浙江大学

博士学位论文

论文题目：流形学习的理论与方法研究

作者姓名 王靖

指导教师 张振跃 教授

学科(专业) 计算数学

所在学院 理学院

提交日期 2006年4月

摘 要

科学的进步，特别是信息产业的发展，把我们带入了一个崭新的信息时代。在信息时代的科学研究过程中，不可避免的会遇到大量的高维数据，如全球气候模型、图像分类系统、文本聚类和基因序列的建模等。在实际应用中，用高维数据来表示的观测点可以模拟成可能带有噪音的低维非线性流形上的样本点或者逼近这些样本点。因此，数据降维尤其是非线性降维成为数据挖掘的一个重要手段，进行降维的目的是为了从高维空间中找出隐藏的低维结构。

过去几年来，非线性降维在包括数据挖掘、机器学习、图像分析和计算机视觉等许多研究领域都吸引了广泛的关注。最近，已经发展出一些有效的算法来进行非线性降维。这些算法包括等距映射（Isomap）、局部线性嵌入（LLE）以及它的变换形式：海赛局部线性嵌入（HLLE）和局部切空间排列（LTSA）等。所有的这些算法都有一个共同的特征：找出每个数据点周围的局部性质以及采用这些所收集到的局部性质信息将流形非线性的映射到一个低维的空间中。然而，这些算法的实现在收集邻域的局部信息以及采用这些收集的局部信息构造全局的嵌入上都是不同的。比如，Isomap 利用每个邻域的邻域点之间的联系和欧氏距离在数据点上构造出一个图，并根据图距离来估计出所有数据点之间的测地距离。构造出的全局低维坐标需要保持估计的测地距离。LLE找出每个点同它的邻域点之间的一个线性组合关系，并且由此决定保持这种线性组合结构的低维嵌入。LTSA将每个点的邻域点投影到它在流形上的局部切空间上，然后排列所有的局部坐标来决定低维的全局坐标。显然的，局部几何结构的恢复效率决定了这些算法的效率。

LLE 是流形学习方面经典的局部非线性方法，它有参数少、计算快、易求全局最优解等优点，并在图像分类、图像识别、谱重建、数据可视化等方面都有着广泛的应用。但是，LLE 可能会将相隔较远的点映射到低维空间中邻近点的位置，从而导致嵌入结果有着比较明显的扭曲。这其中的一个重要原因是，LLE 采用的单个重构权并不能完全的反映出流形的局部几何性质。此外，用以求解重构权的有约束的最小二乘问题的最优解也许不是唯一的，而且 LLE 采用正则化方法求解涉及到正则因子 γ 的选取，难以保证所求的解是最优解。这些都是 LLE 所面临的问题。

有一些共同的因素影响着流形学习方法的效果。流形学习方法首先面临的是邻域选取的问题，需要选取出一个合适的邻域以获取局部的线性信息，邻域选取的结果直接影响着最终的嵌入结果。很显然的，邻域越小可以认为邻域的线性结构越明显，但是我们

需要注意的是，邻域之间需要有足够的交叠以保证较远的点之间有足够的联系，这又使得邻域不能过小。从直观上想像，流形上曲率大的样本点的邻域应该小一些，而流形上曲率小的样本点处的邻域可以大一些。因此，关于邻域选取的主要困难在于：在加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质？此外，流形上的曲率以及样本点密度的变化，不可避免的会使得所寻找的局部邻域结构产生偏差。在利用这些局部邻域结构来构造全局的低维嵌入时，需要将这些偏差计入考虑范围。因此，流形学习还面临着这些问题：如何估计流形上的曲率？如何估计流形曲率和样本点密度的变化对寻找局部邻域结构的影响？在利用局部邻域结构来构造全局的低维嵌入时，应该如何计入这种影响以减轻低维嵌入的偏差？对于上述问题，本文给出了较完善的解答。本文的主要贡献如下：

1. 我们对 LLE 的重构权向量的性质进行详细的分析，在理论上证明了（用正则化方法）确定最优权在数值上是不稳定的，同时在给定精度下，存在着多组线性无关的近似最优权向量。

2. 我们提出了修正的局部线性嵌入方法（Modified Locally Linear Embedding Using Multiple Weights，简称MLLE），采用线性无关的权向量来建立邻域内稳定的局部线性结构，并在低维嵌入中保持这种局部线性结构。MLLE 改善了 LLE 方法的稳定性和有效性。

3. 我们从理论上证明了 MLLE 对采自等距流形的样本点有着理想的结果，通过 MLLE 和 LTSA 之间的详细对比和理论分析，揭示了 LLE、MLLE 和 LTSA 之间的内在联系。这为进一步理解与分析建立了基础。

4. 我们提出了自适应邻域选取的方法，以解决非线性降维方法中面临的邻域选取的难题。基于邻域局部线性逼近的分析，我们给出了决定邻域集是否能在一个给定精度内被一个线性拟合所逼近的标准。进而提出两个算法（采用邻域压缩和邻域扩张策略）来选取能够满足这个标准的邻域。我们的方法从理论上保证了所选出的邻域在匹配流形的局部几何性质的前提下，能够尽可能地扩张邻域以加强样本点之间的关联性。自适应邻域选取方法能适用于所有基于邻域的流形学习方法。

5. 我们给出估计流形局部曲率的方法，并通过引入流形的局部曲率来修正 LTSA 中的极小化模型。这个改进能减少 LTSA 造构造全局嵌入的偏差。结合自适应的邻域选取和曲率修正，我们提出了一种自适应的流形学习方法－自适应局部切空间排列方法（Adaptive local tangent space alignment，简称ALTSA）。虽然曲率模型是针对 LTSA 而设计，但我们相信所提出的基本思想也能适用于其它的流形学习方法。

6. 我们给出了大量的数值例子（模拟例子和实际例子），通过将我们提出的算法与

Isomap、LLE 和 LTSA 的对比和分析，从数值上说明了本文所提出的这些新方法的有效性。

关键词：重构权、局部切空间排列、自适应邻域选取、减少偏差、曲率和切空间

Abstract

With the progress of science, especially the development of informational industry, we enter a brand-new information age. When doing research in information age, one is inevitably confronted with large volumes of high-dimensional data, such as global climate patterns, image classification system, text clustering and gene expression. In real-world applications, observations represented as high-dimensional data or vectors can be modeled as samples lying on or close to a low-dimensional nonlinear manifold possibly with noise. Hence, data reduction especially nonlinear dimensionality reduction is an important tool of data mining, and the goal of dimension reduction is to find the low dimensional structure of the nonlinear manifold from the high dimensional data.

In passed years, the problem of nonlinear dimensionality reduction has aroused a great deal of interest in many research fields including data mining, machine learning, image analysis, and computer vision. Recently, there have been advances in developing effective and efficient algorithms to perform nonlinear dimension reduction. The proposed algorithms include isometric mapping (ISOMAP), locally linear embedding (LLE) and its variations Hessian LLE and local tangent space alignment (LTSA). All these algorithms share a common characteristic: find a local geometry around each data point and then use the collected local geometric information to nonlinearly map the manifold to a lower dimensional space. The performances of these algorithms, however, are different in collecting local information within neighborhoods and in constructing a global embedding using the collected local information. For example, ISOMAP constructs a graph on the data points using the connections and Euclidean distances between neighbors in each neighborhood and estimates the global geodesic distance between all the data points in terms of the graph distance. Global low dimensional coordinates are constructed to preserve the estimated geodesic distances. LLE finds a linear combination of each points with respect to its neighbors and determines a low dimensional embedding that preserves linear combination structures. LTSA projects all neighbors of each point onto its local tangent space of the manifold and then aligns the local coordinates to determine low dimensional global coordinates. Clearly the effec-

tiveness of the local geometry retrieved will determines the efficiency of the methods.

LLE is a classical local nonlinear approach in manifold learning, and it has many applications such as image classification, image recognition, spectra reconstruction and data visualization because of its few parameters, rapid computation, and global optimization. However, LLE may map faraway inputs to nearby outputs in the low dimensional space which lead to a visible distortion in the embedding results. One of the curses that make LLE fail is that only one set of reconstruction weights can not reflect the local geometry of the manifold. Furthermore, the constrained least squares problem for finding the weights may do not have a unique solution. Using regularization method to solve the constrained LS problem is involved in the selection of a regularization term γ, and we can not verify that whether the computed solution is optimal. These questions remain to be answered.

There are some common issues that determine the effectiveness of the manifold learning algorithms. The first step of manifold learning approaches is the selection of neighborhood, and there is a need to select an appropriate neighborhood to find local linear information. The effectiveness of the selection of the neighborhood will determine the embedding results. Clearly the smaller the neighborhood is, the more distinct the linear structure of the neighborhood will be. However, we should notice that the neighborhoods should be fully overlapping such that the faraway points are fully connected, which lead that the neighborhood can not be too small.By intuition, larger curvature tends to shrink the neighborhood, while smaller curvature gives rise to a larger neighborhood. So the main difficulty of the selection of neighborhood is that when imposing the connectivity structure on the sample points, how to adaptively select the neighborhood sizes to match the local geometry of the manifold. Furthermore, the variation in the curvature of the manifold and the sampling density of the data points will inevitably lead to a bias in the collected local structure. There is a need to take the bias into account when using the collected local structures to construct global low dimensional embedding. Hence these questions in manifold learning remain to be answered: how to estimate the curvature of the manifold, and how to account for the bias in the collected structure caused by the variation in the curvature of the manifold and the sampling density of the sample points, and how to reduce the bias in the low dimensional embedding by accounting for the bias in the collected local structure when using the col-

lected local structure to construct the global low dimensional embedding. To confront these proposed problems, we give relatively consummate answers. The contributions of this paper are as follows:

1. We give a detailed analysis to characterize the reconstruction weights and show that (using the regularization method) to obtain the optimal weight is not stable in numerical computation, and there exist multiple linearly independent weights which approximate to be optimal within a given accuracy.

2. The proposed modified locally linear embedding using multiple weights (MLLE) use the linearly independent solutions to establish stable local linear structure and preserves the local linear structure in the low dimensional embedding. MLLE improves the stability and effectiveness of LLE.

3. We show that MLLE can retrieve the ideal embedding for data points sampled from an isometric manifold, and we give the detailed comparisons between MLLE and LTSA to discover the internal connection between LLE, MLLE and LTSA. It lays the foundation of the further comprehension and analysis.

4. We propose the adaptive neighborhood selection strategy to solve the problem of the selection of neighborhood. Based on the analysis of the local linear approximation of neighborhood, we give a criterion for deciding whether a neighbor set can be well approximated within a give accuracy by a liner fitting. Then we propose two algorithms for selecting neighbor sets that satisfy the criterion using the strategies of neighborhood contraction and neighborhood expansion. We show that the selected neighbor set by our methods can be expanded as large as possible to reinforce the connectivity between the sample points which can fit the local geometry of the manifold. These adaptive neighborhood selection methods can be used for all neighborhood-based manifold learning methods.

5. We propose a method to estimate the local curvatures of the manifold and modify the minimization model in LTSA by tacking into account local curvatures of the manifold. This improvement can reduce the bias in construction of global embedding by LTSA. The curvature model can be used together with the adaptive neighborhood selection strategy, and we propose an adaptive manifold learning method, namely, adaptive local tangent space alignment (ALTSA). Although the curvature model is specially designed for LTSA, we believe that the basic ideas we propose can be adapted to other

manifold learning algorithms as well.

6. We give plentiful examples(synthetic examples and real-world examples) to compare our proposed algorithms with Isomap, LLE and LTSA and the experiment results show the efficiency of the proposed new methods in this paper.

Key Words: **reconstruction weights, Local tangent space alignment, adaptive neighborhood selection, bias reduction, curvature and tangent space**

第 1 章　引言

本章要点：

- 数据降维的目的和应用
- 线性降维
- 非线性降维
- 本文的研究目标、范围与组织

§1.1　数据降维的目的和应用

科学的进步，特别是信息产业的发展，把我们带入了一个崭新的信息时代。大量信息资源在给人们带来方便的同时也带来了一大堆的难题：信息过量，难以消化；信息繁杂，真假难以辨识；信息形式不一致，难以统一处理；有价值的信息淹没在海量的数据之中，难以取舍，等等。与此同时，随着数据库技术的迅速发展和广泛应用，人们积累的数据越来越多，但由于缺乏挖掘数据背后隐藏的知识的手段，人们无法发现数据中存在的关系和规则，也无法根据现有的数据预测未来的发展趋势。由于缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。

面对海量数据库和大量繁杂信息，如何才能从中提取出有价值的知识，并进一步提高信息的利用率，引发了一个新的研究方向：基于数据库的知识发现（Knowledge Discovery in Database，简称 KDD）以及相应的数据挖掘（Data Mining）理论和技术的研究。数据挖掘和知识发现是多学科交叉的产物，它综合了人工智能、机器学习、数据库和数据仓库、统计学、高性能计算等技术。作为大规模数据库中先进的数据分析工具，数据挖掘与知识发现的研究已经成为数据库以及人工智能领域研究的一个热点。

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。原始数据可以是结构化的，如关系数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。定义同时包括几层含义：数据源必须是真实的、大量的、有噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用。其实，利用数据挖掘工具从数据集中所发现的知识，是有特定前提和约束条件的、面向特定领域的，同时还要能够易于被用户理解，最好能用自然语言表达所发现的结果。发现的知识可以被用于信息管理、查询优化、决策支持、

过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门很广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。

数据挖掘的一个非常重要的工具和方法是数据降维。数据降维的目的是找出高维数据中的所隐藏的低维结构。现实生活中的许多现象可以用高维数据来描述。比如天气状况，随着气象学的发展，现在用来描述气象特阵的指标非常多，例如温度、湿度、气压、风力、降雨量、辐射强度等，将这些用多个变量描述现象的数据，抽象出来就是高维数据。高维数据提高了有关客观现象的极其丰富、详细的信息。但是，数据维数的大幅度提高给随后的数据处理工作带来了前所未有的困难，即如何从大量的数据特征中找出其本质的或用户有兴趣的特征。这就需要对数据集进行降维处理，并且在降维后能保持数据集原有的一些本质特征不变。我们接下来考虑数据降维的一些典型应用：

1. 图像分类系统。假设有一组数字图像，这组图像共有 N 张图像，每张图像都是 $m*n$ 大小的灰度图。可以对每张图像按行或列堆叠转化成一个列向量，而列向量的维数为 $m*n$，每个分量的大小表示图像的像素点的像素值的大小。对于 $m=n=256$ 的图像，最后转化为 65536 维的向量。对于如此高的维数，常用的分类方法都无法有效的工作，因此，我们需要降低维数，而且在降低维数的同时，能保持图像之间所隐藏的本质关系，比如图像的拍摄角度、光线亮度等等，然后再进行分类识别。

2. 文本分类系统。文本分类是指根据文本的内容和属性，将大量的文本归到一个或多个类别的过程。文本分类技术通过分析待分类对象，提取出分类对象特阵，比较待分类对象和系统预定义类别对象的特征，将待分类对象划归为特征量最相近的一类，并赋予相应的分类号。计算机并不具有人类的智能，从根本上说，它只认识 0 和 1，所以必须将文本转换为计算机可以识别的格式。假定组成文本的字或词在确定文本类别的作用上相互独立，这样，可以就使用文本中出现的字或词的集合来代替文本。在信息处理方向上，向量来表示文本信息：首先，替文本建立一个词库，词库中词的个数 m 为文本的维数；然后就可以利用词库来构造向量 $(w_1,\ldots,w_m)^T$ 表示文本信息，其中 w_i 表示了词库中第 i 个词在文本中出现的次数。很显然，构造文本的词汇量是相当大的，因此，文本向量的维数 m 也是巨大的。对此，我们有必要进行数据降维，以提高算法效率和运行速度，再对降维的结果进行分类。

3. 基因序列的建模。蛋白质是由氨基酸组成的序列，氨基酸分子的个数从几十个到成千上万不等。具有相同空间结构但氨基酸排列不同的蛋白质，被分为同一组中，这就是所谓的蛋白质组（类似于基因组）。通过蛋白质组模型可以了解不同蛋白质组的特殊的性质，能够有助于辨别和发现新组。但由于蛋白质组特征的维数很高，这给辨别和分

析带来了很大的困难。通过数据降维，可以用很少的简单变量来反映蛋白质组的性质，以利于辨别和分析。

由于真实世界中的数据往往是高维的，而高维的数据难以被人理解、表示和处理，因此需要采用数据降维以获得低维的数据。经过降维的数据可以更好的进行分析，因此对降维问题的研究成为机器学习和数据挖掘中的重要主题。数据降维算法可以分成两类，一类是线性降维方法，如主成分分析法 [31]、多维尺度算法 [22, 65]和非负矩阵分解 [46]等；另一类是非线性降维方法，如等距映射算法 [60]、局部线性嵌入法 [56]和局部切空间排列方法 [67]等。

§1.2 线性降维

线性降维是指通过降维所得到的低维数据能保持高维数据点之间的线性关系。假设 N 个维数为 m 的高维数据点为 $x_1, \ldots, x_N$，降维后得到的维数为 d（$d \ll m$）的低维结果为 $\tau_1, \ldots, \tau_N$，若存在线性映射 f 使得 $f(\tau_i) = x_i, i = 1, \ldots, N$，则这个高维数据点 x_i 降到低维的过程为线性降维。

主分量分析法（Principal Component Analysis, 即PCA）是使用最广泛的线性降维方法之一。Hotelling 于1933年提出了主成分分析法，将方差的大小作为衡量信息量多少的标准，认为方差越大提供的信息越多，反之提供的信息越少。PCA 通过原分量的线性组合构造方差大、含信息量多的若干主分量，从而降低数据的维数。PCA 的计算过程可以通过矩阵奇异值分解（Singular Value Decomposition）来实现。假设数据点 x_i 采自 d 维仿射空间，即

$$x_i = c + U\tau_i + \epsilon_i, \quad i = 1, \ldots, N,$$

其中 $c \in R^m, \tau_i \in R^d$，$\epsilon_i$ 表示噪音，$U \in R^{m \times d}$ 是这个仿射空间的一组正交基组成的矩阵。记

$$X = [x_1, \ldots, x_N],\ T = [\tau_1, \ldots, \tau_N],\ E = [\epsilon_1, \ldots, \epsilon_N],$$

