，存在这样的一对点：他们采样自不同的流形[12]，而欧氏距离却很近。本题如果仍然使用谱聚类的方法，那就必须重新设计一种相似度矩阵，利用采样自同意螺旋流形的临近点之间切线方向一致以及中心对称的特点，将聚类问题转化为图分割问题，最后用$normalized-cut$解题。

解题方法：

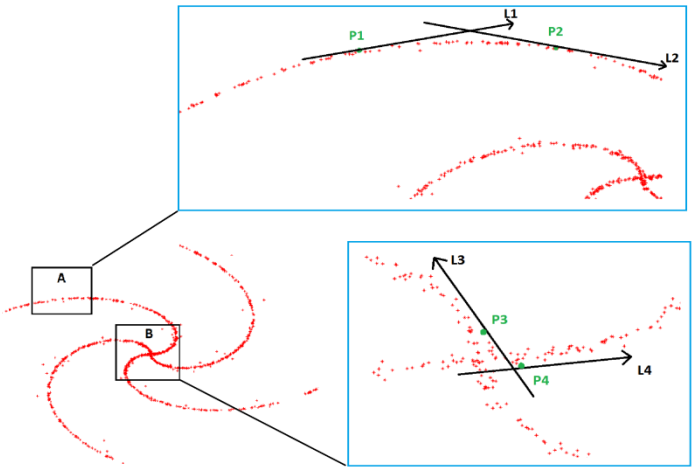


图 14

多流形谱聚类(SMMC)认为，给定一组未标记的数据点$X=\{x_i\in\mathcal{R}^D, i=1,2,\dots,N\}$，从$k>1$个不同的光滑流形$\{\Omega_{\mathrm{j}}\in\mathcal{R}^{\mathrm{D}},\mathrm{j}=1,2,\dots,\mathrm{k}\}$, 其中的一些点可能彼此相交，流形聚类[13]的目的是向每个样本分配给它所属的流形。

如图 14 所示，A区域中的点$P1$，$P2$以及B区域中的点$P3$，$P4$他们之间的欧式距离都很近，但明显他们来自不同的流形，因此一般基于欧式距离的相关度不能区别他们。但是B区域中的$P3$，$P4$点他们的切线方向一致性很低，在欧氏距离的基础上加上切线方向的度量，就能够将$P3$，$P4$点区别开来，这也是$SMMC$算法的基本思路。

$SMMC$设计了如下相关度函数：

$$w_{ij}=\begin{cases}\left(\prod_{l=1}^{d}\cos(\theta_l)\right)^{o} & 若x_i\in knn(\mathrm{x}_j)或\mathrm{x}_j\in knn(\mathrm{x}_i),\\ 0 & 其他\end{cases}\tag{19}$$

上式在knn的基础上增加了临近点切线方向相似度度量，这种相关度函数没有解决knn距离的一个缺点，就是距离比较远但来自同一流形的数据点的相关度在knn中已经被置零，$SMMC$无法重新获得这些点的相关度[14]。

考虑到双曲线在空间中呈对称结构，远距离点如果具有对称性也极有可能来自同一流形，为了knn中丢失的远距离点的相关性重新捕获，设计了如下的相关函数：

$$w_{ij}=\begin{cases}\left(\prod_{l=1}^{d}\cos(\theta_l)\right)^{o} & 若x_i\in knn(\mathrm{x}_j)或\mathrm{x}_j\in knn(\mathrm{x}_i)\\ e^{-|l_i-l_j|}\times cos\,\theta_{ij} & 其他\end{cases}\tag{20}$$

上式中l_i为i点到对称中心距离，θ_{ij}为j，i点分别到对称中心连线的夹角。如图 15 所示，当两点距对称中心距离相近($|l_i - l_j|$)，切线夹角(θ_{ij})很小，也即对称性比较大时，被置较大相关度。

