

# 数学综合课听课报告

彭真 BY1809110

2019年5月15日

通过本次的数学综合课，我学到了很多知识，对量子信息、不确定性量化、和动力系统、数值逼近、偏微分方程数值解、序代数都有了初步的认知，对统计学习理论、信号处理、最优化理论与算法、大数据的算法及应用有了深入的理解，收获很多，非常感谢老师的讲解。由于现阶段正在看流形学习 (Manifold Learning) 在信号处理上的应用这一方面的文章，因此听课报告总结了典型的流形学习方法及在信号处理的应用，记录学习的心得体会。另外需要说明一点，我曾经上陈迪荣老师的统计学习理论课时交的读书笔记也涵盖在本报告中，后期也对其做了补充和完善，所以最后展示到这份听课报告中。

该报告的理论部分主要参考了周志华老师的机器学习 [1]、北大数院 Yuan Yao 老师的专著 *A Mathematical Introduction to Data Science*[2] 和经典教材 *The Elements of Statistical Learning*[3] 以及 CMU 的 Liu Yang 老师关于度量学习做的综述 [4, 5, 6]。

## 1 动机

流形学习是机器学习、模式识别中的一种方法，在维数约简方面具有广泛的应用。它的主要思想是将高维的数据映射到低维，使该低维的数据能够反映原高维数据的某些本质结构特征。流形学习的前提是有一种假设，即某些高维数据，实际是一种低维的流形结构嵌入在高维空间中。流形学习的目的是将其映射回低维空间中，揭示其本质。在此借用周志华老师的对流形学习的解释：

### 流形学习

在高维情况下出现数据样本稀疏、距离计算困难等被称为“维数灾难”的问题和过拟合问题是在提取特征过程中遇到的常见障碍，缓解以上问题的途径之一是降维，即在高维空间中找到更紧凑的低维表示。流形学习是一种借鉴了拓扑流形概念的降维方法，在局部具有欧式空间性质，便于进行距离计算。其目标是寻找到一个低维流形嵌入，保证高维空间样本点间的关系在低维表示中得以保持。

那么在学习流形学习之前首先要明白前人为什么要在流形上花心思，难道是欧式空间不好吗？对于这个问题，文献 [7] 中的作者对欧式几何可能带来的潜在错误总结了两点：

- 不完备

对于对称正定矩阵  $A \in S_{\pm}^2$  可以表示为  $A = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$  且  $ab - c^2 > 0, a > 0, b > 0$ . 那么该矩阵可以在  $R^3$  的点表示出来，并且其约束为一个开的凸锥（见图1）。这就意味着欧式几何不能保证测地线上的点约束在潜在的流形上，导致后续操作（比如插值）会带来无法接受的结果。

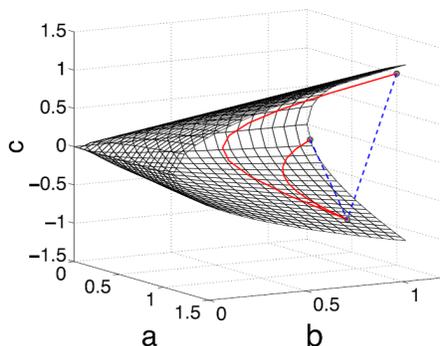


图 1: 比较在  $A \in S_{+}^2$  空间中点之间的欧式距离 (蓝色虚直线) 和黎曼距离 (红色实曲线)

- 膨胀效应

比如求两个矩阵  $A = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$ ,  $B = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$  的平均  $C = \frac{1}{2}(A + B) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ , 那么  $C$  的行列式会大于  $A$  或  $B$  的行列式。此外, 对于样本协方差矩阵用极大似然估计计算时, 较大的特征值往往被高估而较小的特征值往往被低估, 这也可以看做膨胀效应的结果。

## 2 MDS

在学习等度量映射 (Isometric Mapping, ISOMAP) 之前首先要了解多维缩放 (Multidimensional Scaling, MDS), 因为 ISOMAP 算法是以多维标度法 (Multi-dimensional Scaling, MDS) 算法作为理论框架。

### 2.1 MDS 理论基础

MDS 是一组用于分析数据中的相似点或相异点的多元数据分析方法, MDS 的一个很好的特点是它允许我们在低维空间中表示物体对点之间的距离相似性 (或不相似性)。换句话说, MDS 允许我们在低维空间可视化相似性 (或不相似性), 以便进行探索和检查。MDS 是一种将多维空间的研究对象简化到低维空间进行定位、分析和归类, 同时又保留对象间原始关系的数据分析方法。

经典的 MDS 或等距欧氏嵌入的问题: 给定数据点之间的成对距离, 我们可以找到一个欧几里德坐标的系统, 这些点的成对距离满足给定的约束?