则数据点可以用如下矩阵形式表示，即

$$X = c\mathbf{1}_N^T + UT + E,$$

其中 $\mathbf{1}_N$ 表示所有分量为 1 的 N 维列向量。PCA 就是寻找 c, U, T 以极小化重建误差

$$\min \|E\| = \min_{c,U,T} \|X - (c\mathbf{1}_N^T + UT)\|_F,$$

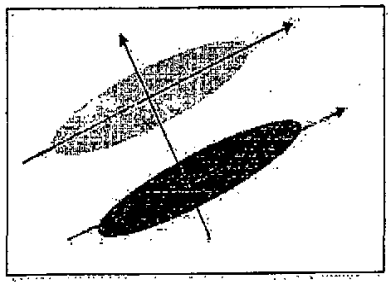

图 1.1: PCA 对椭球状分布的样本集学习所得的主要方向是椭球的主轴方向。

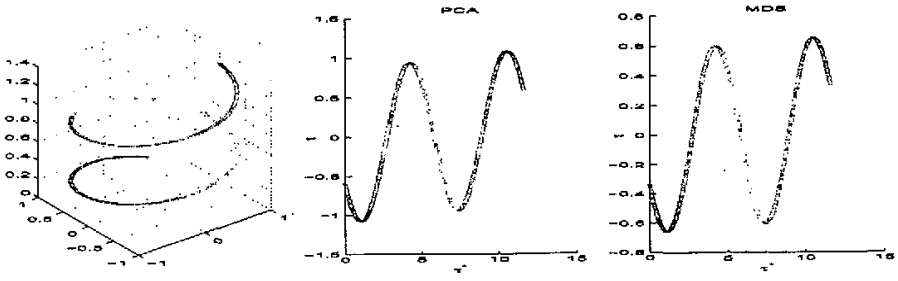

图 1.2: PCA和MDS对圆柱螺线的结果：第一列（圆柱螺线）；第二列（PCA的结果）和第三列（MDS的结果）为嵌入结果 τ vs. 圆柱螺线的弧长 τ^*。

其中 $\|\cdot\|_F$ 表示矩阵的 Frobenius 范数。这个极小化问题通过 SVD 很容易的求解:

1）X 中心化，即选取 $c = X\mathbf{1}_N/N = \bar{x}$。

2）选取低秩矩阵 UT 为中心化矩阵 $X - \bar{x}\mathbf{1}_N^T$ 的最佳秩 d 逼近。这样，最优解可以通过矩阵 $X - \bar{x}\mathbf{1}_N^T$ 的奇异值分解得到，即

$$X - \bar{x}\mathbf{1}_N^T = Q\Sigma V^T, Q \in R^{m\times m}, \Sigma \in R^{m\times N}, V \in R^{N\times N},$$

则$UT = Q_d\Sigma_d V_d^T$，其中 $\Sigma_d = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ 为矩阵 $X - \bar{x}\mathbf{1}_N^T$ 的 d 个最大奇异值所构成的对角阵，Q_d 和 V_d 分别为对应的左右奇异值向量所组成的矩阵。显然，最优解 $U = Q_d$，降维结果 $T = U^T(X - \bar{x}\mathbf{1}_N^T) = \Sigma_d V_d^T$。

如前面所述，PCA 的一个主要性质是它是从 R^d 到 R^m 在投影距离最小的意义下的最佳线性映射。它对于椭球状分布的样本集有很好的学习效果，学习所得的主要方向就是椭球的主轴方向，如图 1.1所示。它的不足之处主要在于：

1. 当样本点具有一些非线性的性质时，采用 PCA 得到的降维结果无法反映出样本点之间所隐藏的非线性性质。如图 1.2的第二列所示，PCA无法恢复出非线性的圆柱螺线样本点之间的弧长关系。

2. PCA 能找到很好的代表所有样本点的方向，但这个方向对于分类未必是最有利的。

3. 对 PCA 所要保持的主分量的个数的估计比较困难。虽然有时我们可以通过样本点的中心化矩阵的相邻奇异值之间的比值大小来估计所要保持的主分量的个数，但是当奇异值的大小变化比较平缓的时，将难以估计应该舍弃哪些分量。此外，有时候 PCA 舍弃的分量也会有意义。

4. 在有些情况下，难以对 PCA 所保持的主分量的意义进行解释。比如，在对图像进行降维处理的时候，每个样本点是由图像的像素值矩阵转化而成，为非负的向量。当对这些非负的样本点采用 PCA 以希望获得一组基图像时，结果中的负值难以进行语义上的解释。

多维尺度变换（Multidimensional Scaling，简称MDS）也是有广泛应用的线性降维方法 [22, 65]。MDS原用于差异性的几何表示，用在降维上，是将高维点的欧氏距离矩阵作为差异性矩阵。用 $d(x_i, x_j)$ 表示样本点 x_i 与 x_j 的距离，即

$$d(x_i, x_j)^2 = \|x_i - x_j\|^2 = x_i^T x_i - 2x_i^T x_j + x_j^T x_j.$$

记 N 维向量 ψ 为 $\psi = [x_1^T x_1, \ldots, x_N^T x_N]^T$，则距离矩阵 $D = (d^2(x_i, x_j))_{i,j=1}^N$ 能重新写成

$$D = \psi 1_N^T - 2X^T X + 1_N \psi^T.$$

不失一般性，假设样本点被中心化，即 $\sum_{i=1}^N x_i = 0$，则有

$$H \equiv -(I - 1_N 1_N^T/N) D (I - 1_N 1_N^T/N)/2 = X^T X.$$

记 H 的特征值分解为

$$H = U \operatorname{diag}(\lambda_1, \ldots, \lambda_N) U^T,$$

其中 $U \in R^{N \times m}$ 为正交阵以及特征值 $\lambda_1 \geq \cdots \geq \lambda_N$ 降序排列，则

$$T = \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2}) U_d$$

为降维的结果，其中 U_d 为最大的 d 个特征值所对应的特征向量所构成的矩阵。由于欧氏距离不能真正反映出非线性的样本点之间的距离关系，而 MDS 的嵌入结果是保持高维样本点之间的欧氏距离，因此对于非线性流形上的数据， MDS 不能恢复出其中的低维结构。如图 1.2的第三列所示，MDS 无法恢复出非线性的圆柱螺线样本点之间的弧长关系。需要注意的是，在降维上 MDS 与 PCA 有着相同的结果。

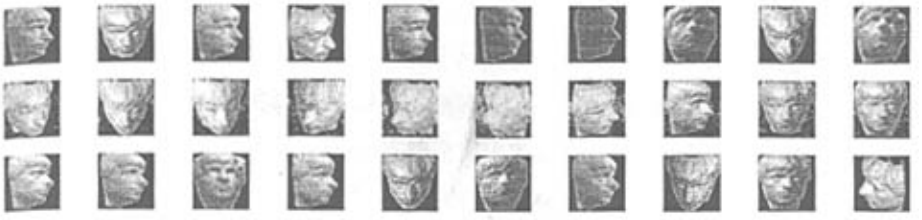

图 1.3: 由拍摄位置、光亮等参数所决定的人脸图像。

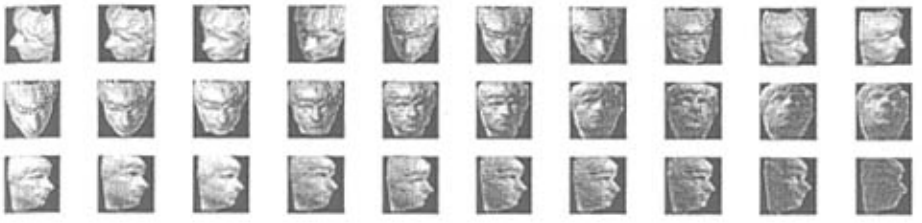

图 1.4: 人脸图像按照从左到右的拍摄角度变化（第一行）、从上到下的拍摄角度变化（第二行）以及按照光线亮度变化（第三行）排列。

经典的线性降维方法，实现简单且容易计算，通过特征的线性组合来降维，可以确保发现处于高维向量空间的线性子空间上的数据集的真实几何结构。但是这类算法的线性本质使其无法揭示复杂的非线性流形结构，并且现实中数据的有用特征往往不是特征的线性组合。如图 1.3中的人脸图像，它由光线亮度、拍摄角度等少数几个参数所决定。在图 1.4中，我们将这些脸部图像按照从左到右、从上到下的拍摄角度变化以及按照光线亮度变化重新排列。从图中我们可以看出，这些脸部图像之间的关系可以用拍摄角度、光线亮度等几个的特征来描述，但这些特征并不能由特阵分量的线性组合决定。为此，人们提出了非线性降维方法以处理非线性的高维空间中的数据点。

§ 1.3 非线性降维

§ 1.3.1 流形学习中的一些数学定义

非线性降维就是通常所说的流形学习，这里先给出流形学习中的一些数学定义[4, 72]:

拓朴。一个拓朴空间就是一个集对 (X,τ)，其中集合 X 为一非空集合，拓扑 τ 是 X 的满足以下性质的子集族:

(1) τ 关于属于它的任意多元素的并运算是封闭的;

(2) τ 关于属于它的有限多元素的交运算是封闭的;

(3) τ 含有空集 $\emptyset$ 和 X 本身作为其元素。

Hausdorff 空间。如果对 X 中任意两个不同点 x, y，都存在 x 的邻域 U 以及 y 的邻域 V 使得 $V \cap U = \emptyset$。此时，称 (X, τ) 为 Hausdorff 空间。

流形。设 M 是一个 Hausdorff拓扑空间，若对每一点 $p \in M$，都有 p 的一个开邻域 U，它与 R^d 的某个开子集同胚，则称 M 为 d 维拓扑流形，简称为 d 维流形。

微分流形。一个 d 维 C^k 流形就是一对 (M, Λ)，其中 M 为 d 维流形，$\Lambda = \{(U_\alpha, \varphi_\alpha)\}_{\alpha \in A}$ 为一 C^k 的**微分结构**，即满足以下条件：

（1）（局部欧氏性）$\{U_\alpha : \alpha \in A\}$ 构成 M 的一个开覆盖，$\varphi_\alpha : U_\alpha \to \varphi_\alpha(U_\alpha) \subset R^d$为同胚映射；

（2）（C^k相容性）若 $U_\alpha \cap U_\beta \neq \emptyset$，则双射

$$\varphi_\alpha \circ \varphi_\beta^{-1} : \varphi_\beta(U_\alpha \cap U_\beta) \to \varphi_\alpha(U_\alpha \cap U_\beta)$$

和它的逆映射都是 k 次可微的，则称$(U_\alpha, \varphi_\alpha)$与$(U_\beta, \varphi_\beta)$ 是相容的；

（3）（最大性）若 U 为 M 中的开集，$\varphi : U \to \varphi(U) \subset R^d$与 Λ 中的每个 $(U_\alpha, \varphi_\alpha)$ 都相容，则 $(U, \varphi) \in \Lambda$。

当 $r = \infty$ 时，称 M 为**光滑流形**。若 $x \in U$，则称 $(U, \varphi) \in \Lambda$ 为 x 处的一个**局部坐标系**，U 为**坐标邻域**，φ为**坐标函数**。

光滑映射。设 M 和 N 为两个光滑流形，$g : M \to N$ 是连续映射。设 $x \in M$，若存在 M 在点 x 处的局部坐标系 (U, φ) 及 N 在点 $g(x)$ 处的局部坐标系 (V, ψ)，使得

$$\psi \circ g \circ \varphi^{-1} : \varphi(U \cap g^{-1}(V)) \to \psi(V)$$

是在点 $\varphi(x)$ 处光滑的映射，则称映射 g 在点 x 处是光滑的。处处光滑的映射称为光滑映射。这种光滑映射的全体记为 $C^\infty(M, N)$。当$N = R$时，记 $C^\infty(M) = C^\infty(M, R)$为光滑流形 M 上的**光滑函数**的全体。

切向量和切空间。光滑流形 M 在点 x 的切向量就是一个映射 $v_x : C^\infty(M) \to R$，且对 $\forall g, h \in C^\infty(M), a, b \in R$ 满足：

(1) $v_x(ag + bh) = av_x(g) + bv_x(h)$;

(2) $v_x(gh) = v_x(g)h(x) + g(x)v_x(h)$。

假设 (U, φ) 为点 x 的一个局部坐标系，则映射

$$(\frac{\partial}{\partial x_i})_x : g \longrightarrow (\frac{\partial g}{\partial x_i})_x \equiv (\frac{\partial (g \circ \varphi^{-1})}{\partial x_i})(\varphi(x)), \quad g \in C^\infty(M)$$

为 x 点的一个切向量。光滑流形的切向量是曲线的切向量的一种推广。x 点的切向量全体记为 $T_x(M)$，它是一个实线性空间，称之为 M 在点 x 的切空间。

黎曼流形。如在光滑流形 M 的每个切空间 $T_x(M)$ 中都给定了内积，则称 M 为黎曼流形。

弧长。设 $C(t), a \leq t \leq b$是黎曼流形 M 中的一条曲线，在 C 上每点的切向量记为 v_t，则可以定义曲线 C 的弧长 $S(C)$ 为 $S(C) = \int_a^b \|v_t\| dt$。

测地距离。设 p、q 是黎曼流形 M 中任何两点，则这两点间的测地距离 $d_M(p,q)$ 为 M 中连接 p、q 的所有分段光滑曲线的弧长的下确界。

等距流形。设 M 为 d 维的黎曼流形，若存在光滑映射 $g: M \to R^d$ 满足：

(1) $g: M \to g(M)$ 为同胚；

(2) 对任意的$p, q \in M$，有 $d_M(p,q) = \|g(p) - g(q)\|$，

则称 M 为 d 维的等距流形。

§1.3.2 流形学习

近年来，流形学习在包括数据挖掘、机器学习、图像分析和计算机视觉等许多研究领域都吸引了广泛的关注并得到了广泛的应用 [14, 30, 38, 49]。关于流形学习方面最有影响的文章是2000年 J.B. Tenenbaum 等和 S. T. Roweis 等人在Science同一期上发表的两篇文章，他们提出了各自的流形学习算法：等距映射（Isomeitric Mapping，简称为Isomap）[60]和局部线性嵌入（Locally Linear Embedding，简称为LLE）[56]。

等距映射 (Isomap) 建立在多维尺度变换 (MDS) 的基础上，力求保持数据点的内在几何性质，即保持两点间的测地距离 [60]。它用流形上点 x_i 和 x_j 的测地距离取代经典的 MDS 方法中的欧氏距离 $d(x_i, x_j)$。测地距离的近似计算方法如下：样本点 x_i 和它的邻域点之间的测地距离用它们之间的欧氏距离来代替；样本点 x_i 和它邻域外的点用流形上它们之间的最短路径来代替。Isomap 是一种全局优化算法，对于等距流形，它能够给出等距低维投影。由于测地距离的整体性，Isomap 的降维效果在整体性上把握得很好，即使流形不完全是等距的。但 Isomap 有如下缺点：

（1）Isomap 要求所对应的低维等距子集是一个凸集。当这个低维子集非凸时，由于非凸性导致了无法得到测地距离的可接受的逼近，Isomap 无法得到理想的嵌入结果。

（2）当样本点的密度不大，或分布不均匀时，测地距离的计算有一定的（较大）误差，这使得 Isomap 的计算结果有较多的空洞现象。

（3）计算流形距离的逼近的计算量较大，导致算法实现所需的时间较多，这一点在在样本点个数较多的时候特别明显。

局部线性嵌入 (LLE) 的基本思想是在样本点和它的邻域点之间构造一个重构权向量，并在低维空间中保持每个邻域中的权值不变，即假设嵌入映射在局部是线性的条件下，最小化重构误差 [56, 57]。LLE 认为所构造的重构权能掌握局部邻域的本质上的几何性质—即那些在平移、旋转、缩放中保持不变的性质。LLE 首先在每个样本点寻找它的最近邻域，然后通过求解一个有约束的最小二乘问题以获得重构权。在求解这样的最小二乘问题时，LLE 将求解最小二乘问题转化成求解一个可能奇异的线性方程组，并通过引入一个小的正则因子 γ 来保证线性方程组系数矩阵的非奇异性。求出重构权后，利用这些重构权可以构造一个稀疏矩阵，LLE 通过求解这个稀疏矩阵的最小的几个特阵向量来获得全局的低维嵌入。同 Isomap 相比，求解特征值所涉及到的只是一个稀疏矩阵，给计算带来了很大的便利。此外，LLE 并不需要计算样本点之间的测地距离，而只需要在每个邻域求解一个小的线性方程组，所需要的计算时间和计算量小的多。但 LLE 也有不足之处，由于 LLE 保持邻近点的几何性质，对于有噪音、样本密度稀疏或者相互关联较弱的数据集，相隔较远的点之间的关联会减弱，这样在从高维到低维的映射过程中，很可能会将相隔较远的点映射到邻近点的位置。而 Isomap 由于是保持样本点间的测地距离而不会出现这样的情况。另外，LLE 通常不能恢复出与流形等距的低维嵌入。

在 Isomap 和 LLE 之后，发展出了许多新的流形学习算法，如拉普拉斯特征映射（Laplacian Eigenmap）[5]、海赛局部线性嵌入（Hessian Locally Linear Embedding，简称为HLLE）[24]、局部切空间排列（Local tangent space alignment，简称为LTSA）[67]等。Laplacian Eigenmap 的降维目标是在高维空间中离得很近的点投影到低维空间中的象也应该离得很近 [5]。它从样本点构建近邻图，图的点为样本点，每个样本点 x_i 同它的邻域点 x_j 之间有边连接。然后给每条边赋予权值，权值为 1 或者 $e^{-\|x_i-x_j\|^2/2\sigma^2}$。最后通过计算图拉普拉斯算子的广义特征向量来得到低维嵌入结果。Laplacian Eigenmap 利用拉普拉斯—贝尔特拉米算子的性质来构造流形的嵌入映射。这个映射试图保持一些局部性质，而这个性质通常并不是局部的等距性。对于一个同低维空间等距的流形，Laplacian Eigenmap 并不一定能得到等距的嵌入结果。LLE 所面临的将远距离点映射到近距离的问题也可能会出现。此外，Laplacian Eigenmap 还涉及到参数 σ 的选取，不同的参数 σ 的嵌入结果可能不同。