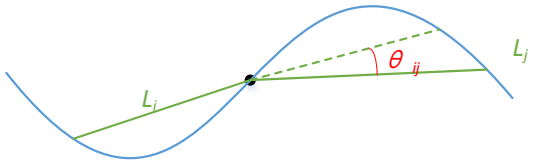


图 15

本文设计的新相关函数经实验证明，适合针对对称的光滑曲线的检测，在本题数据的表现要优于$SMMC$。效果如图 16 所示。

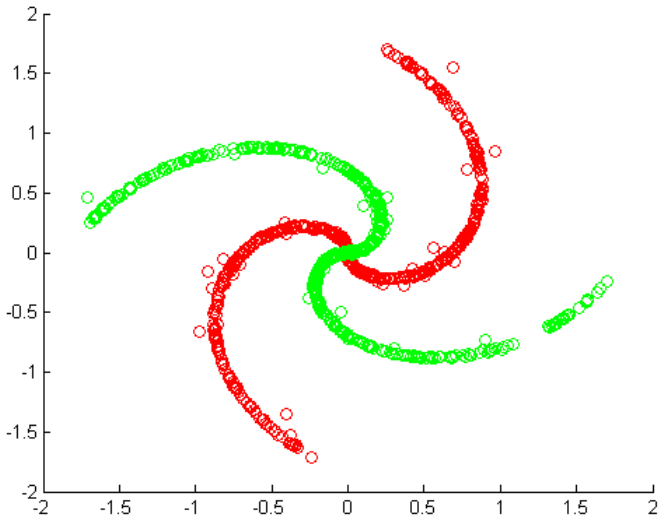


图 16

5.3.1（题 3a）

解题思路：

本题的数据是略有弯曲的相交直线，要将它们分为两类。由于两条相交直线垂直度很高，属于不同直线的点形成的高斯分布在x, y方向形成的方差具有显著差异，因此适合用混合高斯模型，假设不同直线上的点由两个高斯分布产生，直接预估产生两条直线的点的高斯分布的参数，从而达到分类目的[15]。

但是本文采用了另外一种基于$normalized-cut$的更高效的非迭代算法，能够实现快速分类。本文设计了一种度量任意两个数据点是否共线的共线指数($Share-line\ index$)，此方法未在本文参考文献中，是本文的创新点之一。

解题方法：

本题的解题思路仍然是采用基于$normalized-cut$的谱聚类[16]方法，但是本题数据同样存在交叉，因此必须设计一种新的距离度量方法。针对本题数据来自两条直线，因此将两个点处于同一直线作为相关度度量，不失为一种有效手段。

本文设计了如下的共线性度量，我们称之为共线指数，对任意两个数据点p_i，p_j，他们的共线指数$sli(p_i,p_j)$定义如下：

过p_i,p_j作直线l_{p_i,p_j}然后对所有其他数据点p_k计算p_k点到该直线距离d_k，则

$$sli(p_i,p_j)=\sum_{k\in C-\{i,j\}}e^{-\alpha d_k} \tag{21}$$

式中α为常系数，C为所有数据点的集合$C-\{i,j\}$指数据中除了i,j之外的所有点。

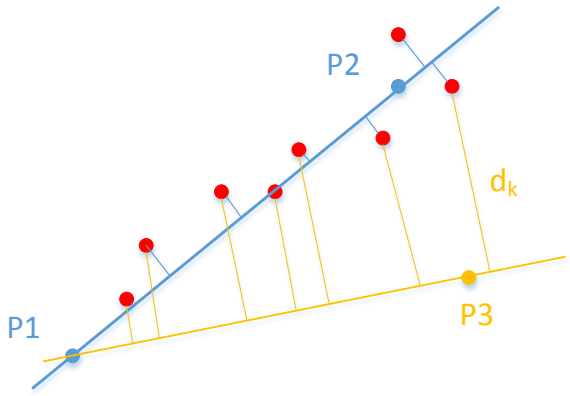


图 17

如图 17 所示，当数据点p_1,p_2共同一直线时，有大量数据点在p_1,p_2连线上或者离p_1,p_2连线很近，因此$sli(p_1,p_2)$相对大，而p_1，p_3不共线，其他数据点离p_1,p_3连线的距离很大，$sli(p_1,p_3)$相对小。