考虑一个简单的问题: 根据样本点计算样本之间距离。给定  $n$  个样本  $x_1, x_2, \dots, x_n \in R^p$ , 令  $X = [x_1, x_2, \dots, x_n]^{p \times n}$ , 两样本点  $x_i$  和  $x_j$  间的欧式距离  $d_{ij}$  (dissimilarity measure) 满足

$$d_{ij}^2 = \|x_i - x_j\|^2 = (x_i - x_j)^T (x_i - x_j) = x_i^T x_i + x_j^T x_j - 2x_i^T x_j$$

现在考虑这个问题的逆问题: 根据样本之间距离估计样本点。如果给定  $d_{ij}$ , 我们能否找到  $x_1, x_2, \dots, x_n$  满足上式。显然这个问题有解且解不唯一。通过最小化如下问题 (被称为 *stress function*) 来寻找待估计的样本点:

$$S_{MDS}(x_1, x_2, \dots, x_n) = \sum_{i \neq j} (d_{ij} - \|x_i - x_j\|)^2$$

类似的，基于上述形式平方最小的缩放变体，*Sammon Mapping* 被提出：

$$S_{SM}(x_1, x_2, \dots, x_n) = \sum_{i \neq j} \frac{(d_{ij} - \|x_i - x_j\|)^2}{d_{ij}}$$

根据特征向量这些问题存在显式解，传统的 MDS 算法的中心思想是：

1. 将距离矩阵  $D = d_{ij}^2$  转换为内积表示
2. 对该内积做特征值分解

接下来我们将看 MDS 算法如何利用  $D$ 。

### MDS 推导 [2]

令  $K$  为内积矩阵

$$K = X^T X$$

其中， $k = \text{diag}(K_{ii})$ 。故

$$D = (d_{ij}^2) = k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K$$

其中  $\mathbf{1} = (1, 1, \dots, 1)^T \in R^p$ 。而样本去中心化处理：

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} X \cdot \mathbf{1}$$

$$\tilde{x}_i = x_i - \hat{\mu}_n = x_i - \frac{1}{n} X \cdot \mathbf{1} \Rightarrow \tilde{X} = X - \frac{1}{n} X \cdot \mathbf{1} \cdot \mathbf{1}^T$$

易得

$$\begin{aligned} \tilde{K} &= \tilde{X}^T \tilde{X} \\ &= K - \frac{1}{n} K \cdot \mathbf{1} \cdot \mathbf{1}^T - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T \cdot K + \frac{1}{n^2} \mathbf{1} \cdot \mathbf{1}^T \cdot K \cdot \mathbf{1} \cdot \mathbf{1}^T \end{aligned}$$

令

$$B = -\frac{1}{2} H \cdot D \cdot H^T$$

其中， $H = I - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T$  为居中矩阵 (centering matrix, 规范化)。因此

$$\begin{aligned} B &= -\frac{1}{2} H \cdot D \cdot H^T \\ &= -\frac{1}{2} H \cdot (k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K) \cdot H^T \\ &= H \cdot K \cdot H^T \\ &= K - \frac{1}{n} K \cdot \mathbf{1} \cdot \mathbf{1}^T - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T \cdot K + \frac{1}{n^2} \mathbf{1} \cdot \mathbf{1}^T \cdot K \cdot \mathbf{1} \cdot \mathbf{1}^T \\ &= \tilde{K} \end{aligned}$$

也就是说

$$B = -\frac{1}{2} H \cdot D \cdot H^T = \tilde{X}^T \tilde{X}$$

以上显示，我们可以利用给定的矩阵向量  $D = d_{ij}^2$  转换为内积矩阵  $B = -\frac{1}{2} H \cdot D \cdot H^T$ ，对  $B$  应用特征值分解可以得到以原地为中心样本点坐标。在现实应用中为了有效降维，往往仅需要降维后的矩阵与

原始空间中距离尽可能接近，而不必严格相等。此时可取  $B$  中前  $k (< p)$  个非零特征值对应的特征向量，得到样本的低维坐标，从而达到降维的目的。