LTSA 的基本思想是利用样本点邻域的切空间来表示局部的几何结构，然后将这些局部切空间排列起来构造流形的全局坐标 [67]。首先，LTSA 寻找出样本点的局部邻域并在每个邻域作 PCA 以获得样本点的切空间和邻域在这个切空间上的投影坐标。其

次，LTSA 认为理想的低维嵌入同局部的投影坐标之间应该只相差一个仿射变换，并由此构造一个最小化重构误差。而求解这个最小化重构误差问题也可以转化成求解一个稀疏矩阵的特征值问题。同 LLE 和 Laplacian Eigenmap 相比，LTSA 能很好的恢复出等距流形的低维嵌入。但它也有一些不足之处。与 Isomap 和 LLE 一样，LTSA 在每个局部邻域计算投影坐标时，会比较大程度地受到流形的曲率和样本点密度的影响，从而导致计算出的局部投影坐标出现偏差，进而影响全局的嵌入结果。此外，当由于数据误差等因素，使得计算流形的局部线性关系时存在较大的偏差，这可能使得 LTSA 的嵌入结果不理想。

这些流形学习方法都有一些共同的特征：1）在每个样本点附近寻找一个邻域，并找出这个局部邻域的几何结构；2）利用这些收集的局部信息将流形非线性的映射到一个低维空间上 [33]。这些算法之间的不同之处只是在于收集的邻域的局部信息不同，并且如何利用这些收集的信息来构造全局的嵌入。它们的共同优势在于：1）它们都是非参数的方法，不需要对流形的很多的参数假设；2）它们是非线性的方法，都基于流形的内在几何结构，更能体现现实中数据的本质；3）它们的求解简单，都转化为求解特征值问题，而不需要用迭代算法，并且避免了局部极值问题。

§1.4 本文的研究动机、目标和范围

LLE 是流形学习方面经典的局部非线性方法，它有参数少、计算快、易求全局最优解等优点，并在图像分类、图像识别、谱重建、数据可视化等方面都有着广泛的应用 [23, 40, 52]。但是，LLE 可能会将相隔较远的点映射到低维空间中邻近点的位置，从而导致嵌入结果有着比较明显的扭曲。这其中的一个重要原因是，LLE 采用的单个重构权并不能完全的反映出流形的局部几何性质。此外，用以求解重构权的有约束的最小二乘问题的最优解也许不是唯一的，而且 LLE 采用正则化方法求解涉及到正则因子 γ 的选取，难以保证所求的解是最优解。本文将对 LLE 的这些缺点进行分析，并提出一个新的算法以克服这些缺点。

有一些共同的因素影响着流形学习方法效果。流形学习方法首先面临的是邻域选取的问题，需要选取出一个合适的邻域以获取局部的线性信息，邻域选取的结果直接影响着最终的嵌入结果。很显然的，邻域越小可以认为邻域的线性结构越明显，但是我们需要注意的是，邻域之间需要有足够的交叠以保证较远的点之间有足够的联系，这又使得邻域不能过小。从直观上想像，流形上曲率大的样本点的邻域应该小一些，而流形上曲率小的样本点处的邻域可以大一些。因此，关于邻域选取的主要困难在于：在加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质？此

外，流形上的曲率以及样本点密度的变化，不可避免的会使得所寻找的局部邻域结构产生偏差。在利用这些局部邻域结构来构造全局的低维嵌入时，需要将这些偏差计入考虑范围。因此，流形学习还面临着这些问题：如何估计流形上的曲率？如何估计流形曲率和样本点密度的变化对寻找局部邻域结构的影响？在利用局部邻域结构来构造全局的低维嵌入时，应该如何计入这种影响以减少低维嵌入的偏差？本文将详细讨论这些问题，并给出方案来解决这些问题。

§1.5 本文的主要结果

1. 我们对 LLE 的重构权向量的性质进行详细的分析，在理论上证明了（用正则化方法）确定最优权在数值上是不稳定的，同时在给定精度下，存在着多组线性无关的近似最优权向量。

2. 我们提出了修正的局部线性嵌入方法（Modified Locally Linear Embedding Using Multiple Weights，简称MLLE），采用线性无关的权向量来建立邻域内稳定的局部线性结构，并在低维嵌入中保持这种局部线性结构。MLLE 改善了 LLE 方法的稳定性和有效性。

3. 我们从理论上证明了 MLLE 对采自等距流形的样本点有着理想的结果，通过 MLLE 和 LTSA 之间的详细对比和理论分析，揭示了 LLE、MLLE 和 LTSA 之间的内在联系。这为进一步理解与分析建立了基础。

4. 我们提出了自适应邻域选取的方法，以解决非线性降维方法中面临的邻域选取的难题。基于邻域局部线性逼近的分析，我们给出了决定邻域集是否能在一个给定精度内被一个线性拟合所逼近的标准。进而提出两个算法（采用邻域压缩和邻域扩张策略）来选取能够满足这个标准的邻域。我们的方法从理论上保证了所选出的邻域在匹配流形的局部几何性质的前提下，能够尽可能地扩张邻域以加强样本点之间的关联性。自适应邻域选取方法能适用于所有基于邻域的流形学习方法。

5. 我们给出估计流形局部曲率的方法，并通过引入流形的局部曲率来修正 LTSA 中的极小化模型。这个改进能减少 LTSA 造构造全局嵌入的偏差。结合自适应的邻域选取和曲率修正，我们提出了一种自适应的流形学习方法－自适应局部切空间排列方法（Adaptive local tangent space alignment，简称ALTSA）。虽然曲率模型是针对 LTSA 而设计，但我们相信所提出的基本思想也能适用于其它的流形学习方法。

6. 我们给出了大量的数值例子（模拟例子和实际例子），通过将我们提出的算法与 Isomap、LLE 和 LTSA 的对比和分析，从数值上说明了本文所提出的这些新方法的有效性。

§1.6 本文的组织结构

第1章为引言，第2章主要对目前的几个流形学习方法作一个简单的介绍和比较；第3章针对LLE的缺点提出新的算法；第4章和第5章将针对流形学习的两个主要问题提出解决方案，并给出一个新的算法；第6章是数值实验。最后一章给出了本文的总结，并指出了未来研究的方向。以下是关于各章内容的较为详细的介绍：

第2章介绍流形学习中几个较为常用的算法，并且分析了几种流形学习方法的优缺点以及它们之间的异同点，以期为后继的章节作一个铺垫。

在第3章，对LLE的重构权的性质进行详细的分析，并且在理论上证明在一个给定精度下，每个邻域最小二乘问题可以存在线性无关的多个解。我们用这组线性无关的权向量来表示样本点的局部线性结构，提出了修正的局部线性嵌入方法（MLLE），试图在低维嵌入中保持这种局部线性结构。在这章中，还把这个新算法同LTSA进行了详细的比较，并在理论上证明新算法对等距流形的有效性。

在第4章中，本文针对流形学习所面临的如何选取邻域的共同问题给出了判断邻域是否合适的标准，并在满足这个标准的条件下给出了自适应的邻域选取方法。自适应邻域选取方法包括邻域压缩策略和邻域扩张策略，在匹配流形曲率的前提下尽可能的扩张邻域。

在第5章，本文针对流形学习中曲率和样本点密度变化使得构造局部邻域结构产生偏差的问题，进行了讨论并提出了专门针对LTSA而设计的解决方案。在解决这个问题的过程中，我们还提出了自适应局部切空间排列方法（ALTSA）。

在第6章，本文将给出一些模拟和实际例子以说明前面几章提出的MLLE、邻域选取策略和ALTSA算法的有效性。

在最后一章，我们对全文所提到的新算法进行了简要的总结，指出了其不足之处；同时还展望了今后的工作。

§1.7 本章小结

本章首先给出了数据降维的目的和应用，并介绍了一些线性降维和非线性降维方法。对这些降维方法，我们大致分析了其中的优缺点，并提出了流形学习中所面临的一些共同的问题：1）在加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质；2）应该如何估计流形曲率同数据集的样本点密度的影响，并且减少由此造成的在构造低维嵌入时产生的偏差。

第 2 章　几种流形学习方法

本章要点：

- 介绍等距映射算法（Isomap）
- 介绍局部线性嵌入算法（LLE）
- 介绍拉普拉斯特征映射算法（LE）
- 介绍海赛局部线性嵌入算法（HLLE）
- 介绍局部切空间排列算法（LTSA）
- 简单分析流形学习方法的一些异同点

在这一章中，我们要介绍几种经典的流形学习方法，包括等距流形映射（Isomeitric Mapping，简称为Isomap）、等距线性嵌入（Locally Linear Embedding，简称为LLE）、拉普拉斯特征映射（Laplacian Eigenmap，简称为LE）、海赛局部线性嵌入（Heissian-based Locally Linear，简称为HLLE）和局部切空间排列（Local Tangent Space Alignment，简称为LTSA）等。我们还要指出这些流形学习方法各自的优缺点以及它们的异同点，为后继的章节作一个铺垫和准备。在开始本章内容之前，我们先记高维数据集为 $\{x_i\}_{i=1}^N \in R^m$，降维目标是为了得到 d 维的嵌入结果 $\{\tau_i\}_{i=1}^N \in R^d$。

§2.1　Isomap：等距映射算法

在前面一章中，我们介绍过多维尺度变换（MDS）方法。它是一种线性降维方法，所构造的欧氏距离矩阵并不能反映流形样本点之间的非线性关系。等距映射（Isomap）建立在 MDS 的基础上，力求保持数据点的内在几何性质，即保持两点间的测地距离 [3, 60]，关于 Isomap 的理论分析和推广应用已有许多研究工作 [10, 26, 37, 44, 48]。它同 MDS 的最大区别在于，MDS 构造的距离矩阵反映的样本点之间的欧氏距离，而 Isomap 构造的距离矩阵反映的是样本点之间的测地距离。因此，Isomap 的关键步骤在于如何计算样本点之间的测地距离。在 Isomap 中，测地距离的近似计算方法如下：样本点 x_i 和它的邻域点之间的测地距离用它们之间的欧氏距离来代替；样本点 x_i 和它邻域外的点用流形上它们之间的最短路径来代替。Isomap 的步骤如下：

等距映射算法（Isomap）

1. **选取邻域，构造邻域图** G. 计算每个样本点 x_i 同其余样本点之间的欧氏距离。当 x_j 是 x_i 的最近的 k 个点中的一个时，认为它们是相邻的，即图 G 有边 x_ix_j（这种邻域称为 k-邻域）；或者当 x_i 和 x_j 的欧氏距离 $d(x_i,x_j)$ 小于固定值 ϵ 时，认为图 G 有边 x_ix_j（这种邻域称为 ϵ-邻域）。设边 x_ix_j 的权为 $d(x_i,x_j)$。
2. **计算最短路径**. 当图 G 有边 x_ix_j 时，设最短路径 $d_G(x_i,x_j)=d(x_i,x_j)$；否则设 $d_G(x_i,x_j)=\infty$。对 $l=1,\ldots,N$，
$$d_G(x_i,x_j)=\min\{d_G(x_i,x_j),d_G(x_i,x_l)+d_G(x_k,x_j)\}.$$
这样可以得到最短路径距离矩阵 $D_G=[d_G^2(x_i,x_j)]_{i,j=1}^N$，它由图 G 的所有样本点之间的最短路径的平方组成。
3. **计算d维嵌入**. 将 MDS 应用到距离矩阵 D_G。记
$$H\equiv-(I-\mathbf{1}_N\mathbf{1}_N^T)D_G(I-\mathbf{1}_N\mathbf{1}_N^T)/2,$$
H的最大 d 个特征值 $\lambda_1,\ldots,\lambda_d$ 以及对应的特征向量 $u_1,\ldots,u_d$ 所构成的矩阵为 $U=[u_1,\ldots,u_d]$，那么 $T=\mathrm{diag}(\lambda_1^{1/2},\ldots,\lambda_d^{1/2})U^T$ 是 d 维嵌入嵌入结果。

从上面的算法可以看出，Isomap 是一种全局优化方法，它的嵌入结果能够反映出高维样本点之间的流形距离。因此，如果高维数据所在的低维流形与欧氏空间的一个子集是整体等距的，那么 Isomap 可以得到很理想的嵌入结果；但是需要注意的是，当流形上有“空洞”，即与流形等距的欧氏空间的子集非凸时，计算流形上样本点间的最短路径时会产生较大的偏差，从而导致嵌入结果产生较为明显的变形 [25]。我们以接下来的例子来说明这种现象。

例 2.1：我们首先生成一个 3 维空间中的 2 维流形 S-曲面，数据点生成如下（用 MATLAB 记号）：

$$\begin{aligned}\mathrm{t} &= (\mathrm{rand}(1,\mathrm{N})*3-2)*\mathrm{pi};\\ \mathrm{s} &= \mathrm{rand}(1,\mathrm{N})*5;\\ \mathrm{X} &= [\cos(\mathrm{t});\mathrm{s};(\sin(\mathrm{t})-1).*\mathrm{sign}(\mathrm{pi}/2-\mathrm{t})];\end{aligned}$$

其中$N=1500$。然后，我们再生成一个有“空洞”的S-曲面。在曲面的生成参数 (s,t) 区域上，我们在点 $(2,\pi/2)$ 处挖去一个长为 2 宽为 π 的长方形区域，并生成一个带“空洞”的S-曲面。我们将Isomap分别用到这两个曲面上，其中邻域选取方式均

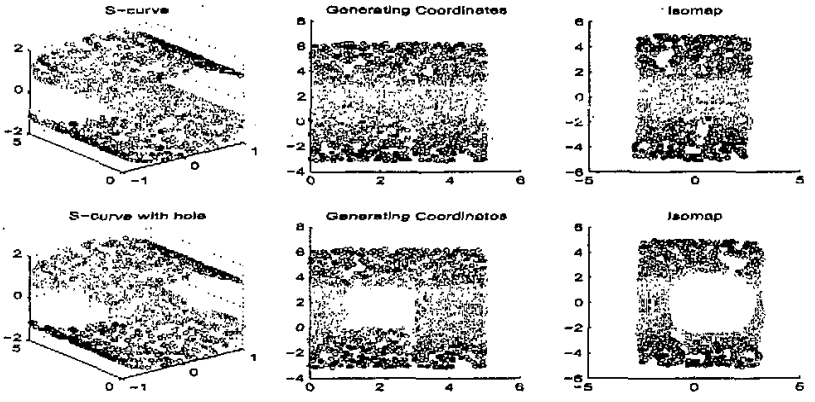

图 2.1: 第一列: S-曲面（第一行）和带“空洞”的S-曲面（第二行）；第二列:生成坐标；第三列: Isomap的嵌入结果。

为$k = 8$的k-邻域。在图 2.1的第一列，我们分别画出了完整的S-曲面（第一行）和有空洞的S-曲面（第二行）；在图 2.1的第二列，分别是它们各自的生成参数 (s,t)。从图 2.1第一行的第三列中我们可以看到，Isomap能很好的恢复出完整的S-曲面的生成坐标。但对于有“空洞”的流形，见图 2.1的第二行第三列，我们可以看到，Isomap的嵌入结果把生成参数平面中的“空洞”放大了。这是因为在计算“空洞”两侧样本点的最短路径时，所计算出的并不是样本点在流形上的测地距离，而是绕过“空洞”的距离，从而导致计算出的最短路径比起真正的最短路径要大，最终使得嵌入结果产生扭曲和变形。

Isomap的另一个问题是所需的计算时间较多，这主要是花在计算样本点之间的最短路径上。由上面给出的 Isomap 算法，我们可以给出 Isomap 的计算复杂度估计：选取邻域的计算复杂度为 $O(mN^2)$，其中 m 为高维样本点的维数；计算最短路径的计算复杂度为 $O(N^3)$；此外，由于距离矩阵的稠密性，计算嵌入结果的计算复杂度 $O(N^3)$。从中我们可以看出， Isomap 所花的计算时间主要是用在计算最短路径和特征值分解上。采用Fibonacci堆的Dijkstra 算法可以将计算最短路径的计算复杂度减小到$O(N^2 logN)$。为了进一步减少 Isomap 的计算时间， V. de Silva 和 J. B. Tenenbaum 提出了 L-Isomap 算法（Isomap with landmark points）[59]，即在样本点中选出 n 个点作为界标点，其中 $n \ll N$。在构造距离矩阵 D_G 时，并不是计算所有样本点之间的测地距离，而是仅仅计算样本点同界标点之间的距离，从而得到一个 $n \times N$ 的距离矩阵 $D_{n,N}$，然后将 L-MDS 用到 $D_{n,N}$ 上来得到嵌入结果。这样，可以将计算最短路径和嵌入结果的计算复杂度减少到 $O(nN^2)$ 和 $O(n^2N)$。 L-Isomap 虽然可以较多的减少 Isomap 所需要的计算

时间，但它同 Isomap 一样，对流形同样有着等距于一个欧氏空间子集以及这个子集是凸的要求。

§2.2 LLE：局部线性嵌入算法

同全局算法 Isomap 相比，局部线性嵌入算法（LLE）是一种局部算法。LLE 在样本点和它的邻域点之间构造一个重构权向量，并在低维空间中保持每个邻域中的权值不变 [56, 57]，目前已有许多对 LLE 的理论分析和推广应用的研究工作 [19, 29, 42, 43, 51, 55]。对于每个样本点 x_i和它的邻域集 $\{x_j, j \in J_i\}$，LLE 需要计算它同邻域点之间的重构权 w_{ji}。重构权 w_{ji} 的选取通过极小化下面的重构误差来实现：

$$\varepsilon(W) = \sum_i \|x_i - \sum_{j \in J_i} w_{ji} x_j\|^2, \tag{2.1}$$

这里权 w_{ji} 表示样本点 x_j 对样本点 x_i 的重构的贡献。为了计算这些权，LLE 对权加入了两个限制：1）若 x_j 不在 x_i 的邻域集中，则 $w_{ji} = 0$；2）对所有的 i，$\sum_j w_{ji} = 1$。这样，这些权组成了一个稀疏矩阵 W，并且矩阵 W 反映了每个样本点 x_i 同它的邻域点之间的局部几何性质。值得注意的是，这些权有着一些重要的性质：在样本点和它的邻域点作平移、旋转和缩放时，重构权保持不变。这种不变性更加表明了通过最小二乘问题（2.1）计算的重构权，能反映出样本点和它邻域点之间本质的性质。

