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

学　校	北京航空航天大学
参赛队号	B10006005
队员姓名	1. 朱日东
	2. 黄 博
	3. 王敬凯

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 数据的多流形和子空间聚类模型研究

摘 要：

本文通过对子空间聚类问题和多流形聚类问题的分析，针对相应的数据结构建立了多种聚类数学模型：K 均值聚类、谱聚类(Spectral Clustering, SC)、基于 CVX 的稀疏子空间聚类(Sparse Subspace Clustering, SSC-CVX) 、基于交替方向乘子法的稀疏子空间聚类(SSC-ADMM)、谱曲率聚类(Spectral Curvature Clustering, SCC)、稀疏流形聚类(Sparse Manifold Clustering and Embedding, SMCE)和谱多流形聚类(Spectral Multi-Manifold Clustering, SMMC)模型，并综合运用这些模型完成了数据的分类，并且实现的聚类效果是非常可观的。

第一问中，数据在两个独立的子空间，聚类相对容易，因此我们首先采用经典的 K 均值聚类与 SC 进行聚类，得出分类结果。为一步验证分类结果，我们采用 SSC-CVX、SSC-ADMM、SCC、SMCE 四个模型分别对数据进行分类，实验结果表明所有算法均可得到相同的分类结果：第 1 个数据到第 40 个数据以及第 141 个数据到第 200 个数据属于第一类，第 41 个数据到第 140 个数据属第二类。

第二问中，要解决四个低维空间中的子空间聚类问题和多流形聚类问题。本文采用 SMMC, SCC, SMCE 三个模型对四组数据分别进行分类。其中通过选取合适的参数，SMMC 非常好的对四组数据进行了分类，得到了题目要求的结果；SCC 对(a),(b)中的数据进行了很好的分类，得到了题目要求的结果；SMCE 对(c)中的数据进行了很好的分类，得到了题目要求的结果。

第三问中，(a)的数据在分布上与第二问中(a)的数据具有一定的相似性，因此对该数据采用 SCC 和 SMMC 模型进行聚类，两种方法都实现了题目要求的分类格式，将数据按照“横”和“竖”分两类。(b)中采用 SSC-ADMM 模型对运动的特征点轨迹进行分类。本文首先采用此算法进行分类。又已知，同一运

动的特征点轨迹在同一个线性流形上。所以我们也尝试采用流形聚类方法：SMCE 和 SMMC 模型对运动特征轨迹分类。为了验证和分析分类结果，我们还采用了 K 均值、SC 模型、SCC 模型进行聚类。实验结果表明，以上方法都一致的将第 1 个数据和第 138 数据分为同一类，将其余的数据归于其它两类。在第 267 个数据，第 275 个数据，第 276 个数据与第 297 个数据的分类中存在差异。我们分析这两类运动的特征点轨迹可能较为相似，在移动中，存在相交，因而在这两类运动的相交边界处的特征点轨迹可能会出现分类上的不同。(c)中采用 SSC-ADMM 模型对人脸进行分类。我们仍先采用此算法对人脸数据进行分类。为验证分类结果，我们采用适宜处理高维数据分类的 SCC、SMCE、SSC-CVX 模型进行分类实验。通过实验可以发现这几种方法均可得到与 SSC-ADMM 算法一致的分类结果：第 1 个到第 5 个人脸数据以及第 11 个到第 15 个人脸数据属于一个人，其余的属于另一个人。

第四问中，(a)问题中圆台数据属于多流形结构，所以我们采用 SMMC 算法对圆台数据进行分类，调整参数的选取，可将圆台的顶、底、侧面分成三类。(b)机器工件外部边缘轮廓的图像数据用 SMCE 算法分成三类和四类，得到较好的聚类效果。

关键词：谱多流形聚类　稀疏流形聚类与嵌入　稀疏子空间聚类
谱曲率聚类　　谱聚类

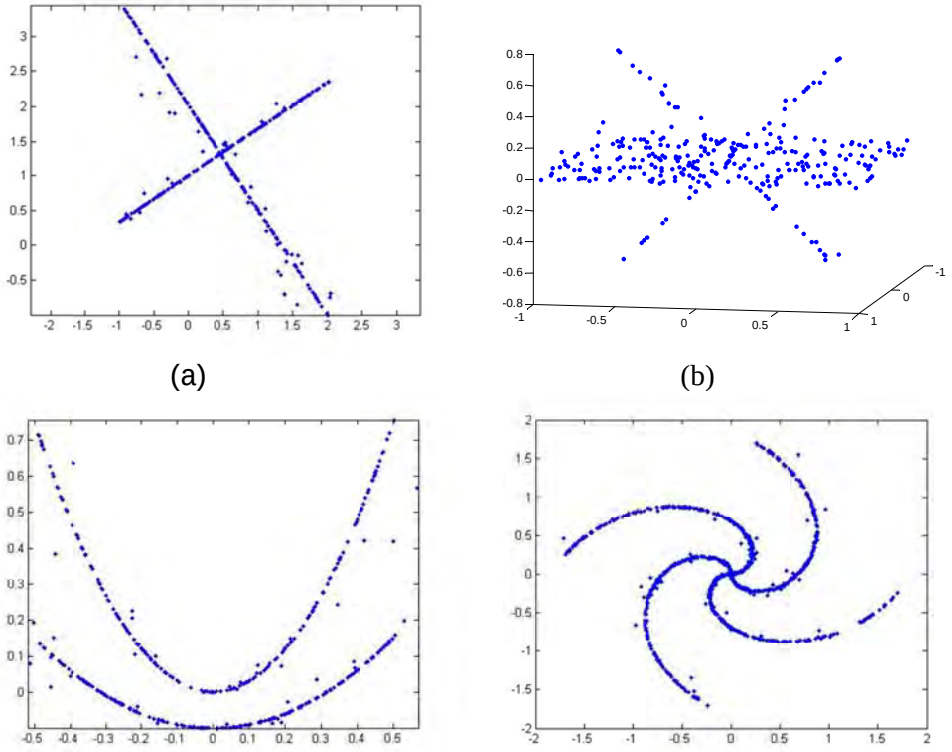
一、问题重述

随着计算机技术和互联网的飞速发展，我们已经进入高维和海量数据的时代，迫切需要对这些大数据进行有效的分析，以至数据的分析和处理方法成为了诸多问题成功解决的关键，涌现出了大量的数据分析方法。几何结构分析是进行数据处理的重要方法,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，此外，如何快速有效的在大规模高维数据上求解各种数学模型，也成为目前数据分析与处理领域亟待解决的重要问题。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。而本次建模就是采用数据处理及其应用领域经典的子空间聚类的求解方法，并着重使用了基于谱聚类的多种方法解决了子空间聚类问题和多流行聚类问题。

在题目给定的数据和参考文献下，求解以下 4 个问题：

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)

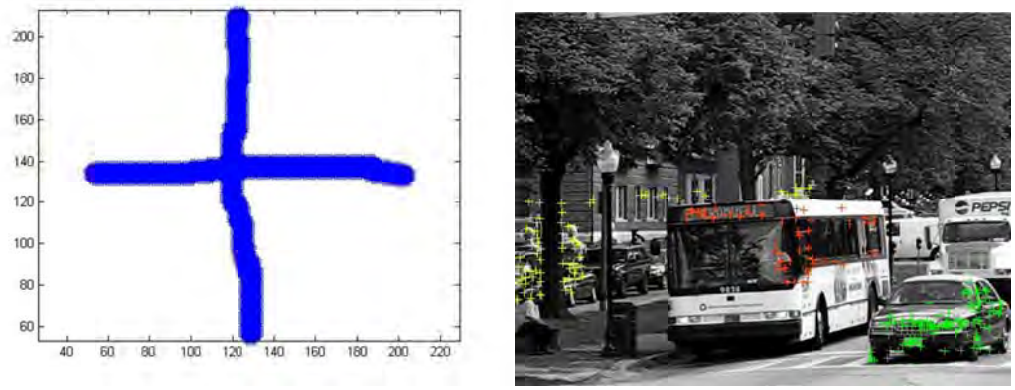
图 1

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 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)