---

### Algorithm 1 经典 MDS 算法

---

**Input:** 距离矩阵  $D^{n \times n}$  其中  $D_{ij} = d_{ij}^2$ 。

- 1: 计算内积矩阵  $B = -\frac{1}{2}HDH^T$ ，其中  $H$  为居中矩阵 (centering matrix, 规范化)；
- 2: 计算特征分解  $B = U\Lambda U^T$ ，其中对角矩阵  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ ；
- 3: 取前  $k$  个最大的非零特征值及其相应的特征向量，

$$\tilde{X}_k = U_k \Lambda_k^{1/2}$$

其中， $U_k = [u_1, \dots, u_k]$ ,  $\Lambda_k = \text{diag}(\lambda_1, \dots, \lambda_k)$

---

## 2.2 MDS 与 PCA 关系

首先回顾一下主成分分析 (Principle Component Analysis, PCA)，其目标是将数据从高维数据投影到低维数据，从而使数据的低维方差最大化。PCA 计算协方差矩阵的奇异值分解 (Singular value decomposition, SVD):  $\hat{\Sigma}_n = \frac{1}{n-1} \tilde{X} \tilde{X}^T = \frac{1}{n} US^2U^T$ ，其中， $\tilde{X} = USV^T$ ，并取前  $k$  个左奇异向量作为嵌入的坐标表示为主成分。

然后重新考虑 MDS 的模型：

$$\min_{x_i \in R^k} \sum_{i,j} \left( \|x_i - x_j\|^2 - d_{ij}^2 \right)^2$$

由前面的推导，构造核矩阵  $B = -\frac{1}{2}H \cdot D \cdot H^T$ ，其中， $H = I - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T$ ， $D = d_{ij}^2$ ，上式等价于

$$\min_{X \in R^{k \times n}} \sum_{i \neq j} \|X^T X - B\|_F^2$$

当  $B$  为正半定矩阵时， $B$  可以表示为内积形式，因此  $X$  可表示为  $B = \tilde{X}^T \tilde{X}$  的  $k$  个最大特征值对应的特征向量，或者说， $\tilde{X} = USV^T$  中前  $k$  个右奇异向量。从中心化的数据矩阵进行 SVD 分解这个角度来说，MDS 和 PCA 是统一的，本质上说，当成对距离是欧几里得距离时，MDS 是 PCA 的对偶问题。

通过比较 MDS 与 PCA 的异同发现：两者都是把观察的数据用较少的维数来表达；不同之处在于，MDS 利用的是成对样本间相似性，目的是利用这个信息去构建合适的低维空间，样本在此空间的距离和在高维空间中的样本间的相似性尽可能的保持一致。

## 3 ISOMAP[8]

Isomap 认为低维流形嵌入到高维空间之后，直接在高维空间中计算直线距离具有误导性，因为高维空间中的直线距离在低维嵌入流形上式不可达的。图2为高维数据分布图，蓝色虚线为高维空间中的直线距离，红色实线为相同两点的测地线距离。

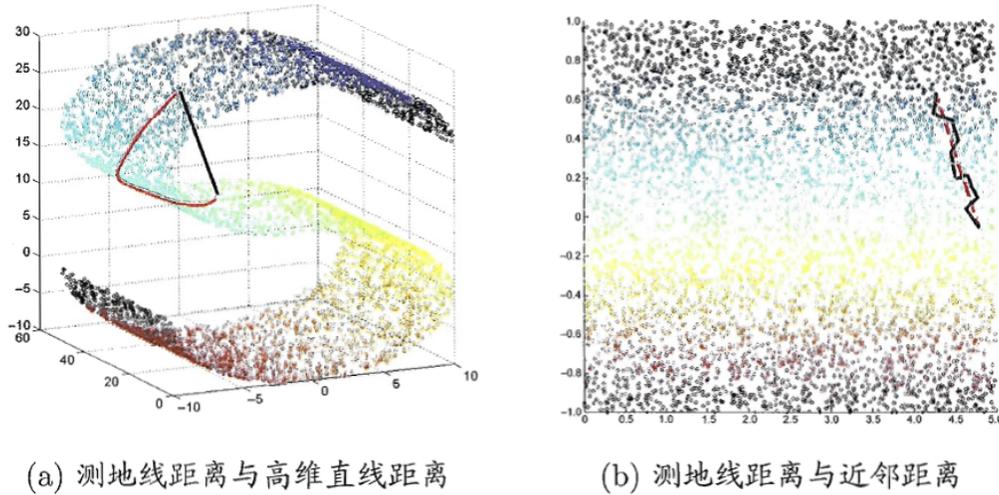


图 2: 测地线距离

利用流形可以进行欧氏距离计算的性质, 对每个点基于欧氏距离找出其近邻点, 于是可以建立一个近邻连接图, 图中近邻点之间存在连接, 而非近邻点不存在连接, 于是计算两点之间测地线距离的问题就转变为计算近邻连接图上两点之间的最短路径问题。