对于样本点 x_i，由于有约束条件 $\sum_{j \in J_i} w_{ji} = 1$，它的重构误差可以写成：

$$\|x_i - \sum_{j \in J_i} w_{ji} x_j\|^2 = \|\sum_{j \in J_i} w_{ji}(x_i - x_j)\|^2. \tag{2.2}$$

记 $G_i = [\ldots, x_i - x_j, \ldots], j \in J_i$以及用 w_i 表示由局部权 $w_{ji}, j \in J_i$ 构成的局部权向量。根据拉格朗日乘子法，（2.2）的最优解满足

$$\begin{cases} G_i^T G_i w_i - \lambda \mathbf{1} = 0, \\ \mathbf{1}^T w_i = 0, \end{cases} \tag{2.3}$$

其中 λ 是拉格朗日乘子。当 G_i 列满秩时，可以通过下述方法来计算重构权：

$$G_i^T G_i y_i = 1, \quad w_i = y_i / \mathbf{1}^T y_i. \tag{2.4}$$

在很多情况下，由于矩阵 $G_i^T G_i$ 是奇异或近似奇异的，方程组（2.4）可能没有解，或者求解这个方程组是不稳定的。LLE 采用正则化的方法，通过求解正则化的线性系统

$$(G_i^T G_i + \gamma \|G_i\|_F^2 I) y_i = 1 \tag{2.5}$$

来获得$w_i = y_i/\mathbf{1}^T y_i$。

LLE 要求低维嵌入 $\tau_i \in R^d$ 与它的邻域点能反映出高维空间中样本点的重构权关系，即极小化下面的价值函数：

$$E(T) = \sum_i \|\tau_i - \sum_j w_{ji}\tau_j\|^2, \tag{2.6}$$

其中 $T = [\tau_1, \ldots, \tau_N]$。为了保证极小化这个价值函数能得到唯一解，LLE 对低维嵌入 T 加上两个约束：1）T 被中心化，即 $T\mathbf{1}_N = 0$；2）T 是标准正交阵，即 $TT^T = I$。注意到价值函数（2.6）又可以写成

$$E(T) = \mathrm{Tr}(T(I-W)^T(I-W)).$$

这样，求解矩阵 $\Phi = (I-W)^T(I-W)$ 的最小 $d+1$ 个特征向量 $u_1, \ldots, u_d$，可以得到低维嵌入 $T = [u_2, \ldots, u_{d+1}]^T$ *。LLE 的算法步骤如下：

局部线性嵌入算法（LLE）

1. **选取邻域**. 计算每个样本点 x_i 的邻域点。记 J_i 为 x_i 的邻域点的下标集，以及 $k = |J_i|$ 表示 x_i 的邻域点的个数。
2. **计算重构权**. 对每个样本点，令 $G_i = [\ldots, x_j - x_i, \ldots]_{j\in J_i}$，求解正则化的线性系统

$$(G_i^T G_i + \gamma\|G_i\|_F^2 I)y_i = \mathbf{1}_k, w_i = y_i/\mathbf{1}_k^T y_i.$$

初使化权矩阵 $W = 0$，再设 $W(J_i, i) = w_i, i = 1, \ldots, N$，得到权矩阵 W。

3. **计算 d 维嵌入**. 计算矩阵 $\Phi = (I-W)^T(I-W)$ 的最小 $d+1$ 个特征向量 $u_1, \ldots, u_d$，则 $T = [u_2, \ldots, u_{d+1}]^T$ 为计算的嵌入结果。

从上面的算法中，可以很容易的估计出 LLE 所需的计算复杂度：选取邻域的计算复杂度为 $O(mN^2)$；计算重构权的计算复杂度为 $O((m+k)k^2N)$；由于矩阵 $(I-W)^T(I-W)$ 很强的稀疏性，计算d维嵌入的计算复杂度只是 $O(dN^2)$ [57]。显然，同 Isomap 相比，LLE 所需要的计算时间要少的多。但是，LLE 也有一些问题。

首先，LLE 的低维嵌入所保持的并不是一个距离关系。因此，对于等距的流形，LLE 并不一定能很好的恢复出同它等距的低维嵌入。如图 2.2所示，我们在图的第二列和第三列画出了对于例 2.1中完整的 S-曲面和有“空洞”的 S-曲面分别采用 LLE 得到的嵌入结果，其中邻域选取策略为 $k = 8$ 的 k-邻域策略。很明显的，对于两个曲面，LLE

*由于$\Phi\mathbf{1}_N = 0$，因此$u_1 = \mathbf{1}_N/\sqrt{N}$，故将$u_1$舍去。

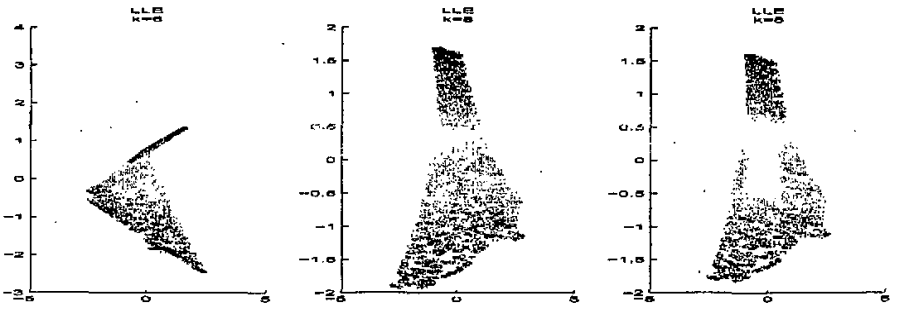

图 2.2: 前两列: 对例 2.1中的 S-曲面采用 LLE 得到的嵌入结果，邻域选取策略分别为 $k=6$（第一列）和 $k=8$（第二列）的 k-邻域策略；第三列:对例 2.1中有“空洞”的 S-曲面采用 LLE 得到的嵌入结果。

得到的嵌入结果同生成坐标相比（见图 2.1的第二列）有着很大的差别，无法恢复出等距的生成坐标。值得注意的是，比较 S-曲面和有“空洞”的 S-曲面的嵌入结果，可以发现在整体的外形上并没有大的差别，这说明 LLE 并没有要求流形所对应的低维空间的子集一定是凸的。

其次，在计算重构权的时候，面临着 γ 的选择问题。对于不同的 γ，最小二乘问题会求出不同的重构权，从而影响最终的嵌入结果。我们会在下一章中对此给出详细的讨论和分析。

最后，由于 LLE 保持邻近点的几何性质，对于有噪音、样本密度稀疏或者相互关联较弱的数据集，相隔较远的点之间的关联会减弱，这样在从高维到低维的映射过程中，很可能会将相隔较远的点映射到邻近点的位置。比如对例 2.1中的 S-曲面，我们设邻域的大小 $k=6$ 以减少相邻邻域之间的交叠和较远的样本点之间的关联。在图 2.1的第一列，我们画出对它采用 LLE 得到的嵌入结果。从图中可以看出，不仅低维嵌入无法恢复出等距的生成坐标，而且出现了将相距远的样本点映射到近的邻域点的现象。对于这个问题，我们针对最小二乘问题可能存在多个解的性质，用多个重构权来代替一个重构权以加强样本点之间的关联，并提出了采用多重权的修正LLE算法，在下一章中我们会给出详细的过程。

§ 2.3 Laplacian Eigenmap：拉普拉斯特阵映射算法

拉普拉斯特征映射（LE）有着很直观的降维目标，即在高维空间中离得很近的点投影到低维空间中的象也应该离得很近 [5]，目前对于 LE 的应用和推广也有一些研究工作 [8, 13, 15, 34, 36]。基于这个出发点，当样本点 x_i 和 x_j 离的很近的时候，LE 用一个正的权 w_{ij} 来联系这两个样本点。通常的，这些权的值被设成 1，即 $w_{ji}=1$；或用指数

衰减函数来设置权，即 $w_{ij} = exp(-\|x_i - x_j\|^2/\sigma^2)$，其中 σ^2 是一个比例参数。用 D 表示对角元素 $D_{ii} = \sum_j w_{ij}$ 的对角矩阵，那么 LE 的低维坐标 τ_i 用来极小化价值函数：

$$E(T) = \sum_{ij} \frac{w_{ij}\|\tau_i - \tau_j\|^2}{\sqrt{D_{ii}D_{jj}}}.$$

同 LLE 相类似，可以通过对 T 加上中心化和标准化限制以得到唯一解，即

$$T1_N^T = 0, \quad TT^T = I.$$

由于价值函数又可以写成

$$E(T) = \mathrm{Tr}(T(I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}})T),$$

因此我们可以通过求解矩阵 $\Phi = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ 的从第 2 小到第 $d+1$ 小的特征向量来得到低维嵌入。LE 的算法总结如下：

拉普拉斯特征值映射（LE）

1. **选取邻域，构造邻域图** G. 计算每个样本点 x_i 同其余样本点之间的欧氏距离。当 x_j 是 x_i 的最近的 k 个点中的一个时，认为它们是相邻的，即图 G 有边 x_ix_j；或者当 x_i 和 x_j 的欧氏距离 $d(x_i, x_j)$ 小于固定值 ϵ 时，认为图 G 有边 x_ix_j。
2. **选择权**. 有两种权的选择方式：
   (1) 当 x_i 和 x_j 相邻时，设 $w_{ij} = e^{-\|x_i - x_j\|^2/\sigma^2}$；否则设 $w_{ij} = 0$；
   (2) 当 x_i 和 x_j 相邻时，设 $w_{ij} = 1$；否则设 $w_{ij} = 0$。
3. **计算** d **维嵌入**. 计算矩阵 $\Phi = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ 的最小 $d+1$ 个特征向量 $u_1, \ldots, u_d$，其中 D 为对角矩阵且 $D_{ii} = \sum_j w_{ij}$，则 $T = [u_2, \ldots, u_d]$ 为计算的嵌入结果。

从给出的算法中可以看出，拉普拉斯特征映射只需要很少的计算量：在第一步和第三步的计算复杂度同 LLE 相同；同 LLE 相比，LE 权的选取方式更为简单直接，无需求解一个线性方程组而是直接设置权值，第二步的计算复杂度最多为 $O(kmN)$（以指数衰减函数设置权值时）。但正是由于权值的简单设置，LE 同 LLE 一样并不适合用来恢复出同流形等距的低维结构。如图 2.3所示，我们对例 2.1中的 S-曲面和有“空洞”的 S-曲面分别采用 LE，嵌入结果完全不能恢复出等距的生成坐标。注意到图 2.3 第一列中出

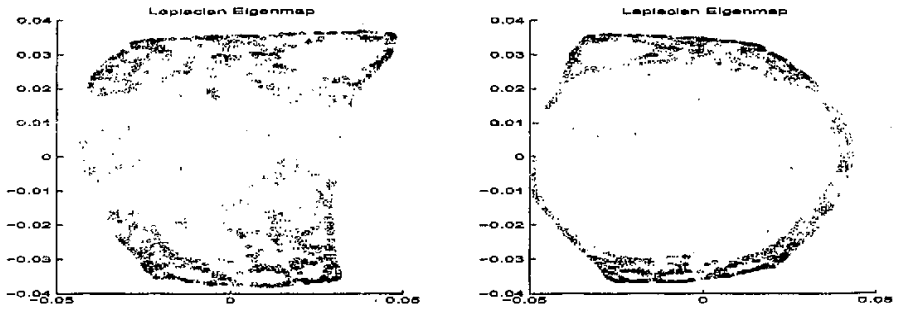

图 2.3: 前两列: 对例 2.1中的 S-曲面采用 LE 得到的嵌入结果（第一列）和对有“空洞”的 S-曲面采用 LE 得到的嵌入结果（第二列），其中邻域选取策略均为$k = 8$的k-邻域策略，权的设置为1。

现一些小的“空洞”而第二列中“空洞”被扩大，这是因为 LE 只是保证流形上邻近的点仍映射到低维空间中邻近的位置，对于流形上距离较远的点，对应低维空间中的距离则可能会放大。LE 直观上的降维目的是为了使邻近的点映射到低维空间上时仍是邻近的，它的一个主要用途是用来处理分类问题。LE 还面临着其它的一些问题，如不同的参数 σ 对嵌入结果有着不同的影响，是否存在最佳的 σ 以及如何选出合适的 σ；作为一种保持局部特征的算法，LE 对于错位点和噪音比较敏感等等，这些问题都值得进一步的研究。

§ 2.4 HLLE：海赛局部线性嵌入算法

海赛局部线性嵌入算法（HLLE）试图恢复出局部等距于低维欧氏空间中开连通子集的流形的生成坐标 [24]。假设 $\mathcal{M} \subset R^m$ 是一个光滑的流形，$T_x(\mathcal{M})$ 是流形中每个点 $x \in \mathcal{M}$ 处的切空间。将 $T_x(\mathcal{M})$ 考虑成 R^m 中的一个仿射子空间，且由于 $x \in \mathcal{M}$ 而有 $0 \in T_x(\mathcal{M})$。记 $\mathcal{N}_x$ 为 x 的邻域，则对每个点 $x' \in \mathcal{N}_x$ 有 $T_x(\mathcal{M})$ 上的唯一的逼近点 $y' \in T_x(\mathcal{M})$ 且映射 $x' \to y'$ 是光滑的。很显然，这些点的坐标可以通过选择 $T_x(\mathcal{M})$ 的正交基来获得。这样，可以得到 x 的邻域 $\mathcal{N}_x$ 的局部坐标，记为 $\theta_1^{(x)}, \ldots, \theta_d^{(x)}$，需要注意的是这些局部坐标依赖于切空间 $T_x(\mathcal{M})$ 中正交基的选取。

接下来，HLLE 用这些局部坐标来定义函数 $f : \mathcal{M} \to R$ 在 x 处的海赛矩阵。用 $g(\theta) = f(x')$ 定义函数 $g : U \to R$，其中 θ 表示 $x' \in \mathcal{N}_x$ 的局部坐标，U 是 R^d 上 0 的邻域。由于映射 $x' \to \theta$ 是光滑的，可以定义 f 在点 x 的海赛矩阵如下：

$$(H_f^{tan}(x))_{ij}) = \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}g(\theta)|_{\theta=0},$$

其中 tan 表示这样的海赛矩阵的建立依赖于切空间 $T_x(\mathcal{M})$。简单说来，在每个点 x 处，HLLE 采用切空间中的坐标和 f 在这样的坐标系统下的导数来建立海赛矩阵。需要

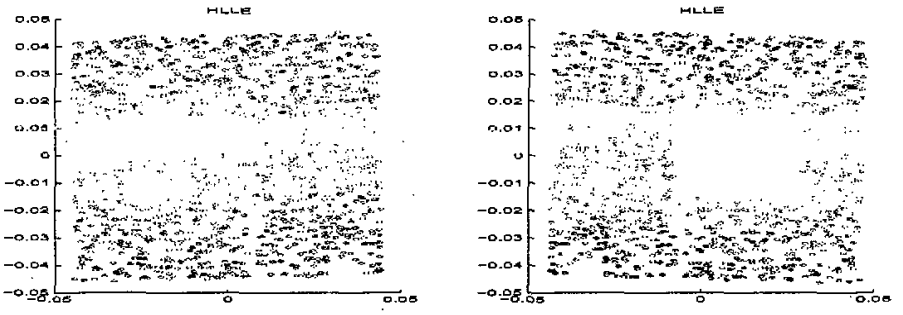

图 2.4: 前两列: 对例 2.1中的S－曲面采用 HLLE 得到的嵌入结果（第一列）和对有“空洞”的S-曲面采用 HLLE 得到的嵌入结果（第二列），其中邻域选取策略均为 $k=8$ 的 k-邻域策略。

特别强调的是，海赛矩阵 H_f^{tan} 依赖于切空间 $T_x(\mathcal{M})$ 中坐标系统的选取。换句话说，如果对 $T_x(\mathcal{M})$ 选用不同的正交基，就可以得到 $\mathcal{N}_x$ 的另一套局部坐标，从而得到不同的海赛矩阵。比较由两套不同的坐标系统生成的海赛矩阵 H 和 H'，如果这两套坐标系统相差一个正交变换 U，那么 H 和 H' 的关系为

$$H' = UHU^T.$$

显然 $\|H'\|_F = \|H\|_F$，其中 $\|\cdot\|_F$ 表示矩阵的 Frobenius 范数。

定义

$$\mathcal{H}(f) = \int_{\mathcal{M}} \|H_f^{tan}(x)\|_F^2 dx,$$

其中 dx 表示 $\mathcal{M}$ 的一个概率测度，这样 $\mathcal{H}(f)$ 就表示了 f 在流形 $\mathcal{M}$ 上的平均弯曲度。在 [24]中，证明了对于局部等距于低维欧氏空间中的开连通子集的流形 $\mathcal{M}$，$\mathcal{H}(f)$ 有 $d+1$ 维的零空间，且这个零空间由一个常函数和等距坐标函数张成的 d 维空间组成。记 $V \subset nullspace(\mathcal{H})$ 为 $\mathcal{H}$ 的零空间中那些正交于常函数的函数组成的子空间，在 V 中可以找出它的一组线性无关的基函数 $\psi_1, \ldots, \psi_d$，并且这组基函数满足：若 x 为 x_0 的邻域点，则有 $\psi_j(x) = \psi_j(x_0) + o(\|x - x_0\|^2)$，$j = 1, \ldots, d$。这样，$\psi(x) = (\psi_j(x))_{j=1}^d$ 为所求的等距低维嵌入。HLLE 的算法步骤如下：

海赛局部线性嵌入算法（HLLE）

1. **选取邻域**. 计算每个样本点 x_i 的邻域点。记 $X_i = [x_{i_1}, \ldots, x_{i_k}]$ 为样本点 x_i 的包括自身在内的最近的 k 个邻域点。
2. **获取切空间坐标**. 对每个样本点的邻域，计算中心化矩阵 $X_i - \bar{x}_i \mathbf{1}_k^T$ 的最大 d 个奇异值对应的右奇异向量，并将这 d 个右奇异向量组成矩阵 V_i。
3. **估计海赛矩阵**. 设

$$M_i = [\mathbf{1},\ V_i,\ (V_i(:,s).*V_i(:,l)_{1\le s\le l\le d})],$$

其中矩阵共有 $1+d+d(d+1)/2$ 列：前 $d+1$ 列由分量为 1 的列向量和 V_i 组成；$V_i(:,s).*V_i(:,l)$ 表示矩阵 V_i 的第 s 列和第 l 列的点积。对矩阵 M_i 进行 Gram-Schmidt 正交化后得到列正交阵 $\tilde{M}_i$，则海赛矩阵 $H^i = \tilde{M}_i(:, d+1 : 1+d+d(d+1)/2)^T$，即为 $\tilde{M}_i$ 的最后 $d(d+1)/2$ 列的转置。