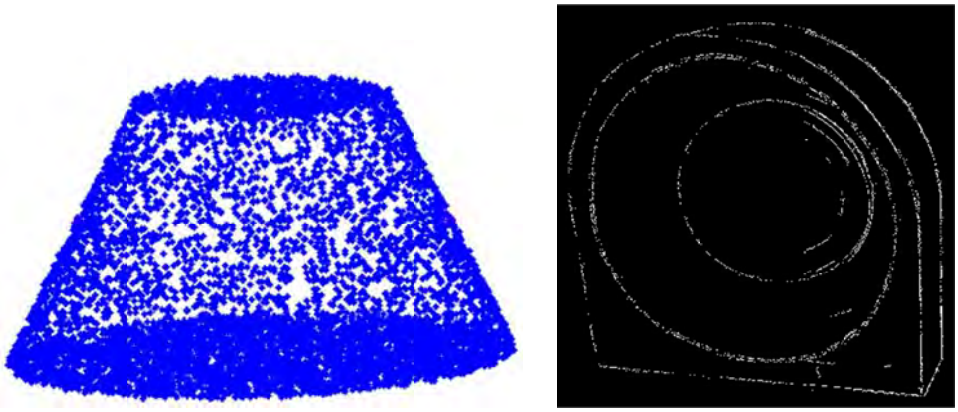
图 2

（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

二、模型假设

1. 假设提供的数据是可信的（只有极小部分受损坏）
2. 假设数据集分布在多个维数不等的流形上
3. 假设数据集服从多项分布
4. 未标记的数据点位于或接近多低维光滑流形

三、符号说明

符号	说明
$X=\{x_1,x_2,\cdots,x_i,\cdots x_n\}$	数据点集合
$J(C)$	聚类结果各类总的距离平方和
$cut(A,B)$	A,B 两个子图的代价函数
$Ncut(A,B)$	A,B 两个子图的割集准则
$assoc(A,A)$	子图 A 内所有顶点连接的权重之和
$q_{ij}=q(\|x_i-x_j\|)$	两个点 x_i,x_j 之间的 Euclidean 距离
w_{ij}	两个点 x_i,x_j 之间的亲和力值
Θ_i	x_i 处的切空间表示
p_{ij}	切空间在 x_i,x_j 两点之间的相似性
$Knn(x)$	x 的 K 个附近邻域
$\{M_l\}_{l=1}^n$	n 个不同的流形

四、问题分析

通过对问题的初步分析可知，本题是要求我们用几何结构方法对数据进行分析处理，而我们知道高维空间的数据往往能够在其低维子空间中进行表示，这样的低维表示对于数据的处理是极有帮助的。而经典的子空间聚类方法恰巧能够准确的在低维空间中表示数据，实现子空间聚类问题的方法有很多，包括代数方法、迭代方法、统计学方法、基于谱聚类的方法。各种方法的理论基础不同，在求解过程上也有很大差异。本文主要采取近几年较为流行的基于谱聚类的多种聚类方法并综合运用得到理想的分类结果。

问题 1: 要求我们对附件一中的数据分成 2 类，由于数据采样于两个独立的子空间，子空间聚类问题相对容易，尝试了 K 均值聚类，SC，SSC 等多种方法进行数据分类，运行结果发现这些方法是合理有效的。

问题 2: 对四个低维空间中子空间聚类问题和多流形聚类问题，由于数据结构性质的变化，简单经典的 K 均值聚类及 SC(谱聚类)方法就无法使用，此时针对问题建立了 SCC、SMCE 与 SMMC 模型，得到理想的分类结果。

问题 3: 分析三个实际应用中的子空间聚类问题，(a)中为确定十字的中心位置可以考虑将十字中的点分成横竖两类，这就与问题 2 中(a)类似。(b)考虑到在文献[5]给出基于 ADMM 的 SCC 模型是一种重要运动的分割方法，所以可以将

其应用到(b)题运动特征点轨迹的分类。由已知，相同运动的特征点轨迹属于同一个流形，进而可用流形聚类方法：SMMC 与 SMCE 对进行数据分类；(c)将数据的每一列看成拉成向量的一幅人脸图像，此时每幅图像即变成一个高维数据。本题即变成对高维数据的分类，因为是在不同光照下的人脸，所以需要提取在不同光照干扰下的人脸特征。此时我们考虑用基于 ADMM 或者 CVT 的 SSC，以及 SCC 方法对人脸数据进行分类。

问题 4：对实际问题中的多流行聚类问题(a)及(b)，这些数据都具有复杂结构且其本身无法使用相互表示的方式，但数据的特征可相互表示，因此基于谱聚类的流形聚类方法仍可处理这种问题。

五、模型建立

5.1 K-Means 聚类算法模型[1]

对于给定的一个包含 n 个 d 维数据点的数据集 $X=\{x_1,x_2,\cdots,x_i,\cdots x_n\}$，其中 $x_i\in R^d$，以及要生成数据子集的数目 K，K-Means 聚类算法将数据对象组织为 K 个划分 $C=\{c_k,i=1,2,\cdots K\}$。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别准则，计算该类内各点到聚类中心 μ_i 的距离平方和

$$J(c_k)=\sum_{x_i\in C_K}\|x_i-\mu_k\|^2 \tag{1}$$

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。

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

其中，$d_{ki}=\begin{cases}1,x_i\in c_i\\0,x_i\notin c_i\end{cases}$，显然根据最小二乘法和拉格朗日原理，聚类中心 μ_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，以减少总的距离平方和。因为 K-Means 聚类算法中总的距离平方和随着类别个数 K 的增加而趋向于减少。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 谱聚类算法模型(SC)[2,3,4,6]

5.2.1 图划分准则

谱聚类算法建立在图论中的谱图理论基础之上，其本质是将聚类问题转化为图的最优划分问题，是一种点对聚类算法，谱聚类算法思想来源于谱图划分理论，假定将每个数据样本看作图中的顶点 V,根据样本间的相似度将顶点间的边 E 赋权重值 W，这样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，就可将聚类问题转化为在图 G 上的图划分问题。基于图论的最优化准则就是使划分成的两个子图内部相似度最大，子图之间的相似度最小，划分结果的好坏直接影响聚类结果的优劣。常见的划分准则有Minimum cut，Average cut，Normalized cut等。

最小割集准则(Minimum cut)：

谱图理论中，将图G划分为A,B两个子图（其中$A\cup B=V, A\cap B=\varnothing$）的代价函数为：

$$cut(A,B)=\sum_{u\in A,v\in B} w(u,v) \tag{3}$$

Wu和Leahy提出最小化上述剪切值来划分图G，这一划分准则被称为最小割集准则。他们用这个准则对一些图像进行分割，并产生了较好的效果，然而该准则容易出现歪斜（即偏性小区域）分割，为了避免切割小集合中出现的不自然的偏见，以下引入解除两分组关联的新方法。

规范割集准则(Normalized cut)：

$$Ncut(A,B)=\frac{cut(A,B)}{assoc(A,V)}+\frac{cut(A,B)}{assoc(B,V)} \tag{4}$$

其中，$assoc(A,V)=\sum_{u\in A,t\in V} w(u,t)$是从A中的节点到图形中的所有节点的总连接。

最小化Ncut函数被称为规范割集准则，该准则不仅能衡量类内样本间的相似程度，也能衡量类间样本间的相异程度。相同情况下，对于一个给定的分割，我们还可以定义一个组内规范关联目标函数（Nassoc）：

$$Nasso(A,B)=\frac{asso(A,A)}{asso(A,V)}+\frac{asso(B,B)}{asso(B,V)} \tag{5}$$

$assoc(A,A)$与$assoc(B,B)$分别是子图A,B内所有顶点之间的连接权重之和。这一无偏函数进一步反映了类内样本间互相连接的紧密程度。同时，可以看出Ncut函数与Nassoc函数是相关的：

$$\begin{aligned} Ncut(A,B)&=\frac{cut(A,B)}{asso(A,V)}+\frac{cut(A,B)}{asso(B,V)}\\ &=\frac{asso(A,V)-asso(A,A)}{asso(A,V)}+\frac{asso(B,V)-asso(B,B)}{asso(B,V)}\\ &=2-(\frac{asso(A,A)}{asso(A,V)}+\frac{asso(B,B)}{asso(B,V)})=2-Nasso(A,B) \end{aligned}$$