Isomap 算法试图保持近邻样本之间的距离不变, 是 MDS 算法在高维流形上的推广, 即将样本点之间的欧氏距离用黎曼距离替代。

---

**Algorithm 2** Isomap 算法

---

**Input:** 样本集  $D = x_1, x_2, \dots, x_n$ , 近邻参数  $k$ , 低维空间维数  $d'$ 。

- 1: **for**  $i = 1$  to  $m$  **do**
- 2: 确定  $x_i$  的  $k$  近邻;
- 3:  $x_i$  与  $k$  近邻点之间的距离设置为欧氏距离, 与其他点的距离设置为无穷大;
- 4: **end for**
- 5: 调用最短路径算法计算两样本点之间的距离  $dist(x_i, x_j)$ ;
- 6: 将  $dist(x_i, x_j)$  最为 MDS 算法的输入;
- 7: **return** MDS 算法的输出;

**Output:** 样本集  $D$  在低维空间的投影  $Z = (z_1, z_2, \dots, z_m)$

---

Isomap 算法的基本特征可以描述为: 我们获取数据的低维嵌入, 保证附近的点被映射到附近, 距离较远的点被映射到距离较远。换句话说, Isomap 算法对数据距离有全局控制, 因此这个方法是一个全局方法。其主要缺点在于其计算复杂性, 根源在于全矩阵特征向量分解。关于收敛性, 在流形上的稠密样本和正则 (dense-sample and regularity) 条件下, Isomap 算法保证收敛以保持流形上的测地距离。

## 4 LLE[9]

在实际应用中，附近的点应该映射到附近，而远处的点不应该施加约束。这是因为在样本点足够接近导致它们是相似的，而点之间距离很远并不能衡量它们相似程度。这激发了另一种算法：局部线性嵌入 (locally linear embedding, LLE)。这是一种局部方法，因为它涉及局部 PCA 和稀疏特征向量分解。LLE 算法认为每一个数据点都可以由其近邻点的线性加权组合构造得到。该算法的主要步骤分为三步：

1. 寻找每个样本点的  $k$  个近邻点；
2. 由每个样本点的近邻点计算出该样本点的局部重建权值矩阵；
3. 由该样本点的局部重建权值矩阵和其近邻点计算出该样本点的输出值。

具体的算法流程如图3所示：

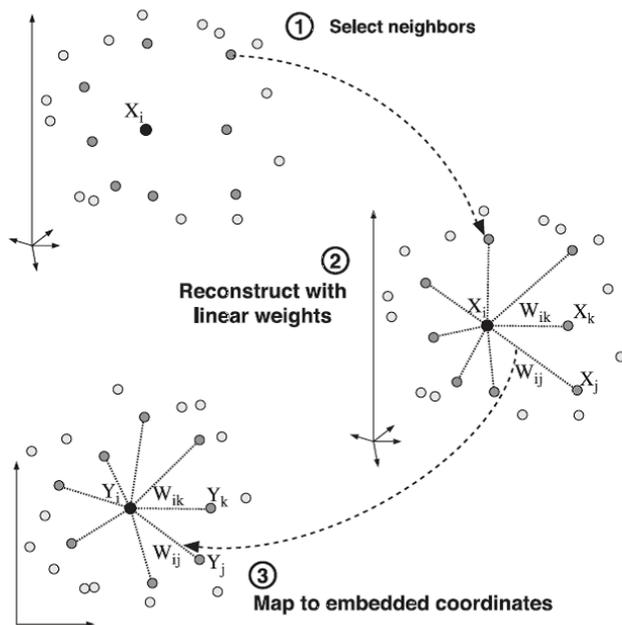


图 3: LLE 算法实现步骤

从理论上讲，LLE 算法尝试保持邻域内样本之间的线性关系。LLE 算法将这个线性拟合的系数当成这个流形局部几何性质的刻画，因此，一个好的低维表示，就应该也具有同样的局部几何，所以利用同样的线性表示的表达式。因此通过以下模型解决：

$$\min_{w_1, w_2, \dots, w_m} \sum_{i=1}^m \left\| x_i - \sum_{j \in Q_i} w_{ij} x_j \right\|_2^2 \quad s.t. \quad \sum_{j \in Q_i} w_{ij} = 1$$

$$\min_{z_1, z_2, \dots, z_m} \sum_{i=1}^m \left\| z_i - \sum_{j \in Q_i} w_{ij} z_j \right\|_2^2, \quad s.t. \quad \sum_{i=1}^m z_i = 0, \quad \frac{1}{m} \sum_{i=1}^m z_i z_i^T = I_d$$