从而谱聚类的关度矩阵 $W[i,j]=sli(p_i,p_j)$，聚类最后结果如图 18 所示：

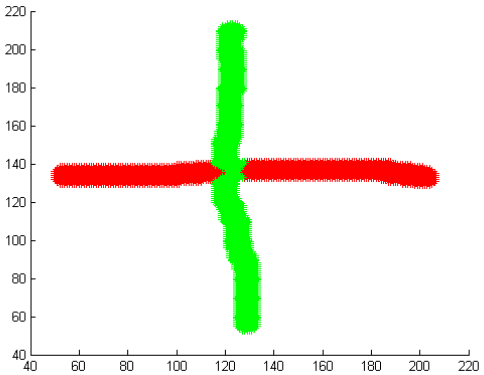


图 18

5.3.2（题 3b）

解题思路：

题设中，指出本题数据为视频中特征点的运动轨迹，要将它们分为三类。运动的特征点处于不同的线性流形[17-20]上，处于不同线性流形上的运动特征

点其运动轨迹的坐标和运动向量具有差异。本文采用了一种度量运动中两点相关性的方法，生成特征点之间的相关性矩阵，对特征空间进行谱聚类。

解题方法：

我们采用相关距离指数$CDI(Correlation\ Distance\ Index)$来描述两个点在运动中的相关性。设$D_{mnk}$表示$m$点和$n$点在$k$时刻的欧氏距离。

CDI的定义：

$$CDI(i,j)=exp\left(-\sum_{k=1}^{n}D_{ijk}\right) \tag{22}$$

其中$n=31$为总运动次数。当两点在多次运动中距离累计越小，其相关性越大。对数据进行谱聚类，谱聚类的相关矩阵

$$W(i,j)=CDI(i,j) \tag{23}$$

对相关矩阵进行谱聚类，得到三类运动结果。

解题结果如图 19 所示，类别标签如表 2 所示：

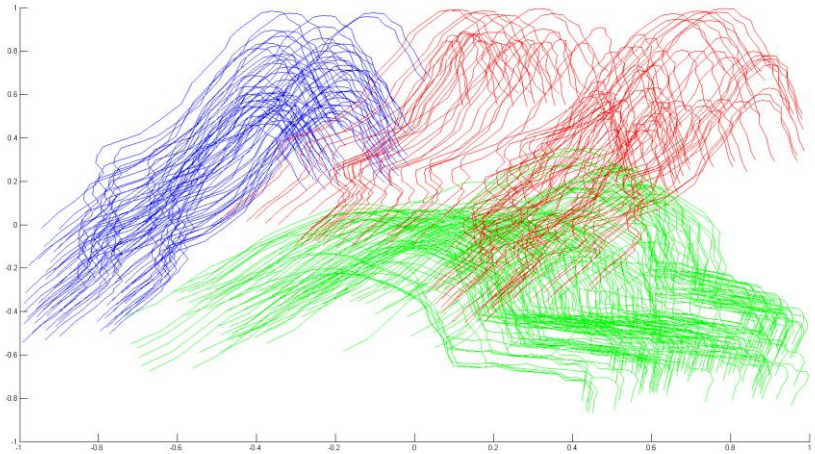


图 19

3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	2	1	2	1	1	1	1	2	1	2	1
1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1
1	1	1	2	1	1	2	2	1	2	1	2	1	1	2	2	1	1	1	1
1	2	1	1	1	1	2	1	1	1	1	1	1	1	1	1	2	0	0	0

表 2

5.3.3（题 3c）

解题思路：

本题给出了 20 张人脸照片，题目要求聚为两类。经可视化发现这 20 幅图片在两个因素上有明显的区分度，可以分为两类：

1　图片有的有光照，有的没有光照

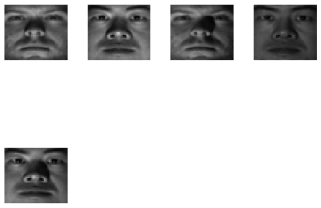
2　图片来自两个不同的人脸

解题方法：

解题思路中已经阐明，照片在两个因素上有明显的区分度，翻译成数学语言就是照片数据在某两个特征维度上具有很强的差异。因此很容易想到使用主成分分析法，将照片数据视为 2016 维的 20 个数据点。经过PCA分析，结果符合预期：数据在前恰好两个主成分上有较大贡献率，之后的主成分贡献率接近于 0[21-23]。