因此，最小化Ncut函数等价于最大化Nassoc函数，但通常情况下都是通过最小化Ncut函数获取图的最优划分。

5.2.2 相似矩阵、度矩阵及Laplacian矩阵

由于图划分问题的本质，求图划分准则的最优解是一个NP难问题。一个很好的求解方法是考虑问题的连续放松形式，这样便可以将问题转化成求解相似矩阵或Laplacian矩阵的谱分解，因此将这类算法统称为谱聚类，可以认为谱聚类是对图划分准则的逼近。

根据矩阵通常用W或A表示，有时也成为亲和矩阵。该矩阵定义为：

$$W_{ij}=\exp(-\frac{d(s_i,s_j)}{2\sigma^2}) \tag{6}$$

其中s_i表示每个数据样本点，$d(s_i,s_j)$一般取$\|s_i-s_j\|^2$，σ为事先给定的参数。

将相似矩阵的每行相加，即得到该点的度，以所有度值为对角元素构成的对角矩阵即为度矩阵，度矩阵常用D表示。

Laplacian矩阵分为非规范Laplacian矩阵和规范Laplacian矩阵。非规范Laplacian矩阵表示为$L=D-W$,规范Laplacian矩阵有两种形式，分别是：

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$
$$L_{rw} = D^{-1} L = I - D^{-1} W \tag{7}$$

5.2.3势函数、Fiedler向量及谱

势函数为表示某样本划分归属的指示向量，其定义为：

$$q_i = \begin{cases} 1, \mathrm{i} \in \mathrm{A} \\ 0, \mathrm{i} \in \mathrm{B} \end{cases} \tag{8}$$

若最终势函数中莫样本对应的值为1，则该样本属于集合A，若为0则该样本属于集合B。但实际划分求解得到的结果 q_i 常为0到1之间的实数值，此时可用k均值聚类等方法进一步决定样本的归属。

许多谱聚类算法都将图划分问题转为求解Laplacian矩阵的第二小特征向量问题。这里的第二小特征向量就是第二小最小特征值对应的特征向量，它代表了最佳图划分的一个解（即势函数），把这一特征向量称为Fiedler向量。与特征向量（不一定是Fiedler向量）对应的特征值称为谱。

5.2.4谱聚类算法步骤

根据不同的准则函数及普映射方法，谱聚类算法发展了很多不同的具体算法，具体步骤可以归纳如下：

表5.1谱聚类算法模型

算法1：谱聚类算法（SC）
输入：数据集 S ,聚类个数 k 。
1. 构造亲和矩阵 $A \in R^{n \times n}$ ，其中 $A_{ij} = \exp(-\left\|s_i - s_j\right\|^2 / 2\sigma^2), (i \neq j), A_{ii} = 0$ ；
2. 定义对角矩阵 D ，其对角元 (i,i) 为 A 的第 i 行的和，构建矩阵 $L = D^{-1/2} A D^{-1/2}$ ；
3. 计算出 L 的前 k 个特征值与特征向量，形成矩阵 $X = [x_1 x_2 \cdots x_k] \in R^{n \times k}$ ；
4. 对矩阵 X 的行规范化处理形成矩阵 Y ；
5. 矩阵 Y 的每一行看成 R^k 中一个点，利用k-means或其他经典聚类算法对它们聚成 k 类；
输出： k 个不相交类的分割。

上述步骤是谱聚类算法的一个框架，在具体实现过程中，不同的算法在数据集矩阵Z的表示上存在着不同。例如根据2-way cut的目标函数， $Z = W$ ；根据随机游动关系，则 $Z = D^{-1} W$ 等。划分准则一般为2-way和k-way，本文根据所使用的划分准则，根据谱聚类算法对数据进行了分类。

5.3 谱曲率聚类算法模型(SCC)[8]

5.3.1 极曲率

令 d, D 为满足 $0 \le d \le D$ 的整数，对 $d+2$ 个 $\overline{R}^D$ 中不相交点 $z_1, \cdots, z_{d+2}$ ，用 $V_{d+1}(z_1, \cdots, z_{d+2})$ ， $d+2$ 个点的极曲率定义如下：

$$c_p(z_1, \cdots, z_{d+2}) = diam(\{z_1, \cdots, z_{d+2}\}) \times \sqrt{\sum_{i=1}^{d+2} (p \sin_{z_i}(z_1, \cdots, z_{d+2}))} \tag{9}$$

其中 $diam(S)$ 表示集合 S 的直径，记 $d = 0$ 时极曲率对应 Euclidean 距离。

5.3.2 亲和力张量及其矩阵表示

假设 R^D 数据集 $X=\{x_1,x_2\cdots,x_N\}$ 形成 d-平面集（可有噪音或奇异点），利用极曲率 c_p 和固定常数 $\sigma>0$ 对 $d+2$ 个点 $x_{i_1},\cdots,x_{i_{d+2}}$ 构造如下多路亲和值：

$$A(i_1,\cdots,i_{d+2}) = \begin{cases} e^{-c_p^2(x_{i_1},\cdots x_{i_{d+2}})/(2\sigma^2)}, i_1,\cdots,i_{d+2}; \\ 0, \quad \text{otherwise} \end{cases} \tag{10}$$

5.3.3 K-means 初始化

求解如下优化问题：

$$\{s_1,\cdots,s_K\} = \arg\max_{1\le n_1<\cdots<n_K\le N} \sum_{1\le i<j\le K} \left\|U(n_i,:)-U(n_j,:)\right\|^2 \tag{11}$$

可应用归纳方案解决上述问题，s_1 是选取 N 行中离中心最远的行，即：

$$s_1 = \arg\max_{1\le n\le N} \left\|U(n,:)-\frac{1}{N}\sum_{i=1}^{N}U(i,:)\right\| \tag{12}$$

假设 $\{s_1,\cdots,s_K\}$ 都已选取则有下述定义：

$$s_{k+1} = \arg\max_{\substack{1\le n\le N \\ n\ne s_1,\cdots,s_k}} \sum_{i=1}^{k}\left\|U(s_i,:)-U(n,:)\right\|^2 \tag{13}$$

表5.2谱曲率聚类算法模型

算法2：谱曲率聚类算法（SCC）
输入：数据集 S ,内嵌维数 d ，样本列数量 c 。 1. 数据集 S 的 c 个随机样本子集，每一子集包含 $d+1$ 个不相交点； 2. 用 (9) 在数据集 S 计算每一子集与其他 $N-d-1$ 个点的极曲率，将 $(N-d-1)\bullet c$ 个极曲率形成向量 c ； 3. 对 $q=1:d+1$ 做如下： • 用 (10) 式及 $\sigma=c(N\bullet c/K^q)$ 计算矩阵 A 的 c 个选择的列，用此 c 列形成矩阵 $A_c\in R^{N\times c}$ ； • 计算矩阵 $D=diag\left\{A_c\bullet(A_c'\bullet 1)\right\}$ 并用此标准化 $A_c: A_c^*=D^{-1/2}\bullet A_c$ ； • 处理矩阵 A_c^* 的左上K个奇异向量按列排成矩阵 U ； • 利用K-means方法，利用 (12),(13) 对矩阵 U 的行初始化，分成独立的K类； 输出： k 个不相交类的分割。

5.4 稀疏子空间聚类模型(SSC)

5.4.1 模型准备

稀疏子空间聚类方法，是结合稀疏表示理论，对子空间表示系数进行稀疏约束的一类子空间聚类方法。子空间聚类的最终结果是要将同一子空间的数据归为一类，在子空间相互独立的情况下，属于某一子空间的数据只有这个子空间的基线性组合表示，换句话说这些数据在其他子空间的表示系数为零。因此，对于高位数据而言，数据在低维子空间的表示系数就有稀疏的特性，通过对表示系数稀疏约束的求解，突出了数据表示系数的这种系数特性，进而为数据的正确聚类提供支持。

子空间表示字典的选择：

字典 D 是 S 的一个基矩阵时， D 中的若干列向量可以张成子空间 S_i。 D 的形式将直接影响数据的子空间表示：一方面对于子空间而言，字典 D 并不是唯一的；另一方面，字典 D 确定了子空间的形态，这就意味着数据在任何子空间的表示取决于 D 的构造。