4. **构造二次项**. 利用每个邻域的海赛矩阵 $H^i, i = 1, \ldots, N$ 来构造对称矩阵 H，它的元素为

$$H_{ij} = \sum_{s=1}^{N} \sum_{l=1}^{d(d+1)/2} (H^s)_{l,i}(H^s)_{l,j}.$$

5. **计算H的零空间**. 计算 H 的最小 $d+1$ 个特征值对应的特征向量 $u_1, \ldots, u_d$，则$U = [u_2, \ldots, u_{d+1}]$ 为所求的零空间。
6. **计算嵌入结果**. 记矩阵

$$R_{ij} = \sum_{l\in J_1} U_{l,i}U_{l,j}, \quad i,j = 1, \ldots, d,$$

其中 J_1 表示某个样本点的邻域，则 $T = R^{-1/2}U^T$ 为嵌入结果。

现在我们考虑一下 HLLE 的计算复杂度：选取邻域和计算零空间的计算复杂度同 LLE 相同；获取切空间坐标的计算复杂度为 $O(mk^2N)$；估计海赛矩阵的过程中构造矩阵 M_i 和进行 Gram-Schmidt 正交化的计算复杂度分别约为 $O(d(d+1)kN/2)$ 和 $O((1+d+d(d+1)/2)^2kN)$；构造二次项的复杂度为 $O(d(d+1)kN/2)$；计算 H 的零空间和最终嵌入结果的计算复杂度分别约为 $O(dN^2)$ 和 $O(d^2N)$。从中我们可以看出，当 d 不小的时候，HLLE 估计海赛矩阵的代价会比较大。HLLE 的出发点就是针对局部等距于低维欧氏空间中开连通的子集的流形，同时对于这样的开子集并没有要求是凸的。对于这样的流形，HLLE 可以恢复出与流形等距的低维嵌入。见图 2.4，HLLE 对例 2.1中的 S-曲面和有“空洞”的 S-曲面，都能很好的恢复出流形等距的生成坐标。HLLE 也有一些缺点：HLLE 获取切空间坐标时需要知道流形的本征维数 d；当流形噪音分布不

一致或流形局部低维特征不明显时，获取的切空间坐标可能有较大的偏差，从而影响嵌入结果。此外，由于 HLLE 选取零空间中一组在某个邻域内保持正交性的基作为嵌入结果，对于不同的邻域，嵌入结果有可能不同。实际上，这些嵌入结果同零空间的正交基只相差一个仿射变换，因此算法中的第 6 步没有必要进行，而只是以零空间中的一组正交基作为嵌入结果。

§ 2.5 LTSA：局部切空间排列算法

局部切空间排列算法（LTSA）的基本思想是利用样本点邻域的切空间来表示局部的几何性质，然后将这些局部切空间排列起来构造流形的全局坐标 [67]。给定一个样本点集 $\{x_1,\ldots,x_N\}, x_i \in R^m$，同其它的流形学习方法一样，LTSA 的第一步是寻找每个样本点的邻域，不妨设 $X_i=[x_{i_1},\ldots,x_{i_k}]$ 为样本点 x_i 包括自身在内的最近的 k 个邻域点所构成的矩阵。接下来，LTSA 计算一个 d 维的仿射子空间来逼近 X_i 中的点，即

$$\min_{x,\Theta,Q}\sum_{j=1}^{k}\|x_{i_j}-(x+Q\theta_j)\|_2^2=\min_{x,\Theta,Q}\|X_i-(x\mathbf{1}_k^T+Q\Theta)\|_F^2, \tag{2.7}$$

其中 $\Theta=[\theta_1,\ldots,\theta_k]$ 且 Q 的列数为 d。记 $\bar{x}=X_i\mathbf{1}_k$ 为邻域矩阵 X_i 的中心点，$Q_i\Sigma_iV_i^T$ 为中心化邻域矩阵 $X_i-\bar{x}\mathbf{1}_k^T=[x_{i_1}-\bar{x},\ldots,x_{i_k}-\bar{x}]$ 的奇异值分解，即 Q_i,V_i 分别为对应于最大的 d 个奇异值 $\sigma_1^{(i)},\ldots,\sigma_d^{(i)}$ 的左右奇异向量所构成的矩阵。这样可以很容易的求出（2.7）的最优解为

$$x=\bar{x}_i, Q=Q_i, \Theta=Q_i^T(X_i-\bar{x}\mathbf{1}_k^T),$$

从而可以得到局部坐标系统 $\Theta_i=[\theta_1^{(i)},\ldots,\theta_{k_i}^{(i)}]=[Q_i^T(x_{i_1}-\bar{x}_i),\ldots,Q_i^T(x_{i_k}-\bar{x}_i)]$。LTSA 的第三步是将所有这些有交叠的局部坐标系统 $\Theta_i=[\theta_1^{(i)},\ldots,\theta_k^{(i)}]$ 排列起来以得到一个全局坐标系统 $T=[\tau,\ldots,\tau_N]$。LTSA 认为全局坐标 τ_{i_j} 应该能反映由局部坐标 θ_{i_j} 所决定的局部几何结构，即满足

$$\tau_{i_j}=\bar{\tau}_i+L_i\theta_j^{(i)}+\epsilon_j^{(i)}, j=1,\ldots,k,\quad i=1,\ldots,N, \tag{2.8}$$

其中 τ_i 是 τ_{i_j} 的中心，L_i 是一个待定的局部仿射变换矩阵，而 $\epsilon_j^{(i)}$ 表示局部的重建误差。记 $T_i=[\tau_{i_1},\ldots,\tau_{i_k}]$ 以及 $E_i=[\epsilon_1^{(i)},\ldots,\epsilon_k^{(i)}]$，则有（2.8）的矩阵形式

$$T_i=T_i\mathbf{1}_k\mathbf{1}_k^T/k+L_i\Theta_i+E_i,$$

并且局部重建误差矩阵 E_i 可以写成

$$E_i=T_i(I-\mathbf{1}_k\mathbf{1}_k^T/k)-L_i\Theta_i.$$

为了尽可能保持局部的低维特征，LTSA 极小化下列的重建误差：

$$E(T) = \sum_i \|E_i\|^2 \equiv \sum_i \min_{L_i} \|T_i(I - 1_k 1_k^T/k) - L_i\Theta_i\|^2. \tag{2.9}$$

为了得到唯一解，同 LLE 一样，LTSA 给全局坐标 T 加上中心化和标准化约束。由于重建误差（2.9）又可以写成：

$$E(T) = \sum_i \|T_i(I - 1_k 1_k^T/k)(I - \Theta_i^{+*}\Theta_i)\|^2 = \mathrm{trace}(T\Phi T^T), \tag{2.10}$$

其中 $\Phi = \sum_{i=1}^N S_i W_i W_i^T S_i^T$ 为排列矩阵，$S_i \in R^{N\times k}$ 是满足 $[x_1, \ldots, x_N]S_i = [x_{i_1}, \ldots, x_{i_k}]$ 的选择矩阵，且

$$W_i = I - [1_k/\sqrt{k}, V_i][1_k/\sqrt{k}, V_i]^T. \tag{2.11}$$

这样极小化重建误差（2.9）的最优解能通过计算矩阵 Φ 的从第 2 到第 $d+1$ 小的特征值所对应的特征向量 $u_2, \ldots, u_{d+1}$ 来获得，即 $T = [u_2, \ldots, u_{d+1}]^T$。LTSA 的算法如下：

局部切空间排列算法（LTSA）

1. **选取邻域**. 计算每个样本点 x_i 的邻域。记 $X_i = [x_{i_1}, \ldots, x_{i_k}]$ 为样本点 x_i 的包括自身在内的最近的 k 个邻域点。
2. **局部线性投影**. 对每个样本点的邻域，计算中心化矩阵 $X_i - \bar{x}_i 1_k^T$ 的最大 d 个奇异值对应的右奇异向量，并将这 d 个右奇异向量组成矩阵 V_i。
3. **局部坐标系统的排列**. 构造排列矩阵 $\Phi = \sum_{i=1,\ldots,N} S_i W_i W_i^T S_i^T$，其中 W_i 的构造如（2.11）。计算 Φ 的最小 $d+1$ 个特征值对应的特征向量 $u_1, \ldots, u_d$，则 $T = [u_2, \ldots, u_{d+1}]^T$ 为计算的嵌入结果。

同 HLLE 一样，LTSA 能很好的恢复出流形等距的低维空间的子集，而且对这个低维空间的子集，LTSA 并不要求它是凸的，即 LTSA 对于带有“空洞”的流形，也能很好的恢复出它的低维结构。比如，对于例 2.1 中的完整的 S-曲面和有“空洞”的 S-曲面，我们将 LTSA 用到这两个数据集上，嵌入结果见图 2.5。从图中我们可以看到，无论对于完整的还是有“空洞”的 S-曲面，LTSA 的嵌入结果只是同它们等距的生成坐标（见图 2.1第二列）相差一个仿射变换，这充分说明了 LTSA 对于等距流形的效果。同 HLLE 相比，LTSA 的计算步骤更为简便，无需进行海赛矩阵的估计和构造二次项，因此它所需要的计算时间更少。从 LTSA 的算法步骤中我们不难计算出它的计算复杂度：选取邻域的计算复杂度为 $O(mN^2)$；计算局部坐标系统的计算复杂度为 $O(mk^2N)$；由

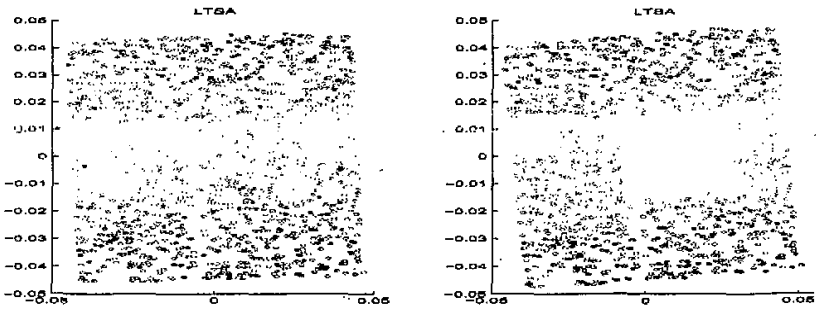

图 2.5: 前两列: 对例 2.1中的S-曲面采用 LTSA 得到的嵌入结果（第一列）和对有“空洞”的S-曲面采用 LTSA 得到的嵌入结果（第二列），其中邻域选取策略均为 $k = 8$ 的 k-邻域策略。

于排列矩阵 Φ 很强的稀疏性，计算 d 维嵌入的计算复杂度为 $O(dN^2)$。此外，LTSA 比 HLLE 在直观上更容易理解，但 LTSA 也面临着同 HLLE 类似的一些问题：由于 LTSA 所反映的局部结构是它的局部 d 维坐标系统，因此由于噪音等因素的影响，数据集的局部低维特征不明显或者不是 d 维的时候，它的局部邻域到局部切空间的投影距离往往并不小。此时，构造的重建误差也不会小，这样 LTSA 可能就无法得到理想的嵌入结果。此外，LTSA 对样本点的密度和曲率的变化的影响比较敏感。样本点的密度和曲率的变化会使得样本点到流形局部切空间的投影产生偏差，而 LTSA 构造排列矩阵的模型并没有将这种偏差计入考虑范围。这使得对于样本点密度和曲率变化较大的流形，LTSA 的嵌入结果可能会出现扭曲现象。在第六章里，我们会详细讨论这种情况，并给出了一种修正的 LTSA 模型，将流形密度和曲率变化的影响计入构造排列矩阵的考虑中。

§2.6 流形学习方法的异同点

从前面的介绍中我们可以看出，流形学习的方法大致上可以分成两类：一类是全局方法（如 Isomap），在降维时将流形上邻近的点映射到低维空间中的邻近点，同时保证将流形上距离远的点映射到低维空间中远距离的点；另一类是局部方法（如LLE、LE、HLLE、LTSA等），这些流形降维方法只是保证将流形上近距离的点映射到低维空间中的邻近点。这两类方法都有各自的缺点：全局方法要求流形所对应的低维空间的子集是凸的，这样才能保证所构造的流形上距离远的点之间的联系是准确的，而且为了找出远距离点之间的联系，需要较多的计算时间；局部方法无需考虑远距离点之间的关系，因此在计算复杂度上要远小于全局方法。此外，由于局部方法只需要考虑流形邻近点之间的关系，因此无需要求流形所对应的低维空间的子集是凸的，有着更广泛的适用对象。但正是由于局部方法只考虑流形上的邻近点，当邻域之间的交叠不够的

时候，流形上远距离点之间的联系较弱。此时会出现将流形上远距离点映射到低维空间中近距离点的现象，而这种现象在邻域关系只是通过单个权来确立的 LLE 和 LE 上显得更加的明显。而且，由于局部方法中流形上较远的点之间的关系不明确，因此全局方法的嵌入结果要比局部方法的嵌入结果有着更直观的意义。

还有其它一些常用的流形学习方法如 Locality Preserving Projections [35]，manifold charting [12]，Local Fisher Embedding [54] 等。同前面介绍过的流形学习方法一样，这些方法都有共同的特征：首先构造流形上样本点的局部邻域结构，然后用这些局部邻域结构来将样本点全局的映射到一个低维空间。它们之间的不同之处主要是在于构造的局部邻域结构不同以及利用这些局部邻域结构来构造全局的低维嵌入方式的不同。比如，Isomap 利用邻域点之间的关联和欧氏距离在数据点上构造一个有权图，然后再利用这个图来估计所有的样本点之间的测地距离，构造的全局低维坐标只是用以保持这个估计的测地距离。LLE 寻找每个样本点同它的邻域点之间的一种线性组合关系，并且使得低维空间中的嵌入坐标之间也保持这种线性组合关系。LTSA 将每个样本的所有邻域点投影到样本点在流形上的局部切空间上，并将所有的局部坐标排列以得到低维的全局坐标。流形学习作为一种非线性降维方法已经在图像处理如人脸图像、手写数字图像以及分类、识别和语音处理等许多方面得到了利用 [30, 37, 40, 49, 66]，它们的共同优势在于：1）它们都是非参数的方法，不需要对流形的很多的参数假设；2）它们是非线性的方法，都基于流形的内在几何结构，更能体现现实中数据的本质；3）它们的求解简单，都转化为求解特征值问题，而不需要用迭代算法，并且避免了局部极值问题。但是，从前面对几种方法的介绍中我们也可以看到它们所面临的一些共同问题：

（1）所有基于邻域的流形学习方法首先面临的都是邻域选取的问题，需要选取出一个合适的邻域以获取局部的线性信息，邻域选取的结果直接影响着最终的嵌入结果。很显然的，邻域越小可以认为邻域的线性结构越明显，但是我们需要注意的是，邻域之间需要有足够的交叠以保证较远的点之间有足够的联系，这又使得邻域不能过小。从直观上想像，流形上曲率大的样本点的邻域应该小一些，而流形上曲率小的样本点处的邻域可以大一些。因此，关于邻域选取我们需要考虑的问题是：在加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质？我们在第四章会详细的讨论这个问题，并提出自适应的邻域选取方法来解决这个问题。

（2）流形上的曲率以及样本点密度的变化，会使得所寻找的局部邻域结构产生偏差。在利用这些局部邻域结构来构造全局的低维嵌入时，需要将这些偏差计入考虑范围。因此，我们需要考虑这些问题：如何估计流形上的曲率？如何估计流形曲率和样本点密度的变化对寻找局部邻域结构的影响？在利用局部邻域结构来构造全局的低维嵌入

时，应该如何计入这种影响以减少低维嵌入的偏差？对于这些问题，我们在第五章会进行详细的讨论，并一一给出回答。我们将以 LTSA 为例，设计一种将曲率和密度计入考虑的模型，这种模型对其它的流形学习方法也有参考作用。最后，结合自适应的邻域选取方法，我们提出了一种自适应的流形学习方法－自适应局部切空间排列方法来解决这两个问题。

§2.7 本章小结

在这一章里，我们介绍了几种经典的流形学习方法，包括等距流形映射（Isomap）、局部线性嵌入（LLE）、拉普拉斯特征映射（LE）、海赛局部线性嵌入（HLLE）和局部切空间排列（LTSA）等。我们指出了这些流形学习方法各自的优缺点以及它们的异同点，并指出了流形学习方法所面临的共同问题：在加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质；应该如何估计流形曲率同数据集的样本点密度的影响，并且减少由此造成的在构造低维嵌入产生的偏差。这些都为后继章节的展开作出了铺垫。

第 3 章　MLLE: 采用多组权的修正 LLE 方法

本章要点：

- LLE 的缺点
- 重构权的性质分析
- 采用多组权的修正 LLE 方法以及分析
- MLLE 和 LTSA 的比较

在前面一章中，我们简单介绍了局部线性嵌入算法（LLE）。LLE 通过每个数据点同其邻域点之间的组合权来决定其局部线性结构，并且在低维的嵌入中保持这种线性结构，而局部权通过求解每个邻域有约束的最小二乘问题而获得。很显然的，LLE 的效果就决定于这种局部重构权的选取。在本章中，我们对 LLE 的局部重构权的性质和特征给出了详细的分析，并且证明了局部有约束最小二乘问题存在一组最优或接近最优的线性无关解。我们用这组线性无关的解作为重构权向量来表示数据点的局部线性结构，并且在低维的嵌入中保持这种线性结构。我们证明了这种采用多重权的修正 LLE 方法（Modified Locally Linear Embedding Using Multiple Weights）对于采自等距流形的样本点可以有理想的结果。我们还将 MLLE 同局部切空间排列方法（LTSA）作了详细的比较，在第六章我们会给出数值实验来说明 MLLE 的效率。