首先解释加入约束条件，根据文献 [9] 的解释，加入这两个约束条件是在不影响损失函数求解的前提下使方程易于求解，且保证解得唯一性。实质上第一个约束是所有点的均值。第二个约束是解的协方差。所有点均值为 0，协方差为单位阵  $I$ 。这就使得解空间分布唯一。

另外值得强调的是，对于每一点处拟合得到的系数归一化的操作特别重要，如果没有这一步，这个算法就没有效果。但是在原始论文中，他们是为了保持数据在平行移动下 embedding 不变。

---

**Algorithm 3** LLE 算法

---

**Input:** 样本集  $D = x_1, x_2, \dots, x_n$ , 近邻参数  $k$ , 低维空间维数  $d'$ 。

- 1: **for**  $i = 1$  to  $m$  **do**
- 2:     确定  $x_i$  的  $k$  近邻;
- 3:     从公式4求得  $w_{ij}$ ,  $j \in Q_i$ ;
- 4:     对于  $j \notin Q_i$ , 令  $w_{ij} = 0$ ;
- 5: **end for**
- 6: 从公式4得到  $M$ ;
- 7: 对  $M$  进行特征值分解;
- 8: **return**  $M$  的最小  $d'$  个特征值对应的特征向量;

**Output:** 样本集  $D$  在低维空间的投影  $Z = (z_1, z_2, \dots, z_m)$

---

然而，与 ISOMAP 不同的是，LLE 算法是否在一定条件下收敛我们尚不清楚。

## 5 统一框架

上面介绍了 ISOMAP 与 LLE 详细步骤，那么这两个流形学习算法及其拓展在本质上是否存在某种关联？能否用一种通用的理论去解释不同的流形学习？在这里，需要预先了解核主成分分析 (Kernel PCA)。

### 5.1 Kernel PCA

现实应用中，数据样本点往往具有非线性的结构，直接使用线性降维方法进行降维可能丢失原始低维空间结构，如：线性投影 PCA，见图4），可能需要非线性映射才能找到恰当的低维嵌入，比如最常用的 kernel PCA。

首先回顾一下主成分分析 (Principle Component Analysis, PCA)，其目标是将数据从高维数据投影到低维数据，从而使数据的低维方差最大化。PCA 计算协方差矩阵的奇异值分解 (Singular value decomposition, SVD)，并取特征向量（数据中方差最大的方向）的最大值。但是，PCA 的局限性在于它仅适用于线性结构，换句话说它不能正确地检测到任何非线性结构。很多人使用这种方法是因为它快速且容易实现，但是在非线性结构数据中应用 PCA 是没有意义的。PCA 在检测潜在曲面结构的数据方面做得不是很好，PCA 将距离较远的数据点映射到平面中的附近点，从图4中可以看出，低维空间中的某些点是重叠的。

值得注意的是，PCA 将原始样本投影到超平面，那么这个超平面在统计意义下应该有两个性质：1) **最近重构性**：样本点到这个超平面的距离都足够近；2) **最大可分性**：样本点在这个超平面的投影能尽可