考虑子空间表示问题， $X=DA$ 是要得到每个数据在其所属子空间中的表示系数。若要得到数据矩阵 X 在其对应子空间中的表示，则 X 中的每一列数据属于 D 的列向量张成的线性空间（记为 $span(D)$ ），而关于 $span(D)$ 的已知量只有数据矩阵 X ；另一方面， X 中的每一列向量属于某一子空间，即 X 的列向量可以表示成子空间基的线性组合，而一个线性空间的基并不唯一，这意味着 X 含有足够多的来自同一子空间中的向量，则有这些向量即可张成其所属的子空间。基于上述分析在一定的条件下，将矩阵 D 取为数据矩阵 X ，能够得到数据在其所属子空间中的表示。

子空间表示系数矩阵的计算：

首先分析在特定条件下子空间表示系数矩阵 A 的形式，这里假设子空间之间是相互独立的，即 $S_i \cap S_j = 0, i \neq j$ 且 $i, j = 1,2\cdots,n$ 。由子空间的性质，可证明以下定理成立：

设 $S_i(i=1,\cdots,n)$ 是线性子空间，如果子空间之间是相互独立的，且子空间 $S_i(i=1,\cdots,n)$ 的维数为 d_i ，对应的基矩阵为 $B_i=[b_{1_i},b_{2_i},\cdots,b_{d_i}]\in R^{M\times d_i}$ ，那么有

$$\dim\left(\sum_{i=1}^{n}S_i\right)=\sum_{i=1}^{n}\dim\left(S_i\right)=\sum_{i=1}^{n}d_i \tag{14}$$

且 $\sum_{i=1}^{n}S_i$ 的基矩阵为 $B=[B_1,B_2,\cdots,B_n]$ 。

由线性代数的知识可知，当 $x\in S_i$ ， B_i 为 S_i 的基矩阵时，有 $x=B_i a$ ，且表示系数由 x 与 B_i 唯一确定，即

$$A=\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n \end{pmatrix}$$

其中， A_i 为 X_i 在子空间 S_i 的表示系数矩阵。

考虑实际情况，基矩阵 B 无法直接得到，将字典 D 取为数据矩阵本身求解问题 $X=XA, X=XA$ 的解存在且有如下性质：

设 n 个线性子空间是相互独立的，且子空间 S_i 的维数为 d_i ，矩阵 $X=\left[X_1,X_2,\cdots,X_n\right]\in R^{M\times N}$, $X_i=\left[x_{1_i},x_{2_i},\cdots,x_{n_i}\right]$ 且 $x_j\in S_i, j=1_i,\cdots,n_i, i=1,\cdots,n$

当 X_i 的秩为 d_i 时， $X=XA$ 具有形如分块对角阵的解

$$A^*=\begin{bmatrix} A_1^* & 0 & \cdots & 0 \\ 0 & A_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n^* \end{bmatrix}$$

其中 A_i^* 为 X_i 在子空间 S_i 的表示系数矩阵。

利用矩阵分解的方法，通过对 X 进行奇异值分解得到了一个具有分块对角结构的解 $V_0V_0^T$，称为形状交换矩阵，其中 V_0 为 X 的瘦形奇异值分解右奇异向量矩阵。

2009 年，Elhamifar 和 Vidal 在子空间聚类中引入稀疏约束，提出稀疏子空间聚类（Sparse Subspace Clustering, SSC）。SSC 模型考虑线性子空间相互独立的情形，设数据矩阵为 $X=[x_1,x_2\cdots,x_N]=[X_1,X_2,\cdots X_n]\Gamma\in R^{M\times N}$,其中 $X_i\in R^{M\times n_i}$ 来自子空间 S_i，$\sum_{i=1}^{n}n_i=N$, i=1,⋯,n 。$\Gamma\in R^{N\times N}$ 是列向量的一个排列，即数据矩阵 X 通过特定的列变换可以将同一子空间的数据排列在一块。X 的列向量 x 的稀疏子空间聚类表示为：

$$\min_a \|a\|_1 \quad \text{s.t. } x=\tilde{X}a \tag{15}$$

其中，$\tilde{X}$ 为 X 中去掉列向量 x 的矩阵。并且(15)有如下性质：

设 $X\in R^{M\times N}$ 的列向量来自 n 个线性子空间 S_i 的并，S_i 之间相互独立且 $X=[X_1,X_2,\cdots,X_n]\Gamma\in R^{M\times N}$。令 x 为第 i 个子空间中的点，则式(15)的解：

$$a=\Gamma^{-1}[a_1^T,a_2^T,\cdots,a_n^T]^T\in R^N \tag{16}$$

具有块稀疏结构，即 $a_i\neq 0, a_j\neq 0, (j\neq i)$。

式子(15)是一个 L_1 优化问题，可利用凸规划方法求解，SSC 模型利用 Lasso 方法求解数据的子空间表示系数。对于(15)，为了避免出现平凡解，实际计算时，字典 X 中将去掉当前计算的列向量 x，对应系数取零，反映到系数矩阵为系数矩阵的对角线元素为零。将模型写为矩阵形式则得到：

$$\min_A \|A\|_1 \quad \text{s.t. } X=XA, \text{diag}(A)=0 \tag{17}$$

其中 $\|A\|_1=\sum_i\sum_j|A_{i,j}|$。

通过求解上述模型，得到 X 的子空间表示系数矩阵 A，由此可以构造图 G 的邻接矩阵 W。由于无向图的邻接矩阵具有对称性，故：

$$W=|A|+|A^T| \tag{18}$$

其中。$|A|$ 表示 A 中每个元素取绝对值。

5.4.2 SSC 模型扩展[5]

实际中的问题，数据往往会受到噪音污染或边界出现奇异值，这种情形下数据并不完全位于子空间里，并可能出现数据点丢失条目的情况，据此对 SSC 模型进行如下推广，令

$$x_i=x_i^0+e_i^0+z_i^0 \tag{19}$$

x_i 为第 i 个无误点 x_i^0 经破坏得到，其完全位于一个子空间中，$e_i^0\in R^D$ 为稀疏边界条目向量，其只有少数的非零大元，即 $\|e_i^0\|\leq k$，k 为某一整数，$z_i^0\in R^D$ 为噪音其范数为有界($\|z_i^0\|_2\leq\varsigma,\varsigma>0$)，由于无误点完全位于子空间中故可有如下表示：

$$x_i^0=\sum_{j\neq i}c_{ij}x_j^0 \tag{20}$$

利用(19)式重新表示 x_i^0 得到：

$$x_i = \sum_{j \neq i} c_{ij} x_j + e_i + z_i \tag{21}$$

这里向量$e_i \in R^D$和$z_i \in R^D$定义分别如下：

$$e_i \triangleq e_i^0 - \sum_{j \neq i} c_{ij} e_j^0 \tag{22}$$

$$z_i \triangleq z_i^0 - \sum_{j \neq i} c_{ij} z_j^0 \tag{23}$$

将e_i和z_i作为列分别拼成矩阵E和Z，则可以将(21)重新写成矩阵形式：

$$X = XA + E + Z, diag(A) = 0 \tag{24}$$

我们的目标是求出(24)式的解(A, E, Z)，A为对应的稀疏系数矩阵，为解决此问题，我们提出下面优化规划问题[]：

$$\begin{aligned} &\min \|A\|_1 + \lambda_e \|E\|_1 + \frac{\lambda_z}{2} \|Z\|_F^2 \\ &s.t.\ \ \text{X=XA+E+Z,diag(A)=0} \end{aligned} \tag{25}$$

其中参数$\lambda_e > 0$，$\lambda_z > 0$，上式为关于变量(A, E, Z)的凸优化问题，故可用凸优化工具有效解决。

综上，SSC模型的算法如下：

表5.3 稀疏子空间聚类算法

算法3：稀疏子空间聚类（SSC）
输入：数据矩阵X,子空间个数n
1.求解稀疏优化问题(17)无误点情形或有误点情形(25)；
2.对矩阵A的列向量进行规范化处理($c_i \leftarrow \frac{c_i}{\|c_i\|_\infty}$)；
3.用N个节点表示出数据的相似图，用$W = \|A\| + \|A\|^T$对两节点的边设置权值；
4.对相似图用谱聚类方法聚类。
输出：数据的分割结果：$X_1, X_2, \cdots, X_n$

5.4.3 基于ADMM的SSC算法模型[5]

ADMM(alternating direction method of multipliers)方法适用于解决如下图优化问题：

$$\begin{aligned} &\min \quad \text{f(x)+g(x)} \\ &\text{s.t.} \quad \text{Ax+Bz=c} \end{aligned} \tag{26}$$