§ 3.1　LLE 的缺点

在前面一章中，我们介绍过局部线性嵌入算法（LLE）。由于 LLE 的参数少、计算快、易求全局最优解等优点，并在图像分类、图像识别、谱重建、数据可视化等方面都有着广泛的应用 [23, 40, 52]。但是，在一些文章中指出，LLE 不够稳定性，尤其当流形维数大于一时，它的嵌入结果可能会有着较明显的扭曲 [53, 58]。这种现象在数据集噪音较大、样本点密度稀疏或者邻域相互关联较弱时更加的明显。这其中的一个原因是 LLE 采用的局部权并不能完全的反映出高维流形的局部几何结构。在前面一章我们提到，对于每个样本点 x_i 和它的邻域集 $\{x_j, j \in J_i\}$，LLE 通过求解最优化问题

$$\min_{\{w_{ji}, j\in J_i\}} \|x_i - \sum_{j\in J_i} w_{ji}x_j\|, \ s.t. \ \sum_{j\in J_i} w_{ji} = 1 \tag{3.1}$$

来构造 x_i 同它的邻域点之间的局部线性关系。为了记号上的方便，我们用 w_i 表示由局部权 $w_{ji}, j \in J_i$ 构成的局部权向量，用 $\mathbf{1}_{k_i}$ 表示所有分量都是 1 的 k_i 维列向量，其中

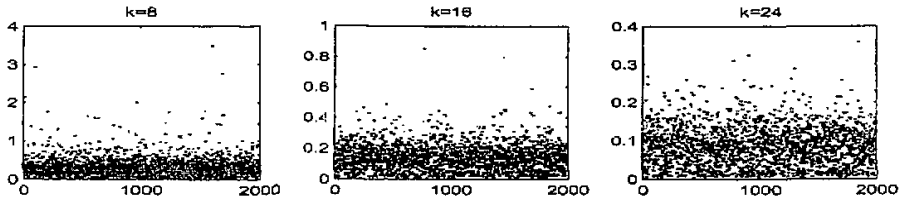

图 3.1: 用 $\gamma = 1.0e-5$ 和 $\gamma = 1.0e-3$ 计算的两组权向量 $w_i^{(1)}$ 和 $w_i^{(2)}$ 之间的距离。

$k_i = |J_i|$ 是邻域点的个数。记 $G_i = [\ldots, x_j - x_i, \ldots]_{j\in J_i}$，由 $\sum_{j\in J_i} w_{ji} = 1$ 可以得到

$$\|\sum_{j\in J_i} w_{ji}(x_i - x_j)\| = \|G_i w_i\|.$$

这样当 G_i 的零空间不正交于 $\mathbf{1}_{k_i}$ 时，w_i 可以通过单位化 G_i 的零空间向量来求得。否则，$w_i = y_i/\mathbf{1}_{k_i}^T y_i$，其中 y_i 是线性系统

$$G_i^T G_i y_i = \mathbf{1}_{k_i} \tag{3.2}$$

的解 [57]。很显然的，当 $G_i^T G_i$ 是奇异或者接近奇异的时候，计算这样的局部权是不稳定的。LLE 往这个线性系统中加入一个小的正数 γ，而通过求解这个正则化的线性系统

$$(G_i^T G_i + \gamma \|G_i\|_F^2 I) y_i = \mathbf{1}_{k_i},\ w_i = y_i/\mathbf{1}_{k_i}^T y_i \tag{3.3}$$

来获得局部权。然而，将 y_i 单位化后的向量 $w_i(\gamma) = w_i$ 很大程度上依赖于 γ 的选取。从而当 γ 有着微小的变化时，最终的嵌入结果可能会有着很大的变化。我们给出一个例子来说明重构权和嵌入结果对 γ 的依赖性。

例 3.1： 我们用 3 维空间的 2 维 swiss-roll 作为例子。数据点生成如下（用 MATLAB 记号）：

$$\begin{aligned} \mathrm{t} &= (3 * \mathrm{pi}/2) * (1 + 2 * \mathrm{rand}(1, \mathrm{N})); \\ \mathrm{s} &= 21 * \mathrm{rand}(1, \mathrm{N}); \\ \mathrm{X} &= [\mathrm{t}. * \cos(\mathrm{t}); \mathrm{s}; \mathrm{t}. * \sin(\mathrm{t})]; \end{aligned}$$

其中 $N = 2000$。在每个点 x_i，我们分别用 $\gamma = 10^{-5}$ 和 $\gamma = 10^{-3}$ 计算两组权向量 $w_i^{(1)}$ 和 $w_i^{(2)}$。在图 3.1，我们画出了 $\|w_i^{(1)} - w_i^{(2)}\|, i = 1, \ldots, N$ 的大小，其中邻域大小 k 分别为8、16、24。从中我们可以看出，采用不同的 γ，在许多的点上权的变化显著。从图 3.2 中我们也可以看到，LLE 的嵌入结果对于不同的 γ 也有着很大的不同。

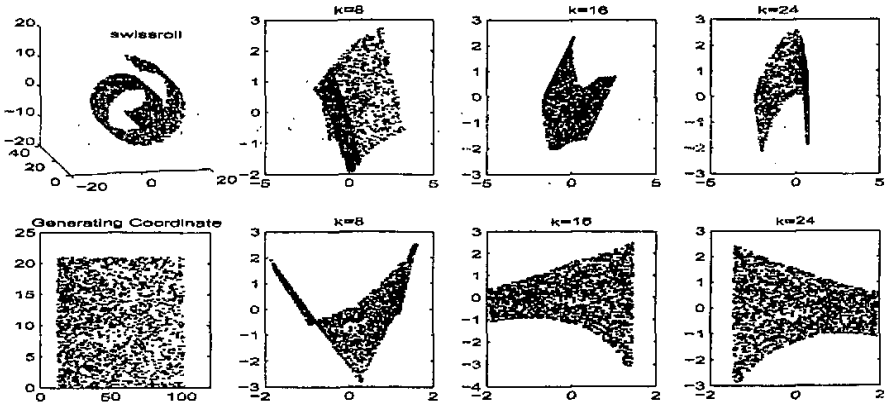

图 3.2: 左列： swiss-roll 的样本点和它的生成坐标。右三列：对于不同的邻域大小 $k = 8, 16, 24$， LLE 分别用 $\gamma = 10^{-5}$（顶行）和 $\gamma = 10^{-3}$（底行）得到的嵌入结果。

在下一节中，我们要给出一些分析来证明 $w_i(\gamma)$ 的收敛性。大致说来，当 γ 并不是充分小的时候，$w_i(\gamma)$ 也许并不能逼近最终的 w_i^* 到一个可以接受的程度。也就是说，很难选择一个合适的 γ 使得 $w_i(\gamma)$ 能很近似的逼近 w_i^*。而另一方面，如果有多个最优权向量，比如当 G_i 的零空间的维数大于 1 时，我们并不知道选择哪个权能最好的得到最终的嵌入结果。我们用下面一个两维的数据集作为例子来说明这种情况。

例 3.2： 我们随机生成一个由 $N = 20$ 个点组成的数据集，并用“$\circ$”点标在图 3.3中。我们设置邻域的大小 $k = 4$。由于每个 G_i 是一个 2×4 的矩阵，G_i 有至少两个线性无关的零空间向量。我们采用如下的方式构造权向量。首先，我们计算出 G_i 的零空间的两个正交基向量；然后，我们通过将它们单位化来得到权向量 $w_i^{(1)}$ 和 $w_i^{(2)}$。我们还通过采用 $\gamma = 10^{-4}$ 的正则化系统（3.3）计算出权 $w_i^{(3)}$。LLE 采用这三组不同的权向量 $\{w_i^{(j)}\}$ 来分别得到嵌入结果 $T^{(j)}, j = 1, 2, 3$。我们对 $T^{(j)}$ 作仿射变换后得到 $Y^{(j)} = c^{(j)}1^T + L^{(j)}T^{(j)}$，其中 $Y^{(j)}$ 满足

$$\|X - Y^{(j)}\| = \min_{c,L} \|X - (c1^T + LT^{(j)})\|.$$

很显然的，如果嵌入结果 T 能够恢复样本点集，那么 T 应该同样本点集 X 之间只相差一个仿射变换，即 $\min_{c,L} \|X - (c1^T + LT)\| = 0$。可是对于这三组权，LLE 都不能恢复样本点集；同样的，误差 $\|X - Y^{(j)}\|, j = 1, 2, 3$ 并不小。在图 3.3的左边三列中，我们分别画出$Y^{(1)}$、$Y^{(2)}$和$Y^{(3)}$（用“.”表示）。很明显，没有 $Y^{(j)}, j = 1, 2, 3$ 可以恢复出点集 X。

对于一个一维流形，LLE 总是能得到很理想的结果。但对于一个更高维的流形，仅仅采用单个权来构造局部线性结构是不够的。这是因为对于更高维的流形，单个权不足

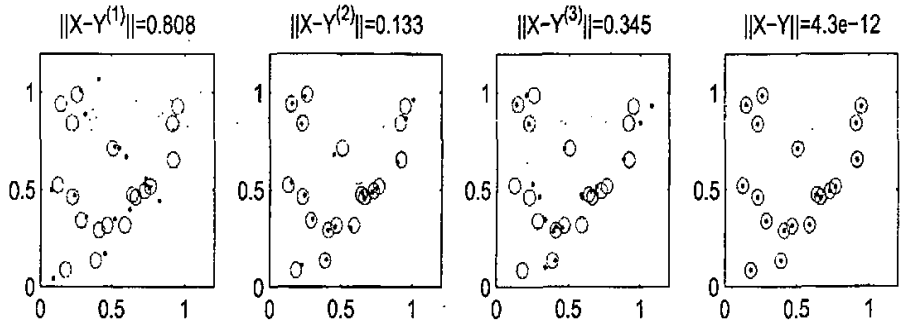

图 3.3: LLE 对于一个两维的样本点集（o-点）的结果。左边三列画出了对 LLE 采用不同的零空间向量作为权（左边两列）或正则化的权（第三列）计算出的低维嵌入作最优的仿射变换后的结果（用 "." 表示）。最后一列画出的是采用两组单位化的零空间向量作为权后 LLE 得到的结果。

以反映出流形上更为复杂的局部几何结构，而且采用单个权建立的样本点之间的关联也比较弱。如果采用多组（线性无关）的权来构造局部线性结构，应该能改善最终的嵌入结果。在图 3.3 中，我们画出了采用两组单位化零空间向量作为权后 LLE 得到的结果。在下一节中，我们将给权向量的性质作一些分析，并且证明一组线性无关的最优或近似最优的权向量的存在性。

§3.2 权向量的性质

在 [57] 中曾指出，当 G_i 是奇异的时候，权向量也许并不是唯一的。在这节中，我们将对权向量的特征作出进一步的分析，包括正则化的权向量 $w_i(\gamma)$ 的收敛性和近似最佳的线性无关权向量的存在性。在这里，出于符号上的便利，我们忽略 G_i 的下标 i 而考虑关于 G 的最小问题

$$\min_{1^T w=1} \|Gw\|. \tag{3.4}$$

在接下来的讨论中，我们会重复用到 G 的奇异值分解 (SVD)

$$G = [U, U_\perp]\,\mathrm{diag}(\Sigma, 0)[V, V_\perp]^T, \tag{3.5}$$

其中 $[U, U_\perp]$ 和 $[V, V_\perp]$ 是正交矩阵，而 $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_s)$ 是由 G 的非零奇异值组成的对角阵（用降序排列）。显然，$V_\perp$ 是 G 的零空间的一组正交基构成的矩阵而 V 是由 G^T 的列空间的一组正交基构成的矩阵。

定理 3.1： 记 $G \in R^{m\times k}$，y_0 为 $\mathbf{1}_k$ 到 G 的零空间上的正交投影，$y_1 = (G^TG)^+\mathbf{1}_k$*。则如下定义的向量 w^* 是 (3.4) 的最优解，

$$w^* = \frac{y^*}{\mathbf{1}_k^T y^*}, \quad y^* = \begin{cases} y_0, & y_0 \neq 0 \\ y_1, & y_0 = 0 \end{cases}. \tag{3.6}$$

证明： 通过 y_0 的定义和 G 的奇异值分解，我们有

$$y_0 = V_\perp V_\perp^T \mathbf{1}_k, \quad y_1 = V\Sigma^{-2}V^T\mathbf{1}_k. \tag{3.7}$$

如果 $y_0 \neq 0$，那么 $\mathbf{1}_k^T y_0 = \|V_\perp^T\mathbf{1}_k\|^2 \neq 0$ 并且 $w^* = y_0/\mathbf{1}_k^T y_0$ 有意义。显然，$Gw^* = 0$。

如果 $y_0 = 0$，那么 $V_\perp^T\mathbf{1}_k = 0$ 并且 $\mathbf{1}_k$ 属于 G^T 的列空间，这样就有 $VV^T\mathbf{1}_k = \mathbf{1}_k$。通过拉格朗日乘子法可以得知，(3.4) 的最优解 w 必定满足 $G^TGw = \lambda\mathbf{1}_k$，这里 λ 是一个缩放因子。将奇异值分解（3.5）代入这个等式后可以得到 $V^Tw = \lambda\Sigma^{-2}V^T\mathbf{1}_k$。这样，$VV^Tw = \lambda(G^TG)^+\mathbf{1}_k = \lambda y_1$，并且我们可以重新把 w 写成

$$w = (I - VV^T)w + VV^Tw = (I - VV^T)w + \lambda y_1.$$

这样就有

$$1 = \mathbf{1}_k^T w = \mathbf{1}_k^T(V_\perp V_\perp^T)w + \lambda\mathbf{1}_k^T y_1 = \lambda\mathbf{1}_k^T y_1$$

和 $\lambda = 1/\mathbf{1}_k^T y_1$。记 $w^* = \lambda y_1 = y_1/\mathbf{1}_k^T y_1$，从而有 $w = (I - VV^T)w + w^*$。通过 $G(I - VV^T) = 0$，我们就得到了 $\|Gw^*\| = \|Gw\|$，可知 w^* 是最优解。 ■

定理 3.2： 记 $y(\gamma)$ 为如下正则化线性系统的唯一解，

$$(G^TG + \gamma\|G\|^2 I)y = \mathbf{1}_k \tag{3.8}$$

其中 $\gamma > 0$。记 $w(\gamma) = y(\gamma)/\mathbf{1}_k^T y(\gamma)$，则有 $\lim_{\gamma\to 0} w(\gamma) = w^*$，其中 w^* 如（3.6）所定义。

证明： 显然，$w(\gamma)$ 能表示成

$$w(\gamma) = \frac{(G^TG + \gamma\|G\|^2 I)^{-1}\mathbf{1}_k}{\mathbf{1}_k^T(G^TG + \gamma\|G\|^2)^{-1}\mathbf{1}_k}.$$

通过 (3.5) 和 (3.7)，有

$$w(\gamma) = \frac{V(\Sigma^2 + \gamma\|G\|^2 I)^{-1}V^T\mathbf{1}_k + \gamma^{-1}\|G\|^{-2}y_0}{\mathbf{1}_k^T V(\Sigma^2 + \gamma\|G\|^2 I)^{-1}V^T\mathbf{1}_k + \gamma^{-1}\|G\|^{-2}\mathbf{1}_k^T y_0}. \tag{3.9}$$

* $(\cdot)^+$ 记为矩阵的 Moore-Penrose 广义逆。

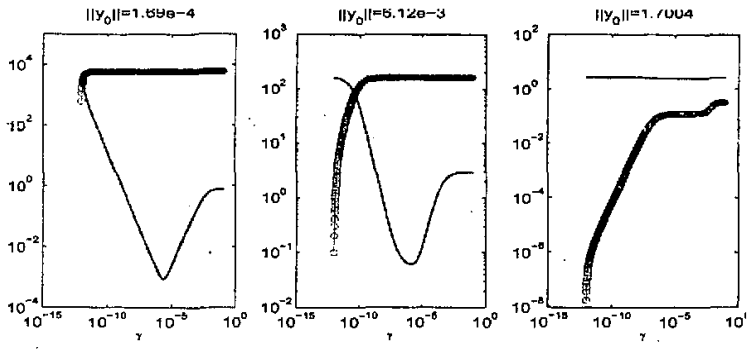

图 3.4: swiss-roll 数据的三个不同样本点的邻域的误差曲线$\|w(\gamma)-\frac{y_0}{1^Ty_0}\|$（用“o”表示）和$\|w(\gamma)-\frac{y_1}{1^Ty_1}\|$（用“.”表示）。

如果$y_0 \neq 0$，则当 $\gamma \to 0$ 有

$$w(\gamma)=\frac{\gamma\|G\|^2V(\Sigma^2+\gamma\|G\|^2I)^{-1}V^T1_k+y_0}{\gamma\|G\|^2 1_k^TV(\Sigma^2+\gamma\|G\|^2I)^{-1}V^T1_k+1_k^Ty_0}\longrightarrow\frac{y_0}{1_k^Ty_0}=w^*.$$

否则，

$$w(\gamma)=\frac{V(\Sigma^2+\gamma\|G\|^2I)^{-1}V^T1_k}{1_k^TV(\Sigma^2+\gamma\|G\|^2I)^{-1}V^T1_k}\longrightarrow\frac{V\Sigma^{-2}V^T1_k}{1_k^TV\Sigma^{-2}V^T1_k}=\frac{y_1}{1_k^Ty_1}=w^*.$$

■

注意到 y_0 和 y_1 是相互正交的。这种正交性意味着对于一个小的正数 γ，在构造权 $w(\gamma)$ 时会有着某种不确定性。我们还需注意到 y_0 同时也正交于（3.9）中分子的第一项。如果 $y_0 \neq 0$ 并且

$$\|V(\Sigma^2+\gamma\|G\|^2I)^{-1}V^T1_k\|>\gamma^{-1}\|G\|^{-2}\|y_0\|, \tag{3.10}$$

那么（3.9）分子的第一项是占优的，从而 $w(\gamma)$ 并不能很好的逼近于极限向量 w^*。通常情况下，当 $y_0 \neq 0$ 且 γ 不是很小的时候，不等式（3.10）成立。所以当 $\gamma \to 0$ 时，$w(\gamma)$ 首先趋向于 $\frac{y_1}{1^Ty_1}$，然后转而趋向于极限值 $w^*=\frac{y_0}{1^Ty_0}$。因此，如果 γ 并不是充分小，$w(\gamma)$ 并不能很精确的逼近于极限向量 w^*。需要注意的是，当 $\|y_0\|$ 比较大的时候，不等式（3.10）不一定成立，这种现象也许不会出现。在数值计算中，我们将 y_0 表示成 1_k 到 G 的对应于最小奇异值的特征值空间上的正交投影。我们用下面的例子来总结上面的分析。