将原始数据投影到第一个主成分上，然后用$k-means$聚类，结果如图$20(a)$、图$20(b)$所示：

图$20(a)$



图$20(b)$

然后将原始数据投影到第二个主成分上，得到的分类结果如图$21(a)$、图$21(b)$所示，类别标签如表 3 所示：

图$21(a)$

图$21(b)$

1	1	1	1	1	2	2	2	2	2	1	1	1	1	1	2	2	2	2	2

表 3

由于在光照上差异的显著性要强于不同人脸照片之间的差异，因此投影到第一个主成分上，结果按照片明暗将照片聚为两类；投影到第二个主成分上则按人的不同将人脸照片聚为两类。实验结果与理论分析以及视觉体验相符合。

5.4.1（题 4a）

解题思路：

本题的数据为三维欧氏空间中的点，空间分布图形为圆台型（两个平面和一个环形曲面）。假设这三种空间面上的采样自不同的流形，本文在 5.2.2 中探讨过，不同流形的点在局部空间邻域点的分布具有显著差异性。

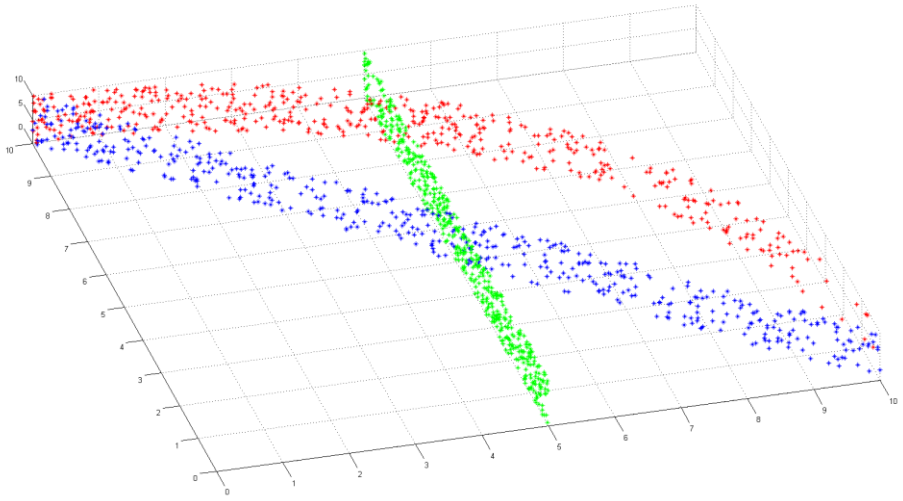


图22

如图 22 所示，假设局部空间中的点集由一个三维高斯分布产生，那么曲面的局部高斯分布参数（均值向量，协方差矩阵）与平面有显著差异(5.2.2 中也有说明)。因此本文分两步解此题，第一步分使用LGD区分出曲面与平面，然后区分两个平行的平面。

解题方法：

假设数据点总数为N。首先，任意一个数据点P，取周围边长为a的正方体A内的所有数据点n个，提取这个局部空间A的LGD特征作为点P的特征。在 5.2.2 中已经说明，LGD特征能够抓取局部空间点的分布特征，利于将周围空间分布不同的点区别开。提取所有N个数据点的LGD特征，将数据转换到特征空间。