变量$x \in R^n$ z$\in$R^m A$\in R^{p \times n}$ B$\in$R$^{p \times m}$ c$\in$R^p

假设f, g是凸的，上述优化问题可表示为：

$$P^* = \inf\{ f(x) + g(x) \,|\, Ax + Bz = c \}$$

可用如下形式增广的拉格朗日乘子：

$$L_p(x, z, y) = f(x) + g(z) + y^T (Ax + Bz - c) + (\rho / 2) \|Ax + Bz - c\|_2^2$$

ADMM算法包含如下迭代：

$$x^{k+1} := \arg\min_x L_\rho(x, z^k, y^k)$$

$$z^{k+1} := \arg\min_z L_\rho(x^{k+1}, z, y^k)$$

$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c), \rho > 0$$

故要解决的优化问题具有形式：

$$(x^{k+1}, z^{k+1}) := \arg\min_{x,z} L_\rho(x, z, y^k)$$
$$y^{k+1} : y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \tag{27}$$

5.5 谱多流形聚类模型(SMMC)[9]

Spectral Clustering on Multiple Manifolds(SMMC)尽管 SC 方法已经成功地应用于很多有挑战性的聚类场景中，但不同聚类交叉式往往会出现错误。基于 SC 方法可算的较好当不同类点的亲和值较小时，提出 SMMC 了方法，可用来处理交叉，假设数据位于或接近多个平滑的低维流形，一些数据流形独立或交叉，样本数据的局部几何信息是通过构建合适的亲和矩阵呈现的，最后谱方法应用于亲和力矩阵对数据完成分组。分类结果显示，SMMC 分类性能较好。

流形聚类的目的是把输入流行数据集分为若干个类别，使得每个类别中的数据点都来自单一、简单、低维嵌入流形首先假设低维流形的数目和维数是已知的。

给定无标记数据集 $X = \{x_i \in \Re^D, i = 1, \cdots, N\}$ 来自 $k > 1$ 个不同的光滑流形 $\{\Omega_j \subseteq \Re^D, j = 1, 2, \cdots, k\}$，流形聚类的目标是对各样本点找出其所属的流行。在两个点 x_i, x_j 之间考虑两个亲和力函数，其中一个为它们定义的切空间，另一个为定义的 Euclidean 距离 $q_{ij} = q(\|x_i - x_j\|)$，则这两个函数以下列亲和力值融合在一块：

$$w_{ij} = f(p_{ij}, q_{ij}) \tag{28}$$

现对 p, q 给出具体的公式，假设切空间在 $x_i (i = 1, \cdots, N)$ 为 Θ_i，则切空间在 x_i, x_j 两点之间的相似性可定义为：

$$p_{ij} = p(\Theta_i, \Theta j) = (\prod_{l=1}^{d} \cos(\theta_l))^o \tag{29}$$

(29) 中，$o \in N^+$ 为调和参数，$0 \le \theta_1 \le, \cdots, \le \theta_d \le \pi / 2$ 为两个切空间 Θ_i, Θ_j 主要角序列，其定义为：

$$\cos(\theta_1) = \max_{\substack{u_1 \in \Theta_i, v_1 \in \Theta_j \\ \|u_1\| = \|v_1\| = 1}} u_1^T v_1 \tag{30}$$

$$\cos(\theta_l) = \max_{\substack{u_l \in \Theta_i, v_l \in \Theta_j \\ \|u_l\| = \|v_l\| = 1}} u_l^T v_l, l = 2, \cdots, d \tag{31}$$

其中 $u_l^T u_i = 0, v_l^T v_i = 0, i = 1, \cdots, l-1$。

局部相似简单定义为：

$$q_{ij} = \begin{cases} 1, x_i \in Knn(x_j) \text{ or } x_j \in \mathrm{Knn}(x_i) \\ 0, otherwise \end{cases} \tag{32}$$

其中 $Knn(x)$ 代表 x 的 K 附近领域数。

最后两个函数简单增加在一块给出下列亲和值：

$$w_{ij} = p_{ij} q_{ij} = \begin{cases} (\prod_{l=1}^{d} \cos(\theta_l))^o, x_i \in Knn(x_j) \text{ or } x_j \in Knn(x_i) \\ 0, \quad \text{otherwise} \end{cases} \tag{33}$$

容易检验(33)中定义的亲和力值能达到预期效果，即属于不同聚类的点亲和力值相对较小。

表5.4谱多流形聚类模型

算法4：谱多流形聚类算法（SMMC）
输入：数据集 X ,聚类个数 k ，流行的维数 d ，混合模型数 M ，领域数 K ,调优参数 o 。 1.用MPPCA原理近似基流行使 M 变为 d 维的局部线性流形； 2.对每一个点确定局部切空间； 3.用 (29) 对两个成对的切空间计算它们的成对亲和力； 4.用 (33) 计算亲和矩阵 $W \in \Re^{N\times N}$ ； 5.计算对角矩阵 $E, E_{ij} = \sum_j w_{ij}$ ； 6.计算方程 $(E-W)u = \lambda eu$ 前 k 个广义特征向量 $u_1, \cdots, u_k$ ； 7.利用K-means算法对 $\Re^k$ 中 U 的行向量聚类。 输出： k 个不相交类的分割。

5.6 稀疏流形聚类和嵌入模型(SMCE)[10]

5.6.1 方法提出：

在许多机器学习、模式识别、信息检索和计算机视觉中，我们会在流行及其附近遇到高维数据，这种情况下，重要的是进行降维，找到数据的紧致表示以破坏它们的自由度。大多数降维方法的第一步是建立一个邻域图，固定数量的连接给定半径邻域内所有点，局部方法有：LLE,Hessian LLE 和拉普拉斯特征映射，试图在数据点的领域内用权值表示局部的关系。全局方法：半定的嵌入，最小体积嵌入等试图维护全部数据点的局部与全局关系。两种方法选取合适的领域大小是构建邻域图的关键，具体地说，小的邻域捕获不了足够的流行几何信息，而大的邻域获取多方面的信息可能违反原则。

流形聚类的目的是把输入流行数据集分为若干个类别，使得每个类别中的数据点都来自单一、简单、低维嵌入流行首先假设低维流行的数目和维数是已知的。

假设给定集合 N ，数据点 $\left\{x_i \in R^D\right\}_{i=1}^N$ 位于 n 个不同的流形 $\left\{M_l\right\}_{l=1}^n$ 上，且流行对应的维数为 $\left\{d_l\right\}_{l=1}^n$ ，以下着手解决基于基流形对数据进行聚类及数据在聚的每一低维类中的表示问题。我们用谱聚类和嵌入算法处理这个问题，特别的我们建立了一个相似图其节点代表数据点边代表数据之间的近似，但困难处是节点按照何种方式选取，为实现聚类，在同一流行中每一点与其他点相连接，为做到降维，将每一点用权值（代表信息）与附近点连接，为同时实现这两种操作，同一流行中我们选取附近的点。进而对此问题我们提出一个基于稀疏表示的计算优化算法问题，该方法背后的基本假设是每一数据点都有个小邻域，在同一流行中该点的此小领域中最小数量的点形成一个低维的仿射子空间。其精确的假设为：

假设：对每一数据点 $x_i \in M_l$ ，考虑最小的球域 $\mathrm{B}_i \subset R^D$,其包含 d_l+1 维的最近邻域(数据 x_i)，令邻域集 N_i 为所有数据点（除 x_i ）的集合，一般地，此领域包含 M_l 或其他流行中的点，假定对所有 i 存在 $\varepsilon \geq 0$ 使得下述稀疏解的非零项:

$$\left\| \sum_{j\in N_i} c_{ij}(x_j - x_i) \right\|_2 \leq \varepsilon, \sum_{j\in N_i} c_{ij} = 1 \tag{34}$$

对应流行 M_l 中包含数据点 x_i 的 d_l+1 维的领域。

5.6.2 优化算法

我们的目标是选取一个方法，对每个数据点 x_i 以及同流行中的一些领域，若邻域 N_i 已知且相对较小，则可以找到最小数量的点满足(34)，若 N_i 未知，则在 N_i 中找到数据点满足(34)就变得复杂（因为领域在变大），为处理这类问题，我们让领域的大小是任意的，可是，通过用稀疏优化程序，我们选取靠近 x_i 的一些点形成低维的仿射子空间通过 x_i 的附近的方法是有偏见的。