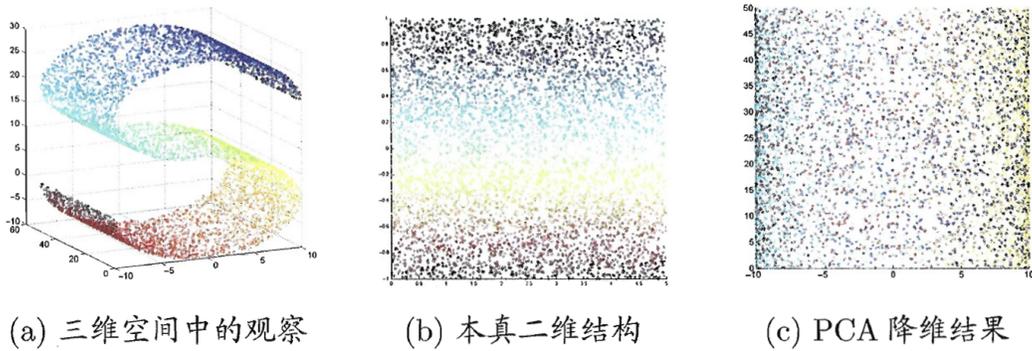


图 4: PCA 降维丢失原始低维空间结构

能分开。在欧式空间中，以这两个为目标能推导出 PCA 的两种等价推导 [1]，然而在非欧式空间（比如黎曼流形）上，这两者并不等价 [7]。

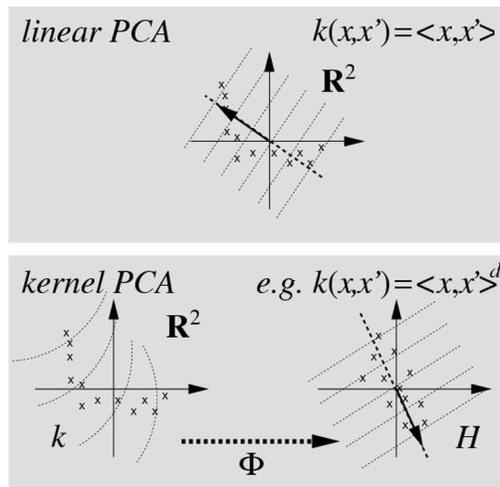


图 5: KPCA 的基本思想

作为 PCA 的推广, kernel PCA 的确有能力处理非线性结构。该方法引入核函数  $\mathcal{K}(x_i, x_j) = \phi(x_i)^T \phi(x_j)$  将原始数据集映射到更大的维特征空间, 将非线性结构映射到线性, 然后应用 PCA 进一步减小尺寸。在这里, 我们使用高斯核方程来计算两点之间距离的平方  $L_2$  范数:  $\mathcal{K}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$ 。我们将该核函数应用于数据集中的每一对数据点, 然后得到规范化的核矩阵的特征向量。

## 5.2 大统一思想

文献 [10] 总结出: ISOMAP、MDS、LLE 等非线性降维方法均可用 Kernel PCA 统一框架 (Unified Framework) 表示, 这些方法可表示为不同核函数下的 kernel PCA。

- **MDS:** 对于给定距离矩阵  $D$ , MDS 的核函数定义为

$$\mathcal{K}_{MDS} = -\frac{1}{2} (I - ee^T) D (I - ee^T)$$

- **Isomap:** 对于给定测地线距离矩阵  $D^G$ , ISOMAP 的核函数定义为

$$\mathcal{K}_{ISOMAP} = -\frac{1}{2} (I - ee^T) D^G (I - ee^T)$$

- **LLE:** 令  $\lambda_{\max}$  是  $L = (I - W)^T (I - W)$  最大的特征值, LLE 的核函数定义为

$$\mathcal{K}_{LLE} = \lambda_{\max} I - L$$

总的来说, ISOMAP、MDS、LLE 等非线性降维方法本质上是 kernel PCA 的不同表现形式。值得注意的是, 以上的核矩阵只在训练数据上定义, 而 kernel PCA 以内核的分析形式开始, 因此定义整个空间。

### 5.3 非线性降维的缺点

流形学习可以保证降维后的点保留原始高维空间全局（或局部）的几何结构, 那么实际应用上为什么 PCA 等线性降维的运用更广泛。我大概总结一下原因:

1. **自身的局限:** 流形学习欲有效进行邻域保持则需样本密采样, 而这是高维情况下面临的重大障碍; 用 kNN 算法寻找邻域是可能会出现短路 (short-circuit) 问题; 此外, 对于某些特定的非线性结构 (如: 曲率较大的闭合结构) 效果不佳;
2. **对噪声和异常值敏感:** 实际数据在获取和计算过程中不可避免会遇到噪声和异常值, 总所周知 kNN 算法对噪声和异常值异常敏感, 直接导致空间结构的紊乱, 因此基于 kNN 算法的流形学习对含噪声和异常值的样本点处理效果不好;
3. **计算量大:** 亲身体会, 对于下面的 Swiss Roll 实验, 虽然 ISOMAP 和 LLE 得到的效果会更好, 但流形学习 (ISOMAP) 的运行时间跟 PCA 不是一个数量级;
4. **高维流形到低维嵌入的显式映射:** 针对非线性流形学习, 我们无法通过线性投影获取降维后的点, 而是通过求解特征值问题得到, 因此对于新的样本点, 则需要重新按照算法步骤计算新获得的低维点。该点点已被前人进行了解决: 将非线性的方法线性化, 如 LLE 线性化得到邻域保持嵌入 (NPE), 拉普拉斯特征映射 (LE) 线性化得到局部保持投影 (LPP)。线性流形学习算法既可以保留全局或局部的几何信息, 还能学习到一个投影映射矩阵, 将高维空间上的样本点直接映射到低维空间上。这也引出了第六节的内容: 流形学习与度量学习在本质上的关系。
5. **估计本征维度:** 对未知的潜在非线性结构需要降到多少维比较合适这个问题疑惑我很久了, 文献 [11] 虽然给出 6 种方法估计本征维数, 但并没有给出一个自适应的方法来估计本征维度。