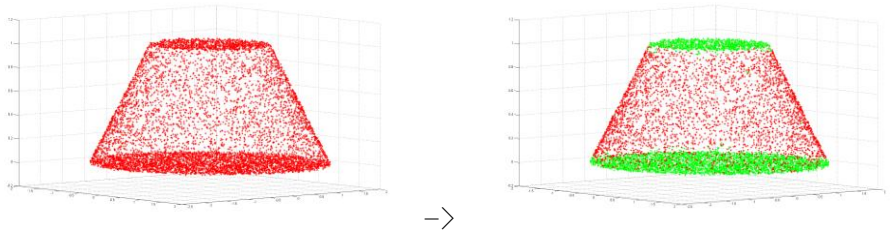
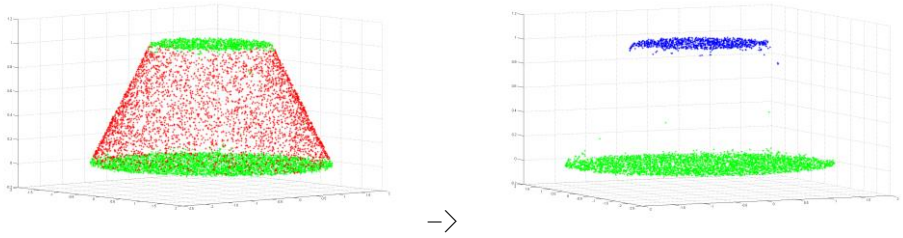


图23(a)　　图23(b)

分别计算两组数据的LGD特征，已知平面的LGD特征拥有更小的方差，提取出处于两个平面上的的数据点M个。已知两个平面上的点位置没有交叉，不存在一对采自不同流形的点，其欧氏距离比某对采样自统一流形的点小，利用谱

聚类对M个点的欧氏距离进行聚类，效果如图 24 所示。



图$24(a)$　　图$24(b)$

最终实验结果如图 25 所示：

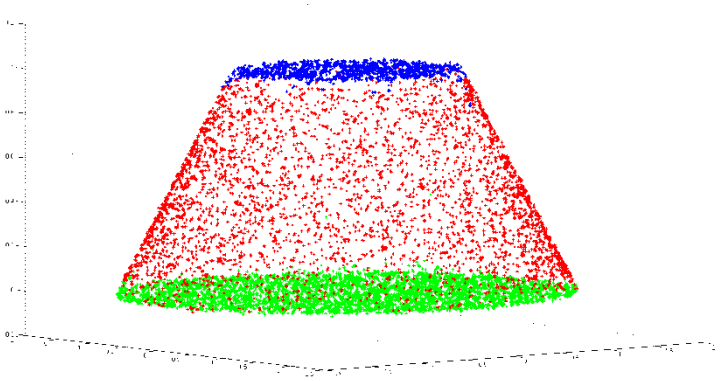


图 25

六、模型评价

题 1：*PCA* 分析必须满足线性子空间条件，如果数据具有更复杂的结构，*PCA* 降维之后信息会丢失，影响聚类结果。但满足线性子空间条件之后 *PCA* 能将高维数据降低至很低的维度。

题 2：针对题 2 中数据的分布我们使用了不同的模型。例如 2a，霍夫变换只能用于直线检测，无法解决曲线检测问题。但我们的模型可以推广到高维，高维空间中的交叉直线问题我们的模型仍然有效。2b 中所提出的 *LGD* 特征具有较好的泛化能力，在其它面面相交等情况下也能表现良好的性能。但丢失了全局信息，例如 2b 的结果图中有部分直线上的点局部表现为平面点的分布，就被 *LGD* 分到了平面那一类。2c 中基于 *knn* 的 *N-cut* 算法的局限在于无法处理交叉的数据，例如 2d 的情况。但它的优势在于无论线条呈何种形态，曲线、直线、高维空间或者二维平面都能将同一线条上的点割为一类。2d 中的 *SMMC* 模型适用于光滑曲线相交情况，不适用于存在激烈角度变化的情况，而我们改进的 *SMMC* 算法在运算速度上超越了 *SMMC*，并在对称曲线的检测上表现出很好的性能，但是损失了一般性。不能处理曲线没有对称性或者曲线存在拐角的情况。

题 3a 中利用 *PCA* 的一个主成分成功将不同人脸的照片分开，方法较为简单运算迅速，而且理论可解释性也很强。但是利用了人的先验，因为经过观察发现光照影响因素最大，而不同人脸的影响其次，因此使用了贡献率第二大的主成分。对于新的未知的数据其性能不能收到保证。