例3.1（续）：我们选择例 3.1中的三个不同样本点 x_i 的邻域为例。在图 3.4中，我们分别画出了 $\|w(\gamma)-\frac{y_0}{1^Ty_0}\|$（用“o”表示）和$\|w(\gamma)-\frac{y_1}{1^Ty_1}\|$（用“.”表示）对于这三个

邻域集的不同的误差曲线。在图3.4的左边两幅中，$\|y_0\|$ 很小，此时 $w(\gamma)$ 先趋向于 $\frac{y_1}{1^T y_1}$ 再趋向于极限向量 $w^* = \frac{y_0}{1^T y_0}$。然而，当 $\gamma > 10^{-10}$ 时，$w(\gamma)$ 也不能逼近 w^*到一个很好的精度。而当 $\|y_0\|$ 不小的时候，从图3.4的第三幅中，我们可以看到这种现象并不出现。

当 x_i 和它的邻域点处于或近似在一个 d 维的超平面时，G_i 一定有着小的奇异值 $\sigma_{d+1}(G_i)$ 并且 $\sigma_d(G_i)$ 相对来说并不小。在这种情况下，当 $\gamma \to 0$ 时 $w(\gamma)$ 会出现如上讨论不定性。这也可以用来部分的说明为什么用正则化方法来近似求解重构权向量时，LLE可能会得到不理想的嵌入结果。

即使是这些重构权能在计算时能有很好的精度，由单一权（3.1）所决定的线性结构是很弱的，这在流形的本征维数较高时尤为明显。因此必须在 x_i 和它的邻域点之间建立一种更强的线性结构用以反映高维流形复杂的局部几何结构。出于这个目的，一种简单的方法是采用多个权 $w_i^{(j)}$ 来构造这种线性结构。这里 $w_i^{(j)}$ 不一定要正好是（3.1）的最优解，但每个 $\|G_i w_i^{(j)}\|$ 应该能逼近最小值*。对于有噪音的数据，这些权向量比起单一的最优权也似乎更加的合理。下面的定理证明了当 G_i 有小的奇异值向量时，一组线性无关且近似最优的权向量是存在的。

定理3.3: 若 $G \in R^{m\times k}$，$\sigma_1(G) \geq \cdots \geq \sigma_k(G)$ 是 G 的 k 个奇异值，那么对于 $r < k$，存在 $k-r$ 个线性无关向量 $w^{(j)}, j = 1, \ldots, k-r$ 使得

$$\|Gw^{(j)}\| \leq \min_{1_k^T w = 1} \|Gw\| + \sigma_{r+1}(G), \quad 1_k^T w^{(j)} = 1, \tag{3.11}$$

并且对于 $W_* = [w^{(1)}, \ldots w^{(k-r)}]$，有

$$\|GW_*\|_F \leq \sqrt{k-r} \min_{1_k^T w = 1} \|Gw\| + \sqrt{\sum_{j=r+1}^{k} \sigma_j^2(G)}. \tag{3.12}$$

证明: 我们通过奇异值分解将 G 分成 $G = [U_1, U_2]\,\mathrm{diag}(\Sigma_1, \Sigma_2)[V_1, V_2]^T$，这里 Σ_1 是由 G 的 r 个最大的奇异值组成的对角阵，$\|\Sigma_2\|_2 = \sigma_{r+1}(G)$ 并且 $\|\Sigma_2\|_F = \sqrt{\sum_{j=r+1}^{k} \sigma_j^2(G)}$。我们能构造一个 Householder 矩阵 $H = I - 2hh^T \in \mathcal{R}^{(k-r)\times(k-r)}$ 将向量 $V_2^T 1_k$ 变换到平行于 1_{k-r} [28]，即有

$$HV_2^T 1_k = \alpha 1_{k-r}, \quad \alpha = \frac{1}{\sqrt{k-r}} \|V_2^T 1_k\|.$$

*如果 x_i 和它的 k_i 个邻域点在一个 m 维欧氏空间的一个 d 维超平面上，则 G_i 的列秩为 d，并且 G_i 的零空间的一组正交基向量是 $k_i - d$ 个线性无关的最优权。

向量 h 能用如下的方式轻易的构造。记 $h_0 = \alpha \mathbf{1}_{k-r} - V_2^T \mathbf{1}_k$。如果 $h_0 \neq 0$，则 $h = h_0/\|h_0\|$；否则，$h = 0$。矩阵 H 显然是正交的。我们设

$$w^{(j)} = (1-\alpha)w^* + V_2 H(:,j), \quad j = 1, \ldots, k-r,$$

其中 w^* 是（3.4）的最优解。显然有 $\mathbf{1}_k^T w^{(j)} = 1$。注意到 $GV_2 = U_2\Sigma_2$ 和 $\alpha \leq 1$，我们有

$$\|Gw^{(j)}\| \leq (1-\alpha)\|Gw^*\| + \|U_2\Sigma_2 H(:,j)\| \leq \min_{\mathbf{1}_k^T w = 1} \|Gw\| + \sigma_{r+1}(G).$$

显然的，有 $W_* = [w^{(1)}, \ldots w^{(k-r)}] = (1-\alpha)w^* \mathbf{1}_{k-r}^T + V_2 H$ 和 $\mathbf{1}_k^T W_* = \mathbf{1}_{k-r}^T$。这就证明了

$$\begin{aligned} \|GW_*\|_F &= \|(1-\alpha)Gw^* \mathbf{1}_{k_i-r}^T + GV_2 H\|_F \\ &\leq (1-\alpha)\sqrt{k_i - r}\|Gw^*\| + \|U_2\Sigma_2 H\|_F \\ &\leq \sqrt{k_i - r} \min_{\mathbf{1}_k^T w = 1} \|Gw\| + \sqrt{\sum_{j=r+1}^{k_i} \sigma_j^2(G)}. \end{aligned}$$

我们现在证明 $w^{(1)}, \ldots, w^{(k-r)}$ 的线性无关性。假设有向量 $y \in R^{k-r}$ 使得 $W_* y = 0$，那么由 $\mathbf{1}_k^T W_* = \mathbf{1}_{k-r}^T$ 可得 $\mathbf{1}_{k-r}^T y = (\mathbf{1}_k^T W_*)y = 0$。这样我们就有

$$V_2 H y = W_* y - (1-\alpha)w^* \mathbf{1}_{k-r}^T y = 0.$$

由于 $V_2 H$ 是正交的，这样就有 $y = 0$。所以 $w^{(1)}, \ldots, w^{(k-r)}$ 是线性无关的。 ■

通常来说，如果 $\{x_i\}$ 采自于一个 d 维流形（也许有噪音）并且 $\{x_i\}$ 的邻域集相对的比较小，那么 $\{x_i\}$ 和它的邻域点近似的处在一个 d 维的超平面上。这样，$\sigma_{d+1}(G_i)$ 通常也比较小。

下面的推论是显然的。

推论 3.1: 记 s 为 G 的零空间的维数。那么有 s 个线性无关的最优权。

证明: 记 $r = k - s$。由 $\sigma_{r+1}(G) = 0$ 和定理 3.3可直接得出结论。

备注. 当 $\|w^*\|$ 不大的时候，权矩阵 W_* 是有意义的。实际上，对于任何 $\|y\| = 1$ 的向量

$$\|W_* y\| = \|(1-\alpha)w^* \mathbf{1}_{k-r}^T y + V_2 H y\| \geq 1 - (1-\alpha)\|w^*\|\|\mathbf{1}_{k-r}^T y\|.$$

注意到 $\|\mathbf{1}_{k-r}^T y\| = \|\mathbf{1}_k^T W_* y\| \le \sqrt{k}\|W_* y\|$。我们有 $\|W_* y\| \ge 1 - \sqrt{k}(1-\alpha)\|w^*\|\|W_* y\|$，从而有

$$\|W_* y\| \ge \frac{1}{1+\sqrt{k}(1-\alpha)\|w^*\|}.$$

这意味着 $\sigma_{k-r}(W_*) \ge 1/(1+\sqrt{k}(1-\alpha)\|w^*\|)$。另一方面，

$$\|W_*\| \le 1 + \sqrt{k-r}(1-\alpha)\|w^*\|.$$

我们可以得到 W_* 的条件数 $\mathrm{cond}(W_*) = \|W_*\|/\sigma_{k_r}(W_*)$ 的一个界

$$\mathrm{cond}(W_*) \le (1+\sqrt{k}(1-\alpha)\|w^*\|)^2.$$

这个界通常并不大。

§3.3 MLLE :修正的局部线性嵌入方法

§3.3.1 权的计算

很显然的，用多组线性无关权向量构造的线性结构比起用单一的权向量构造的线性结构有着更好的稳定性，而且更能反映高维流形复杂的局部几何结构。这使得通过对每个邻域采用多组权来改善 LLE 成为可能。在这一节中，我们会给出采用多组重构权的修正局部线性嵌入方法。下一节中，我们会给出一些分析来说明这种想法的可行性。

考虑 x_i 的邻域集。将定理 3.3中的 G 改成 G_i，我们可以得到 $k_i - r_i$ 个线性无关的权向量 $w_i^{(1)}, \ldots, w_i^{(k_i-r_i)}$，即

$$w_i^{(\ell)} = (1-\alpha_i)w_i^* + V_2^{(i)} H_i(:,\ell), \quad \ell \le k_i - r_i,$$

其中 w_i^* 是（3.6）中所定义的最优解，$V_2^{(i)}$ 是对应于 G_i 的 $k_i - r_i$ 个最小奇异值的右奇异向量，$\alpha_i = \frac{1}{\sqrt{k_i - r_i}}\|(V_2^{(i)})^T \mathbf{1}_{k_i}\|$，$H_i = I - 2h_i h_i^T$，且 h_i 如下构造

$$h_{i0} = \alpha_i \mathbf{1}_{k_i - r_i} - (V_2^{(i)})^T \mathbf{1}_{k_i}, \quad h_i = \begin{cases} \frac{h_{i0}}{\|h_{i0}\|} & h_{i0} \ne 0, \\ 0 & h_{i0} = 0. \end{cases}$$

我们寻找一个 d 维嵌入 $\{\tau_1, \ldots, \tau_N\}$，其中 $\tau_i \in \mathcal{R}^d$ 能保持 x_i 和它的邻域点之间更强的线性结构，即极小化下列的嵌入价值函数

$$E(T) = \sum_{i=1}^{N} \sum_{\ell=1}^{k_i - r_i} \| \sum_{j \in J_i} w_{j,i}^{(\ell)} \tau_j - \tau_i \|^2.$$

记 $W_i = [w_i^{(1)}, \ldots, w_i^{(k_i - r_i)}]$ 为局部的权矩阵，并且将其嵌入到 N 维空间。记为 $\hat{W}_i \in \mathcal{R}^{N \times (k_i - r_i)}$，并有

$$\hat{W}_i(J_i, :) = W_i,\ \hat{W}(i, :) = -\mathbf{1}_{k_i - r_i}^T,\ \hat{W}(j, :) = 0,\ j \notin I_i,$$

其中 $I_i = J_i \cup \{i\}$。这样我们可以将嵌入价值函数重新写成

$$E(T) = \sum_i \|T\hat{W}_i\|_F^2 = \mathrm{Tr}(T\Phi T^T), \tag{3.13}$$

其中

$$\Phi = \sum_i \hat{W}_i \hat{W}_i^T = [\hat{W}_1, \ldots, \hat{W}_N][\hat{W}_1, \ldots, \hat{W}_N]^T. \tag{3.14}$$

$E(T)$ 的极小解可以也就是对应于 Φ 的从第 2 到第 $d+1$ 个最小特征值的特征向量所构成的矩阵 $T = [u_2, \ldots, u_{d+1}]^T$。

现在我们考虑将如何选取 r_i。通常说来，如果样本点采自一个 d 维的流形，r_i 可以是满足 $d \le r_i < k_i$ 的任何一个整数。而之所以让 $r_i \ge d$ 是因为如果 $r_i < d$，则 G_i 的奇异值 $\sigma_{r_i+1}(G_i)$ 也许并不小，这样得到权并不是最佳或近似最佳的。在理想的情况下，最佳的 r_i 似乎应该是 d，因为增加线性无关的权的个数可以增强线性结构。可是，由于有着噪音的影响或者当邻域点选的并不好的时候，低维的局部性质也许并不能很好的由其邻域点所表示出来。所以，我们采用下面的方法来选取 r_i，

$$r_i = \min_{\ell} \left\{ \ell \ge d,\quad \frac{\sum_{j=\ell+1}^{k_i} \lambda_j(G_i^T G_i)}{\sum_{j=1}^{\ell} \lambda_j(G_i^T G_i)} < \eta \right\}, \tag{3.15}$$

其中 $\lambda_j(G_i^T G_i) = \sigma_j^2(G_i)$ 是 $G_i^T G_i$ 的特征值而 $\eta < 1$。

接下来，我们要考虑如何选取合适的 η。很显然，每个邻域的权的个数受 η 的控制：当 η 较大的时候，有 $r_i = d$，每个邻域采用 $k_i - d$ 个权构造它的局部线性结构；当 η 很小的时候，有 $r_i = k_i - 1$，每个邻域只采用单个权构造它的局部线性结构。因此，我们可以用 η 来控制流形的曲率和噪音对低维嵌入结果的影响：当流形的局部曲率和噪音比较大的时候，我们减少表示局部线性结构的重构权的个数，这样可以减小局部的重构误差；当流形的局部曲率和噪音小的时候，我们用较多的重构权表示局部线性结构，在保持局部重构误差较小的同时尽可能采用多的重构权来增强线性结构的稳定性。记

$$\rho_i = \frac{\sum_{j=d+1}^{k_i} \lambda_j(G_i^T G_i)}{\sum_{j=1}^{d} \lambda_j(G_i^T G_i)},\quad i = 1, \ldots, N$$

且 $\rho_1 \leq \cdots \leq \rho_N$ 为升序排列。很显然，无论流形的曲率以及噪音的大小是多少，当我们认为一个流形是 d 维的，它的多数样本点的局部应该是 d 维或近似 d 维的，即多数的 ρ_i 都比较小。因此我们可以选取 $\eta = \rho_{\lceil N/2 \rceil}$，其中 $\lceil N/2 \rceil$ 表示对 $N/2$ 向上取整。这种 η 的选取方式表明：对于流形上 ρ_i 小的半数样本点，我们将用 $k_i - d$ 组权表示局部线性结构以增强邻域点之间的关联；对于其余的样本点，我们适当的减少局部邻域重构权的个数以保证局部邻域的重构误差比较小。在第六章中，我们会给出数值实验来说明这种选取方式是合适的，并且能处理曲率高度变化的流形。

§3.3.2 构造Φ的计算要点

通过设置初使 $\Phi = 0$，Φ 可以通过逐次的更新来构造，即

$$\Phi \longleftarrow \Phi + \hat{W}_i \hat{W}_i^T, \quad i = 1, \ldots, N. \tag{3.16}$$

充分利用 $\hat{W}_i$ 的稀疏结构，第 i 次可以如下更新，

$$\begin{cases} \Phi(i,i) & \leftarrow & \Phi(i,i) + k_i - r_i, \\ \Phi(J_i,J_i) & \leftarrow & \Phi(J_i,J_i) + W_i W_i^T, \\ \Phi(J_i,i) & \leftarrow & \Phi(J_i,i) - W_i \mathbf{1}_{k_i - r_i}, \\ \Phi(i,J_i) & \leftarrow & \Phi(i,J_i) - (W_i \mathbf{1}_{k_i - r_i})^T. \end{cases} \tag{3.17}$$

虽然有

$$W_i = (1-\alpha_i) w_i^* \mathbf{1}_{k_i - r_i}^T + V_2^{(i)} H_i, \tag{3.18}$$

但在（3.17）中并不要求计算 H_i。实际上，若 $\alpha_i = 0$，则 $H_i = I$ 并且有 $W_i = w_i^* \mathbf{1}_{k_i - r_i}^T + V_2^{(i)}$。否则，将等式 $(V_2^{(i)})^T \mathbf{1}_{k_i} = \alpha H_i \mathbf{1}_{k_i - r_i}$ 代入（3.18），我们可以得到

$$\begin{aligned} W_i \mathbf{1}_{k_i - r_i} &= (k_i - r_i)(1-\alpha_i) w_i^* + \frac{1}{\alpha_i} V_2^{(i)} V_2^{(i)^T} \mathbf{1}_{k_i} \\ W_i W_i^T &= (1-\alpha_i)^2 (k_i - r_i) w_i^* (w_i^*)^T + \frac{1-\alpha_i}{\alpha_i} w^* \mathbf{1}_{k_i}^T V_2^{(i)} V_2^{(i)^T} \\ &\quad + \frac{1-\alpha_i}{\alpha_i} V_2^{(i)} V_2^{(i)^T} \mathbf{1}_{k_i} w^{*T} + V_2^{(i)} V_2^{(i)^T}. \end{aligned} \tag{3.19}$$

这里 w_i^* 可以是（3.3）的一个正则解。

我们归纳修正局部线性嵌入算法如下。

MLLE算法 (Modified Locally linear Embedding).