考虑 d_l 维流行 M_l 中的点和集合 $\{x_j\}_{j\neq i}$，由上述假设，M_l 中靠近 x_i 的一些点形成 d_l 维的仿射子空间 R^D 通过 x_i 的附近,数学描述为：

$$\|[x_1-x_i\cdots x_N-x_i]c_i\|_2\le\varepsilon,1^Tc_i=1 \tag{35}$$

有一解 c_i 其 d_l+1 个非零项对应过 M_l 中的点 x_i 的 d_l+1 个领域。

注意：解 c_i 为最小数量的非零项，不再是唯一的。规范化向量 $\{x_j-x_i\}_{j\neq i}$ 令

$$X_i\triangleq[\frac{x_1-x_i}{\|x_1-x_i\|_2}\ \cdots\frac{x_N-x_i}{\|x_N-x_i\|_2}]\in R^{D\times N-1} \tag{36}$$

下考虑阀点接近 x_i 的目标函数，即接近 x_i 的点其阀值比远点的低于是有如下 l_1 优化问题：

$$\min\|Q_ic_i\|_1\quad\text{s.t.}\quad\|X_ic_i\|_2\le\varepsilon,\ 1^Tc_i=1 \tag{37}$$

6、模型求解

6.1 问题一的求解

由已知，问题 1 中数据在两个独立的子空间，因此聚类相对容易,因此我们选用经典的 K 均值聚类与谱聚类(SC)均可以得出分类结果。根据第 5 节建立的稀疏子空间聚类(SSC)、谱曲率聚类(SCC)、基于交替方向乘子法的稀疏子空间聚类(SSC+ADMM)、稀疏流形聚类(SMCE)、谱多流形聚类(SMMC)模型也可以得到相同的分类结果。

在 SCC 算法中令 d=1, K=2, c=200.

在 SSC-ADMM 算法中，需要先利用 PCA 方法对数据进行降维，我们仿真实验可以发现，把原数据的维数降到 2-7 维时可以得到较好的分类结果，与其它分类一致。

在 SMCE 算法中，$\lambda\in(0,100)$，其中特征集 $L=200/10=20$.

SMMC 算法中，$k=2$，流形维数 $d=7,8,9,10$，M 为混合流形数，K 为数据点邻域的数量，o 为调节参数，N 为数据个数。

$$M=\lceil N/(10d)\rceil=2,K=\lceil 2\log N\rceil=10,o=8$$

用 Matlab 编程实现上述算法，并将参数代入，运行得分类结果见表 6.1.

表 6.1

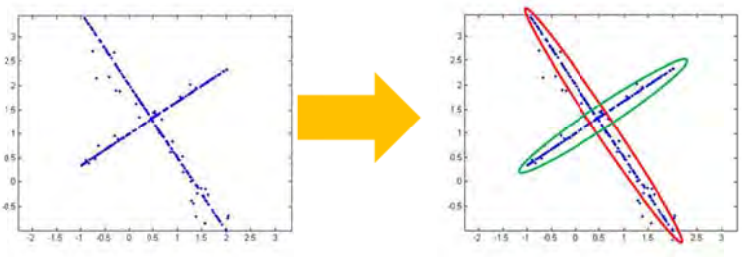
数据	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

数据	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	101	102	103	104	105	106	107	108	109	110	111	112	113	104	105	106	107	108	109	120
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

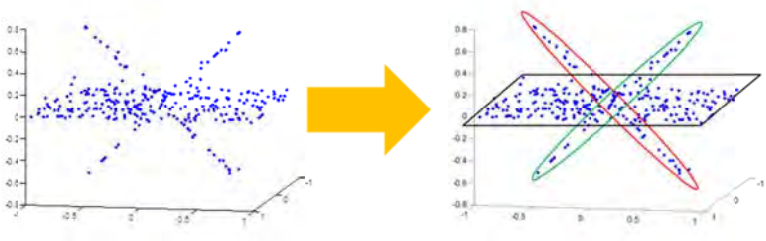
6.2 问题二的求解

问题 2 是对四个数据进行分类，根据题意可知，图(a), (b), (d)是不同的类具有交叉的情况，(c)为无交叉的情况；(a), (b)是线性的，(c),(d)是非线性的。

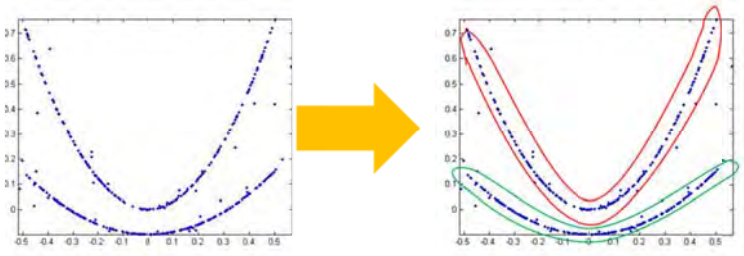
本题主要是解决四个低维空间中的子空间聚类问题和多流形聚类问题，我们的预期分类结果为：



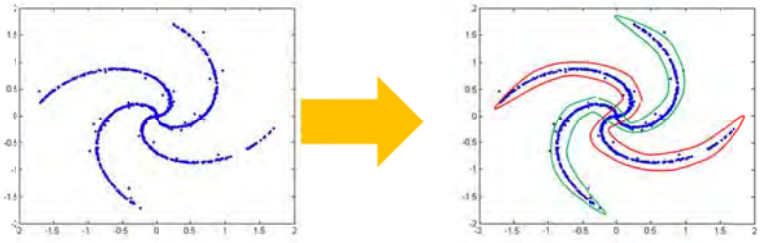
(a)



(b)



(c)



(d)

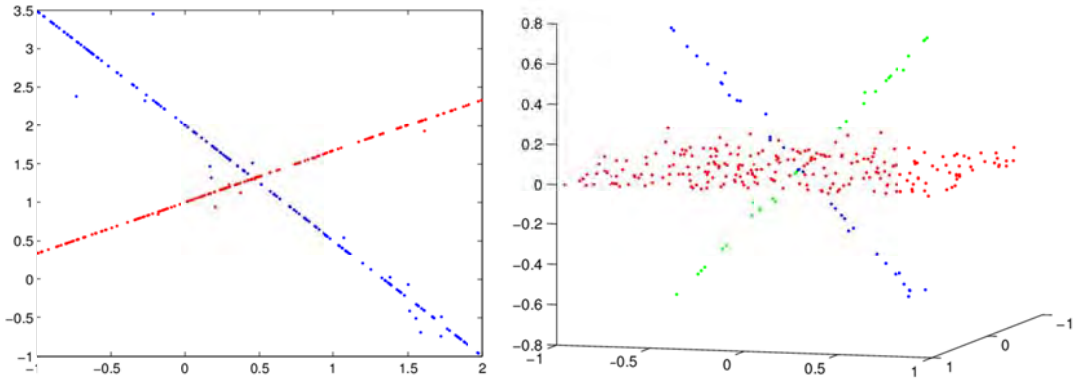
图 6.1

6.2.1 SMMC 算法对各问题实现结果情况：

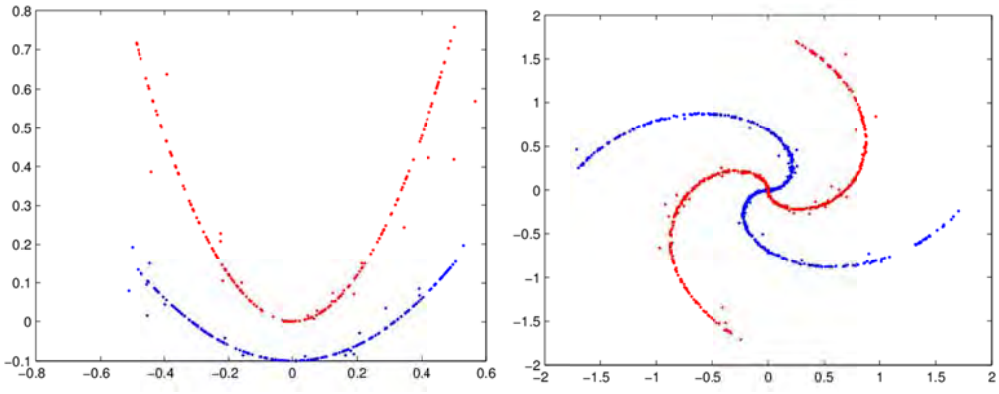
表 6.2 SMMC 算法参数

问题	聚类数	流形维数	局部化模型数	近邻点数	调节参数
(a)	2	1	4	7	13
(b)	3	1	12	17	11
(c)	2	1	2	7	2
(d)	2	1	50	20	10