题 3b 以及题 4a 的模型不一一赘述，但与题 2 中各模型相似，那就是在具体领域具有高效率和高效能，且在同一领域泛化性能也不错，但是在其它领域泛化能力一般。

本文所采用（提出）的解题模型对于细分领域的数据处理问题均表现出较好的效果，但有个共同的短处，就是在领域见泛化能力较差，且利用了较多的人的先验，难以用于其它流形数据的处理。但数据本身就是多样化的，处理的方式也应该多样化。针对不同结构的数据应采用不同的算法，算法在细分领域的性能是首先考虑的，然后兼顾泛化能力。

近年来有监督学习算法例如图像领域的 *CNN*(卷积神经网络)，自然语言处理领域的递归神经网络(*rCNN*)在数据处理领域十分流行，基于大量训练样本的有监督学习模型具有非常强的泛化能力和数据处理准确率。如果能有效结合有监督学习与无监督学习方法，那么一定会在性能与泛化能力之间同时提升。

最后，由于代码较长，已于论文名作为附件发送至 B 题指定邮箱。

六、参考文献

[1]吴晓婷,闫德勤. 数据降维方法分析与研究[J]. 计算机应用研究,2009,08:2832-2835.
[2]胡洁. 高维数据特征降维研究综述[J]. 计算机应用研究,2008,09:2601-2606.
[3]陈黎飞. 高维数据的聚类方法研究与应用[D].厦门大学,2008.
[4]王卫卫,李小平,冯象初,王斯琪. 稀疏子空间聚类综述[J]. 自动化学报,2015,08:1373-1384.
[5]王千,王成,冯振元,叶金凤. K-means 聚类算法研究综述[J]. 电子设计工程,2012,07:21-24.
[6]姚磊. 基于 Hough 变换和不变矩的图像模式识别技术研究[D].燕山大学,2006.
[7]曾接贤,张桂梅,储珺,鲁宇明. 霍夫变换与最小二乘法相结合的直线拟合[J]. 南昌航空工业学院学报(自然科学版),2003,04:9-13+40.
[8]李涛,王卫卫,翟栋,贾西西. 图像分割的加权稀疏子空间聚类方法[J]. 系统工程与电子技术,2014,03:580-585.
[9]刘建华. 基于隐空间的低秩稀疏子空间聚类[J]. 西北师范大学学报(自然科学版),2015,03:49-53.
[10]刘应东,牛惠民. 基于 k-最近邻图的小样本 KNN 分类算法[J]. 计算机工程,2011,09:198-200.
[11] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Transactions Pattern Analysis Machine Intelligence, 22(8):888–905, 2000.
[12]雷迎科. 流形学习算法及其应用研究[D].中国科学技术大学,2011.
[13]闫妍. 子空间聚类改进方法研究[D].大连理工大学,2008.
[14]单世民,闫妍,张宪超. 基于 k 最相似聚类的子空间聚类算法[J]. 计算机工程,2009,14:4-6.
[15]蔡晓妍,戴冠中,杨黎斌. 谱聚类算法综述[J]. 计算机科学,2008,07:14-18.
[16]张亚平. 谱聚类算法及其应用研究[D].中北大学,2014.
[17] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(2):218–233, 2003.
[18] R. Vidal. Subspace clustering. IEEE Signal Processing Magazine, 28(2):52–68, 2011.
[19] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):171–184, 2013.
[20] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications.IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11):2765–2781, 2013.
[21] Y. Wang, Y. Jiang, Y. Wu, and Z. Zhou. Spectral clustering on

multiple manifolds. IEEE Transactions on Neural Networks, 22(7):1149－1161, 2011.
[22] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, Multi-task low rank affinity pursuit for image segmentation, ICCV, 2011.
[23] C. Lang, G. Liu, J. Yu, and S. Yan, Saliency detection by multitask sparsity pursuit, IEEE Transactions on Image Processing, 21(3): 1327－1338, 2012.