## 6 流形学习与度量学习

度量学习 (Distance Metric Learning, DML) 与流形学习类似, 通过寻找一个合适的距离度量从而寻找合适的低维空间, 使得在此空间中进行学习能比原始空间性能更好。因此度量学习也可以达到降维的目的。在此参照 CMU 的 Liu Yang 老师总结的关于 DML 的综述 [5] 和对 DML 的经典算法进行了分类总结 [4] 作为入门, 阅读文献 [6] 得到了流形学习与度量学习的本质关系。

对于两点  $x, y \in R^M$  间的距离定义为:

$$\begin{aligned}d(x, y) &= (x - y)^T A (x - y) \\ &= (Px - Py)^T (Px - Py)\end{aligned}$$

上式表明: 度量学习的目标——度量矩阵  $A \in R^{M \times M}$  转换到线性投影矩阵  $P = A^{1/2}$ , 说明学习线性投影  $P$  的线性流形学习方法可以被解释为学习如上的投影  $A^{1/2}$  矩阵, 从本质上解决了度量学习问题。因此, 任何能够学习显式投影映射的线性流形学习算法具有学习距离度量学习的等价目标。

那么非线性流形学习是否也与度量学习有关联? 对于非线性流形学习无法得到没有明确的投影映射, 为了分析它们与距离度量学习的关系, 首先, 虽然距离约束的具体形式可能会有所不同, 但我们应该认识到保持距离约束的共同性。距离度量学习方法保留二进制距离 (1 表示连接, 0 表示不连接), 例如成对约束。ISOMAP 保留了成对数据点之间的最短路径来估计的测地线距离, 而 LLE 基于邻域的局部线性组合保留距离。其次, 由 ISOMAP 和 LLE 计算的非线性表示可以理解为在距离度量学习中获得的数据表示  $(x_i - x_j) A^{1/2}$ 。最后, LLE 可以被线性近似为 NPE, 而 NPE 作为线性流形学习算法在前面的分析可知与度量学习本质上相同。

总之, 线性流形学习解决着与距离度量学习相似的问题; 而非线性流形学习也与距离度量学习有着本质的联系 [6]。

## 7 流形学习在信号处理中的应用

目前流形学习在信号处理中的研究主流方向通常是利用空间几何特性。在实际应用中, 如信号、图像等高维数据往往都分布在某个低维子流形上, 所以利用对称正定矩阵来表示图像、信号等高维信息具有优势 [12]。在早期的信号分类方法中, 信号的二阶统计信息, 即协方差矩阵, 能为信号分类提供具有鉴别能力的特征 [13]; 随着研究发现, 协方差矩阵作为对称正定矩阵, 更应该从黎曼流形中处理高维数据, 因此, 运用黎曼距离和切空间等黎曼几何中的知识用于信号分类比早期在欧式空间进行信号分类方法更好 [14]。另外, 在信号的稀疏表示上也有用对称正定矩阵做文章的, 例如: 对称正定矩阵结合字典学习能提高信号的稀疏表示的性能 [15, 16], 对称正定矩阵结合核稀疏表示 (kernel sparse representation) 提高稀疏编码的性能 [17, 18]。在矩阵低秩表示上, 在流形上处理低秩表示 (LRR) 模型从而达到聚类的目的在计算精度和时间上均有改善 [19]。