用 Matlab 运行 SMMC 程序实现如下的分类结果：



(a)　　(b)



(c) (d)

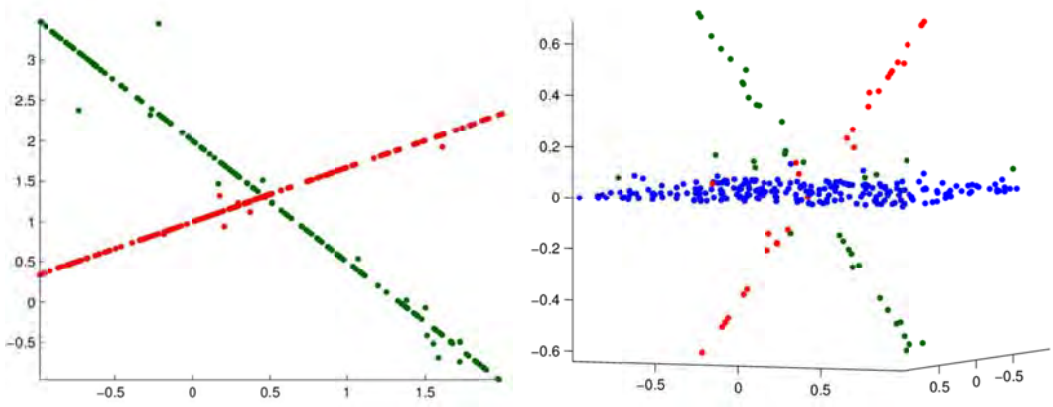
图 6.2

6.2.2 SCC 算法对各问题实现结果情况：

表 6.3 SCC 算法参数

问题	聚类数	流形维数	采样点数
(a)	2	1	200
(b)	3	2	300

用 Matlab 运行 SCC 程序实现如下的分类结果：



(a) (b)

图 6.3

从实现的结果可以看出：SCC 算法对于问题(c)和(d)的分类情况达不到预期的效果。

6.2.3 SMCE 算法对各问题实现结果情况：

表 6.4 SMCE 算法参数

问题	聚类数	流形维数	近邻点数	调节参数
(a)	2	2	20	1

用 Matlab 运行 SMCE 程序实现如下的分类结果：

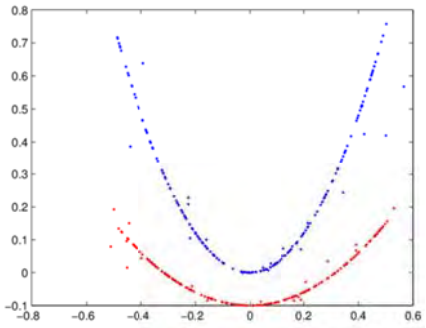


图 6.4

从实现的结果可以看出：SMCE 算法对于问题(a),(b)和(d)的分类情况达不到预期的效果。

6.3 问题三的求解

6.3.1 问题三图 2(a)的求解

图 2(a)的中为确定十字的中心位置，我们分别用谱多流形聚类(SMMC)和曲率谱聚类(SCC)算法，通过调节参数，对十字上的点很好的聚成了两类。

运用谱多流形聚类算法(SMMC)对问题实现结果如下：

表 6.5 SMMC 算法参数

问题	聚类数	流形维数	局部化模型数	近邻点数	调节参数
(a)	2	1	43	32	9

用 Matlab 运行 SMMC 程序实现如下的分类结果：

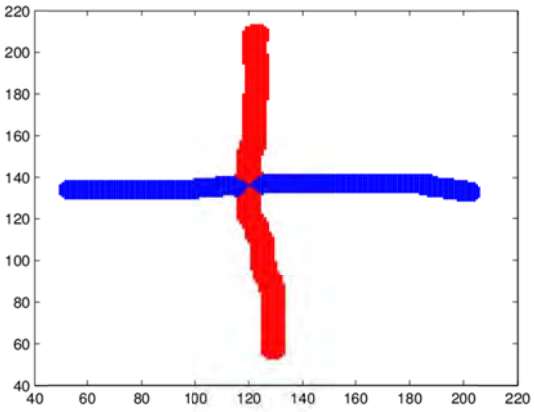


图 6.5

运用谱曲率聚类(SCC)算法对问题实现结果如下：

表 6.6 SCC 算法参数

问题	聚类数	流形维数	采样点数
(a)	2	1	200

用 Matlab 运行 SCC 程序实现如下的分类结果：

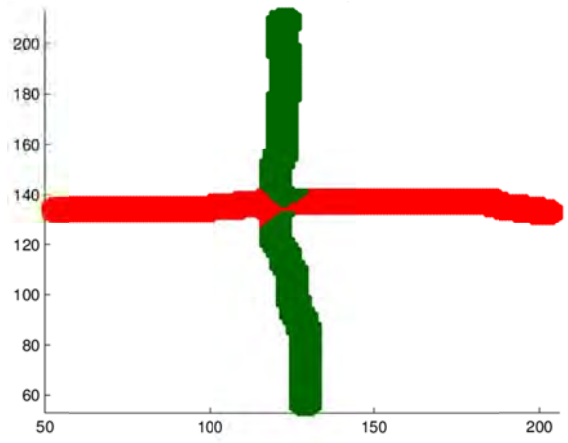


图 6.6

6.3.2 问题三图 2(b)的求解

对运动特征轨迹点的分类，我们采用六种方法：K 均值聚类，SC，SCC，基于 ADMM 的 SSC，SMMC 与 SMCE，具体分类结果见表 6.7.

其中，K 均值聚类将第 275 个数据，第 276 个数据归于第 3 类。SC 算法与 SCC 算法与表格一致。基于 ADMM 的 SSC 算法将第 267 数据，第 275 个数据，第 276 个数据，第 297 个数据归于第 3 类。SMMC 算法将第 275 个数据，第 276 个数据，第 297 个数据归于第 3 类.SMCE 在 λ 取不同的值时, 将第 275 个数据归为第 2 类或者第 3 类。

分析上述分类结果，我们判断第 2 类和第 3 类的特征轨迹有可能相交，第 267 数据，第 275 个数据，第 276 个数据，第 297 个数据可能位于第 2 类和第 3 类的边界处，所以可能出现分类的不同。

SMMC 算法，$k=3$, 流形维数 $d=9$,混合流形数 M , 数据点领域数 K , 调节参数 o

$$M=\lceil N/(10d)\rceil=3\ ,\ K=\lceil 2\log N\rceil=10\ ,\ o=8$$

SMCE 算法，令 $\lambda\in(0,150)$ ，其中特征集 $L=30$.

表 6.7

数据	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数	10	10	10	10	10	10	10	10	10	11	11	11	11	10	10	10	10	10	10	12

据	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
数据	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
类别	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
数据	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
数据	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220
类别	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
数据	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
类别	3	3	3	3	3	3	3	3	3	2	3	2	3	3	3	3	2	3	2	3
数据	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
类别	3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3	3
数据	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280
类别	3	3	3	2	3	3	2	2	3	2	3	2	3	3	2	2	3	3	3	3
数据	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297			
类别	3	2	3	3	3	3	2	3	3	3	3	3	3	3	3	3	2			

6.3.3 问题三(c)的求解

采用 SCC 算法、基于 CVX 的 SSC 算法，基于 ADMM 的 SSC 算法与 SMCE 算法，可以得到一致的分类结果，见表 6.8.

SMCE 算法，令 $\lambda \in (0,50)$，其特征集 $L = 10$.

表 6.8

数据	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
类别	1	1	1	1	1	2	2	2	2	2	1	1	1	1	1	2	2	2	2	2

6.4 问题四的求解

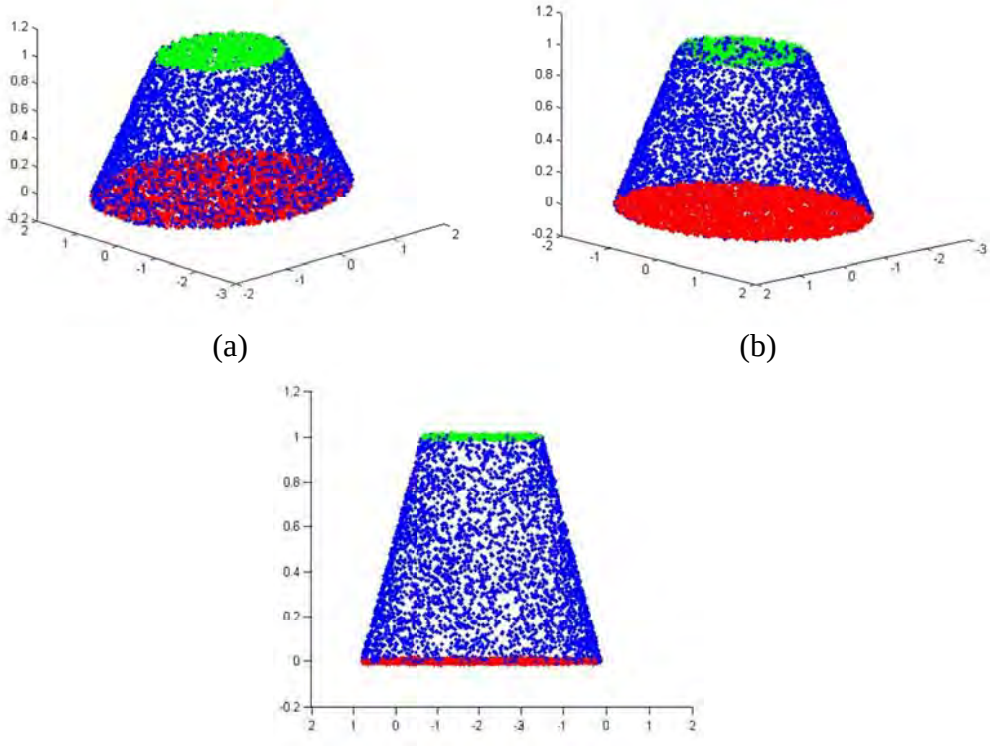
6.4.1 问题四图 3(a)的求解

为将圆台的顶、底、侧面分成三类，我们用了多流形谱聚类算法(SMMC)，通过调节参数，实现了可观的分类效果。

表 6.9 SMMC 算法参数

问题	聚类数	流形维数	局部化模型数	近邻点数	调节参数
4(a)	3	1	42	22	13

用 Matlab 运行 SMMC 程序将圆台的顶、底、侧面分成三类，如下为实现的一个聚类结果，其为站在不同视角下的三个效果图。



(a) (b)

(c)

图 6.7

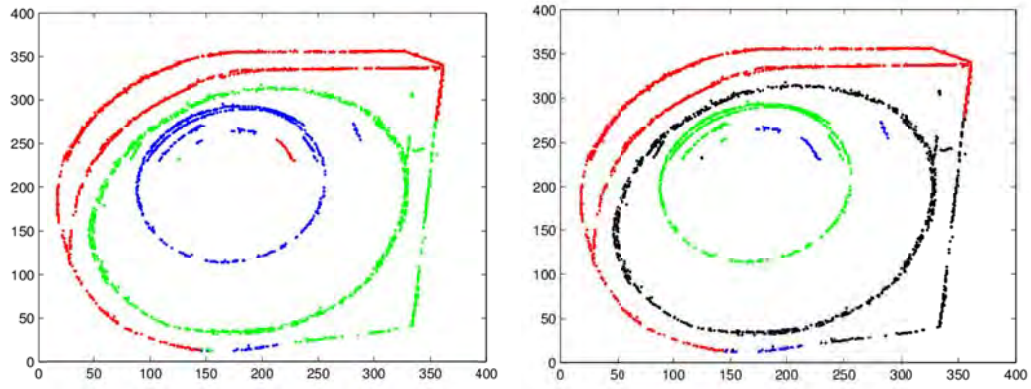
6.4.2 问题四图 3(b)的求解

根据问题的要求，我们用 SMCE 算法分别将轮廓线中的不同直线和圆弧分为 3 类和 4 类，得出较好的分类效果。

表 1 SMCE 算法参数

问题	聚类数	流形维数	调节参数	近邻点数
4(b)	3	2	5	10
4(b)	4	2	5	10

用 Matlab 运行 SMCE 程序,聚类数分别为 3 和 4 时的效果图如下：



三类结果图　　四类结果图

图 6.8

7、模型评价与总结

本文针对问题中给出的数据的结构特点，建立不同的模型进行数据分类。并综合运用各种方法得到较好聚类效果。

在问题一中，本文给出的七种聚类方法均可给出一致的分类结果，这表明位于不同独立子空间的数据分类较为容易，可选用的聚类方法多。

在问题 2 中，针对四个低维空间的聚类问题，给出三种方法进行分类，其中 SMMC 可以给出四种问题的理想分类结果。而 SCC 与 SMCE 均只能对部分问题进行分类。这表明 SMMC，SCC，SMCE 均适用位于不相互独立的子空间上的数据分类；SMMC，SCC 适用于位于独立流形的数据的分类。SMMC 适用于曲线类的位于不独立流形的数据的分类。SMMC 适用范围最广。

问题三中(a)的分类可应用问题二中(a)给出的方法，实验结果表明可以得到理想的分类效果，也对问题 2 中的结论进行了有效验证。在运动轨迹点的分类中，用六种方法进行了分类，其中有些方法在某些点存在差异，在合理范围内。在人脸数据分类中，所采用四种聚类方法均可得出一致的分类结果。不同模型的实验结果的相似性说明本问题中的实验结果具有很高的可靠性。

在问题 4 对圆台的分类中，本文使用上文中的各种模型，进行了大量实验，最后只有 SMMC 模型得出理想结果。同样，在进行机器的轮廓边缘分类实验中，本文也使用上文所提供的各种模型进行了大量实验，最后只有 SMCE 得到了较为理想的效果。综上，基于数据流形结构的分类方法具有更广的适用范围，可以在大多数情况下得到理想的分类效果。

本文介绍的对高维数据聚类分析的方法在市场分析、信息安全、金融、娱乐、反恐等方面都有很广泛的应用。

参考文献

[1] J. A. Hartigan and M. A. Wong, A k-means clustering algorithm, Appl. Stat., 28(1):100-108,1979.

[2] A. Ng, Y. Weiss, and M. Jordan, On Spectral Clustering: Analysis and an Algorithm, Proc. Neural Information Processing Systems Conf. 849-856, 2001.

[3] 蔡晓妍等，谱聚类算法综述[J]. 计算机科学， 35（7）：14-17,2008.

[4] Wen-Yen Chen, Yangqiu Song, Parallel Spectral Clustering in Distributed Systems. 33(3):568-586,2011.

[5] 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.

[6] J. Shi and J. Malik, Normalized cuts and image segmentation, IEEE Transactions Pattern Analysis Machine Intelligence, 22(8):888–905, 2000.

[7] R.Vidal. Subspace clustering. IEEE Signal Processing Magazine, 28(2):52–68, 2011.

[8] Guangliang Chen, Spectral Curvature Clustering. Int J Comput Vis, 81:317-330,2009.

[9] Y. Wang, Y. Jiang, Y. Wu, and Z. Zhou. Spectral clustering on multiple manifolds. IEEE Transactions on Neural Networks, 22(7):1149–1161, 2011.

[10] Ehsan Elhamifar, http://www.eecs.berkeley.edu/~ehsan.elhamifar/code.htm, 2015.9.20.

[11] S. Wang, J. M. Siskind. Image segmentation with ratio cut. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(6):675-690, 2003.

[12] M. Belkin and P. Niyogi, Laplacian eienmaps for dimensionality reduction and data representation, Neural Comput., 15(6):1373-1396,2003.

[13] M. E. Tipping and C. M. Bishop, Mixtures of probabilistic principal component analyzers, Neural Comput., 11(2):443-482, 1999.

[14] S. T. Roweis and L. K. Saul, Nonlinear dimensionality reduction by locally embedding, Science, 290(5500):2323-2326, 2000.

[15] R.Vidal , Visonlab, http://vision.jhu.edu, 2015.9.20.