信号在分类中通常被认为是一种高维数据, 因此, 高维数据的分类方法被运用到信号分类中。目前对高维数据分类通常都先采用某种数学变换将原始高维属性空间转变为一个低维“子空间”, 在这个子空间中样本密度大幅度提高, 距离计算变得更加更容易。原始高维空间中的样本点, 在某个低维嵌入子空间中更容易进行提取特征和分类。因此在信号处理方面, 先对信号进行预处理, 降维并提取其特征以便后续操作。以下是两个研究热点:

- ♣ **空间** 早期的信号分类方法都是在欧式空间内完成的, 比如经典的 CSP[1], 该方法仅利用协方差阵获得空间滤波器, 而忽视了协方差矩阵是对称正定矩阵这一重要性质。对称正定矩阵空间是在黎曼流形上的, 故黎曼几何中的很多概念和工具可以用来处理分类问题, 比如黎曼距离 [20] 和切空间 [11]。与在欧式空间分类相比, 在黎曼流形上对信号分类结果的正确率更高, 因此基于流形学习的方法被广泛用于高维数据分类上。

♠ **降维** 在高维情况下出现数据样本稀疏、距离计算困难等被称为“维数灾难”的问题和过拟合问题是在提取特征过程中遇到的常见障碍，缓解以上问题的途径之一是降维，即在高维空间中找到一个更紧凑的低维表示。流形学习是一种借鉴了拓扑流形概念的降维方法，在局部具有欧式空间的性质，便于进行距离计算。其目标是寻找到一个低维流形嵌入，保证高维空间的样本点间的关系在低维表示中得以保持。

## 参考文献

- [1] 周志华. 机器学习 (Machine learning)[M]: 清华大学出版社, 2016.
- [2] YAO Y. A mathematical introduction to data science[J], .
- [3] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical Learning[J]. Journal of the Royal Statistical Society, 2009, 167(1): 267–268.
- [4] LIU Y. Distance Metric Learning: A Comprehensive Survey[J]. Michigan State University, 2006.
- [5] YANG L. An Overview of Distance Metric Learning[J]. Proc Computer Vision & Pattern Recognition, 2007.
- [6] LIU Y. The Connection Between Manifold Learning and Distance Metric Learning[J], 2007.
- [7] HOREV I, YGER F, SUGIYAMA M. Geometry-Aware Principal Component Analysis for Symmetric Positive Definite Matrices[C]. Asian Conference on Machine Learning. 2015.
- [8] TENENBAUM J B, DE SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. science, 2000, 290(5500): 2319–2323.
- [9] ROWEIS S T, SAUL L K. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science, 2000, 290(5500): 2323–6.
- [10] GHODSI A. Dimensionality Reduction A Short Tutorial[J]. General Information, 2006, 22(2): 183–207.
- [11] MAASTRICHT U. An Introduction to Dimensionality Reduction Using Matlab[J], 2007.
- [12] HARANDI M, SALZMANN M, HARTLEY R. Dimensionality Reduction on SPD Manifolds: The Emergence of Geometry-Aware Methods[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99): 1–1.
- [13] BLANKERTZ B, TOMIOKA R, LEMM S, et al. Optimizing Spatial filters for Robust EEG Single-Trial Analysis[J]. IEEE Signal Processing Magazine, 2007, 25(1): 41–56.
- [14] BARACHANT A, BONNET S, CONGEDO M, et al. Multiclass brain-computer interface classification by Riemannian geometry[J]. IEEE Transactions on Biomedical Engineering, 2012, 59(4): 920–928.

- [15] HARANDI M T, HARTLEY R, LOVELL B, et al. Sparse Coding on Symmetric Positive Definite Manifolds Using Bregman Divergences[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 27(6) : 1294–1306.
- [16] CHERIAN A, SRA S. Riemannian Dictionary Learning and Sparse Coding for Positive Definite Matrices[J]. IEEE Transactions on Neural Networks & Learning Systems, 2015, PP(99) : 1–13.
- [17] WU Y, JIA Y, LI P, et al. Manifold Kernel Sparse Representation of Symmetric Positive-Definite Matrices and Its Applications.[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2015, 24(11) : 3729–3741.
- [18] JAYASUMANA S, HARTLEY R, SALZMANN M, et al. Kernel Methods on the Riemannian Manifold of Symmetric Positive Definite Matrices[C]. Computer Vision and Pattern Recognition. 2013 : 73–80.
- [19] WANG B, HU Y, GAO J, et al. Low Rank Representation on SPD Matrices with Log-Euclidean Metric[J], 2017.
- [20] CHU J. Introducing Locally Linear Embedding (LLE) as a Method for Dimensionality Reduction[J], .