1. 对于$i=1,\cdots,N$,

   1.1 决定 x_i 的邻域 $\mathcal{N}_i=\{x_j,\ j\in J_i\}$，其中 $J_i=\{i_1,\ldots,i_{k_i}\}$ 以及 $i\notin J_i$。记 $G_i=[x_{i_1}-x_i,\ldots,x_{i_{k_i}}-x_i]$。

   1.2 将一个小的正数 γ 代入（3.3）并计算它的正则化解 w_i^*。

   1.3 计算 $G_i^TG_i$ 的特征值 $\lambda_i^{(1)},\ldots,\lambda_i^{(k_i)}$ 和其对应的特征向量 $v_i^{(1)},\ldots,v_i^{(k_i)}$。计算 $\rho_i=\frac{\sum_{j=d+1}^{k_i}\lambda_i^{(j)}}{\sum_{j=1}^{d}\lambda_i^{(j)}}$。

2. 将 $\rho_i, i=1,\ldots,N$ 按升序重新排列，并设 $\eta=\rho_{\lceil N/2\rceil}$。

3. 对于$i=1,\cdots,N$,

   3.1 利用 η 和（3.15）设置 r_i。记 $V_2^{(i)}=[v_i^{(r_i+1)},\ldots,v_i^{(k_i)}]$ 以及 $\alpha_i=\|1_{k_i}^TV_2^{(i)}\|$。

   3.2 利用（3.17）更新 Φ。当 $\alpha_i<\epsilon$ 时，用 $W_i=w_i^*1_{k_i-r_i}^T+V_2^{(i)}$；否则用（3.19）。

4. 计算 Φ 的 $d+1$ 个最小特征向量并且选取其中对应于第 2 到第 $d+1$ 个最小特征值的特征向量组成特征矩阵，记为 $T=[u_2,\ldots,u_{d+1}]^T$。

由 MLLE 的算法我们可以很容易的给出它的一个计算复杂度的估计。注意到 MLLE 的算法步骤同 LLE 相比主要是在每个邻域多计算了 $G_i^TG_i$ 的特征值分解，它的计算复杂度为 $O(k_i^3)$。假设 $k_i=k$，则 MLLE 比 LLE 更多的计算复杂度为 $O(k^3N)$，同选取邻域的 $O(DN^2)$ 和计算嵌入结果的 $O(dN^2)$ 相比，它只是一个小量。因此，MLLE 在计算时间上同 LLE 并没有多大的差别。

§3.4 MLLE 在等距流形上的分析

为了显示出 MLLE 的效率，我们给出了 MLLE 应用在等距流形上的分析。我们假设数据采自于一个参数化的流形，$x_i=f(\tau_i), i=1,\ldots,N$，其中 $f:\Omega\subset\mathcal{R}^d\to\mathcal{R}^m$ 是充分光滑且等距的映射，Ω 是一个开集。不失一般性，我们假设 $\{\tau_i\}$ 的均值是0。对于等距流形，在样本空间上的重建误差和在参数空间上的重建误差通常是近似相等的。

定理 3.4: 记 $\varepsilon_i=\max_{j\in J_i}\|\tau_j-\tau_i\|$。那么对于任何有界权 w_{ji}，有

$$\|\sum_{j\in J_i}w_{ji}x_j-x_i\|=\|\sum_{j\in J_i}w_{ji}\tau_j-\tau_i\|+O(\varepsilon_i^2). \tag{3.20}$$

进一步的，如果 1_{k_i} 并不正交于 $[\tau_{i_1},\ldots,\tau_{i_{k_i}}]$ 的零空间，那么

$$\min_{1^Tw_i=1}\|\sum_{j\in J_i}w_{ji}x_j-x_i\|=O(\varepsilon_i^2). \tag{3.21}$$

证明： 在 x_i 作一阶泰勒展开，对于它的邻域点 x_j，有

$$x_j = x_i + J_f(\tau_i)\cdot(\tau_j - \tau_i) + O(\|\tau_i - \tau_j\|^2),$$

其中 $J_f(\tau_i) \in \mathcal{R}^{m\times d}$ 是 f 在 τ_i 的雅科比矩阵。我们有

$$\sum_{j\in J_i} w_{ji}x_j = x_i + J_f(\tau_i)\cdot(\sum_{j\in J_i} w_{ji}\tau_j - \tau_i) + \sum_{j\in J_i} w_{ji}O(\|\tau_i - \tau_j\|^2).$$

由于 $w_{ji}, j \in J_i$ 是有界的并且对于 $j \in J_i$ 有 $\|\tau_j - \tau_i\| \le \varepsilon_i$，那么有

$$\|\sum_{j\in J_i} w_{ji}x_j - x_i\| = \|J_f(\tau_i)\cdot\sum_{j\in J_i} w_{ji}(\tau_j - \tau_i)\| + O(\varepsilon_i^2).$$

由 f 的等距性可知 $J_f(\tau_i)$ 是正交的，从而等式（3.20）成立。进一步有

$$\min_{1^T w_i=1} \|\sum_{j\in J_i} w_{ji}x_j - x_i\| = \min_{1^T w_i=1} \|\sum_{j\in J_i} w_{ji}\tau_j - \tau_i\| + O(\varepsilon_i^2).$$

如果 1_{k_i} 不正交于$[\tau_{i_1},\ldots,\tau_{i_{k_i}}]$的零空间，那么由推论 3.1，可得

$$\min_{1_{k_i}^T w_i=1} \|\sum_{j\in J_i} w_{ji}\tau_j - \tau_i\| = 0.$$

这样立刻可推出 (3.21)成立。 ■

定理 3.4所说的是，数据点 $\{x_i\}$ 的重建权向量近似于在低维空间的投影点 $\{\tau_i\}$ 的重建权向量。而且，用 $G = G_i$ 代入定理 3.3，我们可以得到权矩阵 $W_i = [w_i^{(1)},\ldots,w_i^{(k_i-r_i)}]$ 满足

$$\|G_iW_i\| \le \sqrt{k_i - r_i}\min_{1^T w_i=1} \|\sum_{j\in J_i} w_{ji}x_j - x_i\| + \sqrt{\sum_{j=r_i+1}^{k_i}\sigma_j^2(G_i)}.$$

我们可以推出

$$\|G_iW_i\| \le \sqrt{\sum_{j=r_i+1}^{k_i}\sigma_j^2(G_i) + O(\varepsilon_i^2)}$$

以及

$$E(T^*) = \sum_{i=1}^{N}\sum_{\ell=1}^{k_i-r_i} \|\sum_{j\in J_i} w_{ji}^{(\ell)}\tau_j - \tau_i\|^2 \le \sum_{i=1}^{N}\sum_{j=r_i+1}^{k_i}\sigma_j^2(G_i) + O(\max_i \varepsilon_i^2).$$

我们注意到 T^* 的行向量也许不是相互正交的。我们将这些行正交化来得到正交阵 U，并且可以将 T^* 表示成 $T^* = LU$，其中 $L = T^*U^T \in \mathcal{R}^{d\times d}$。由 $\sigma_d(L) = \sigma_d(T^*)$，我们可以得到

$$E(U) \le E(T^*)/\sigma_d^2(T^*) \le \sum_{i=1}^{N}\sum_{j=r_i+1}^{k_i} \sigma_j^2(G_i)/\sigma_d^2(T^*) + O(\max_i \epsilon_i^2).$$

通过对 r_i 的选取，对于一个小的 η 可以使得 $\sum_{j=r_i+1}^{k_i}\sigma_j^2(G_i) \le \eta\sum_{j=1}^{r_i}\sigma_j^2(G_i)$ 也是一个小值。由我们在前一节中关于 η 的设置可知 η 是一个小值。这样 $E(U)$ 也是一个小值，这意味着 MLLE 能恢复出等距的嵌入。

§ 3.5 与LTSA的比较

§ 3.5.1 邻域点的线性相关性

MLLE 用 $k_i - r_i$ 个线性无关的权向量 $w_i^{(1)}, \ldots, w_i^{(k_i-r_i)}$ 来构造邻域点之间的线性关系，其中 $w_i^{(1)}, \ldots, w_i^{(k_i-r_i)}$ 是极小化问题

$$\min_{\sum_{j\in J_i} w_{ji}=1} \|\sum_{j\in J_i} w_{ji}x_j - x_i\|$$

的解或近似解。出于分析和比较上的便利，我们假设对所有的 i 有 $r_i = d$。用 $\mathcal{N}_i$ 表示包括 x_i 自身的邻域点的集合。误差

$$\epsilon^{MLLE}(\mathcal{N}_i) = \sum_{\ell=1}^{k_i-d} \|\sum_{j\in J_i} w_{ji}^{(\ell)}x_j - x_i\|^2 = \|G_iW_i\|_F^2$$

定义了一个邻域集合线性相关性的测度。如果我们记

$$\bar{X}_i = [\ldots, x_j - \bar{x}_i, \ldots]_{j\in I_i},$$

其中 $I_i = \{i\}\cup J_i$，$\bar{x}_i = \frac{1}{|I_i|}\sum_{j\in I_i} x_j$ 是 x_i 包括自身的邻域的中心，那么可以得出 $G_iW_i = \bar{X}_i\tilde{W}_i$，其中 $\tilde{W}_i = \hat{W}_i(I_i,:)$。这样就有 $\epsilon^{MLLE}(\mathcal{N}_i) = \|\bar{X}_i\tilde{W}_i\|_F^2$。

在 LTSA 中，局部的线性结构是用下列的最佳线性拟合来构造，

$$\min_{c_i,\theta_j, U_i^TU_i=I} \sum_{j\in I_i}\|x_j - (c_i + U_i\theta_j)\|^2.$$

记 $\bar{X}_i = [Q_i, Q_i^\perp]\,\mathrm{diag}(\Sigma_i, \tilde{\Sigma}_i)[V_i, V_i^\perp]^T$ 为 $\bar{X}_i$ 的奇异值分解，其中 Q_i 是 $\bar{X}_i$ 对应最大的 d 个奇异值的左奇异向量构成的矩阵。则上面的最优问题的解为 $c_i = \bar{x}_i$，$U_i = Q_i$ 以及 $\theta_j = \theta_j^{(i)} = Q_i^T(x_j - \bar{x}_i)$。这样 $\mathcal{N}_i$ 的线性相关性也能用下列的式子度量，

$$\epsilon^{LTSA}(\mathcal{N}_i) = \sum_{j\in I_i}\|x_j - \bar{x}_i - Q_i\theta_j^{(i)}\|^2 = \|\bar{X}_i - Q_i\Theta_i\|_F^2 = \|\bar{X}_iV_i^\perp\|_F^2,$$

其中 $V_i^{\perp}$ 是 span$([1_{k_i+1}, \Theta_i^T])$ 的零空间的正交基。在 MLLE 和 LTSA 中用以估计邻域点之间相关性的测度函数 ϵ^{MLLE} 和 ϵ^{LTSA} 是很相似的：

$$\epsilon^{MLLE}(\mathcal{N}_i) = \|\bar{X}_i\tilde{W}_i\|_F^2, \quad \epsilon^{LTSA}(\mathcal{N}_i) = \|\bar{X}_i V_i^{\perp}\|_F^2.$$

§ 3.5.2 排列矩阵

MLLE 和 LTSA 都是通过极小化一个排列矩阵 Φ 的迹函数来得到嵌入结果，

$$\min_{TT^T=I} \text{trace}(T\Phi T^T).$$

这个排列矩阵可以同样的写成 $\Phi = \sum_{i=1}^{N} S_i\Phi_i S_i^T$，其中 Φ_i 是一个由所构造的局部线性结构决定的正半定矩阵，S_i 是一个满足 $XS_i = X(:, I_i)$ 的选择矩阵。我们分别用 Φ^{MLLE} 和 Φ^{LTSA} 表示 MLLE 和 LTSA 的排列矩阵，并且 Φ_i^{MLLE} 和 Φ_i^{LTSA} 分别表示相应的局部算子且有

$$\Phi_i^{MLLE} = \tilde{W}_i\tilde{W}_i^T, \quad \Phi_i^{LTSA} = V_i^{\perp}(V_i^{\perp})^T,$$

其中 $V_i^{\perp}$ 是零空间 span$([1, \Theta_i^T])$ 上的正交投影，见 [69]。下面定理给出了 span$(\tilde{W}_i)$ 和 span$(V_i^{\perp})$ 的距离一个上界.

定理 3.5： span$(\tilde{W}_i)$ 和 span$(V_i^{\perp})$ 之间的距离 dist$(\tilde{W}_i, V_i^{\perp})$ 有如下的界，

$$\text{dist}(\tilde{W}_i, V_i^{\perp}) \leq \frac{\|G_iW_i\|}{\sigma_d(\tilde{W}_i)\sigma_d(\bar{X}_i)}$$

证明： 记 Q_i, V_i 为对应 $\bar{X}_i$ 的 d 个最大的奇异值的左奇异向量和右奇异向量所构成的矩阵。记 $\tilde{W}_i = \tilde{Q}_iR_i$ 为 $\tilde{W}_i$ 的 QR 分解。距离 dist$(\tilde{W}_i, V_i^{\perp})$ 定义为

$$\text{dist}(\tilde{W}_i, V_i^{\perp}) = \|\tilde{Q}_i^T V_i\|.$$

（见 [28] 中关于子空间距离的讨论。）注意到 $G_iW_i = \bar{X}_i\tilde{W}_i = \bar{X}_i\tilde{Q}_iR_i$ 以及 $Q_i^T\bar{X}_i = \Sigma_i V_i^T$，我们有

$$Q_i^T G_iW_i = \Sigma_i V_i^T\tilde{Q}_iR_i.$$

由于 $\tilde{W}_i$ 是满秩的，从而 R_i 是非奇异的。这就得出 $V_i^T\tilde{Q}_i = \Sigma_i^{-1}Q_i^T G_iW_iR_i^{-1}$。因此

$$\|V_i^T\tilde{Q}_i\| \leq \|\Sigma_i^{-1}\|\|G_iW_i\|\|R_i^{-1}\| = \|G_iW_i\|/(\sigma_d(\tilde{W}_i)\sigma_d(\bar{X}_i)).$$

得出结论。 ■

若 x_i 和它的邻域点近似在一个 d 维超平面，G_iW_i 很小。这样由定理 3.5可知，$\tilde{W}_i$ 和 $V_i^{\perp}$ 的列空间近似相同的。所不同之处在于 Φ_i^{LTSA} 是正交投影，而投影 Φ_i^{MLLE} 不是正交的。

§3.6 本章小结

LLE 是流形学习方面经典的局部非线性方法，它有参数少、计算快、易求全局最优解等优点，并在图像分类、图像识别、谱重建、数据可视化等方面都有着广泛的应用[23, 40, 52]。但是，LLE 可能会将相隔较远的点映射到低维空间中邻近点的位置，从而导致嵌入结果有着比较明显的扭曲。这其中的一个重要原因是，LLE 采用的单个重构权并不能完全的反映出流形的局部几何性质。此外，用以求解重构权的有约束的最小二乘问题的最优解也许不是唯一的，而且 LLE 采用正则化方法求解涉及到正则因子 γ 的选取，难以保证所求的解是最优解。在本章中，我们对 LLE 的重构权向量的性质进行详细的分析，在理论上证明了（用正则化方法）确定最优权在数值上是不稳定的，同时在给定精度下，存在着多组线性无关的近似最优权向量。我们用这组线性无关的权向量来表示数据点的局部线性结构，并且在低维的嵌入中保持了这种线性结构。这种修正的 LLE 方法（Modified Locally Linear Embedding Using Multiple Weights）有着很好的稳定性。我们从理论上证明了 MLLE 对采自等距流形的样本点有着理想的结果，通过 MLLE 和 LTSA 之间的详细对比和理论分析，揭示了 LLE、MLLE 和 LTSA 之间的内在联系。这为进一步理解与分析建立了基础。我们将在第六章用数值例子说明 MLLE 的效率。

第 4 章　自适应邻域选取

本章要点：

- 邻域选取对流形学习效果的影响
- 邻域选取的标准
- 邻域压缩策略
- 邻域扩张策略

在前面几章中，我们介绍了流形学习的一些算法，包括 Isomap 、 LLE 、 HLEE、 LTSA 、 MLLE 等。所有的流形学习方法首先面对的都是邻域选取问题，需要选取出一个合适的邻域来获取局部的线性信息。很显然的，邻域越小可以认为邻域的线性结构越明显，但是我们需要注意的是，邻域之间需要有足够的交叠以保证较远的点之间有足够的联系，这又使得邻域不能过小。从直观上想像，流形上曲率大的样本点的邻域应该小一些，而流形上曲率小的样本点处的邻域可以大一些。因此，关于邻域选取我们需要考虑的问题是：在加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何结构？在本章中，我们会提出解决这个问题的方法，并由此给出相应的算法。我们将主要基于局部切空间排列 （LTSA） 算法来讨论这个问题。我们会给出关于邻域的局部线性逼近的分析，并由此给出一个判断邻域集是否能在给定精度内被一个线性拟合所逼近的标准。接下来，我们会采用邻域压缩和邻域扩张两个策略来选取邻域集并使得选出的邻域集能满足这个标准。我们的方法从理论上保证了所选出的邻域在匹配流形的局部几何性质的前提下，尽可能的扩张邻域以加强样本点之间的关联性。

§ 4.1　邻域选取对流形学习效果的影响

我们再来回顾一下邻域选取的两种策略。一种策略是 k-邻域策略，对每个样本点，选其在欧氏距离下最近的 k 个点作为邻域点，即

$$\mathcal{N}_k(x_i) = \{x_i\text{在数据集中最近的 } k \text{ 个点}\}.$$

另一种是 ϵ-邻域策略 [56, 60]，

$$\mathcal{N}_\epsilon(x_i) = \{x_j \mid \|x_j - x_i\| \le \epsilon\},$$

其中 $\|\cdot\|$ 表示一个向量的 2-范数。对一致分布的样本点，k-邻域和 ϵ-邻域策略是大致等价的：只要选取合适的 k 和 ϵ，两种策略都能找到同样的邻域点。通常情况下，k-邻域

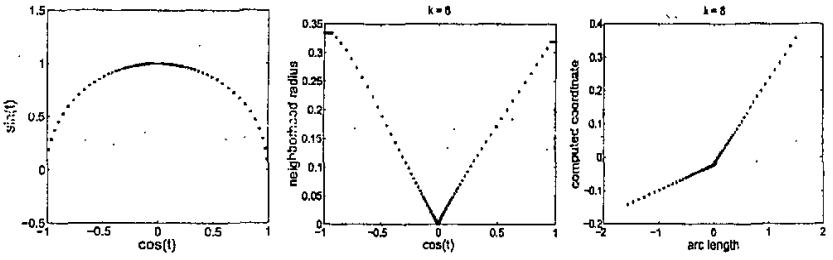

图 4.1: 例 4.1: 不同密度的数据集（左），邻域大小（中）以及 LTSA 计算的结果（$k=8$）vs. 中心化的弧长（右）。

策略能更好的控制非一致分布的样本点，因此 k-邻域策略更多的被采用。如同在 [69] 所示，局部邻域之间更多的交叠能够产生更好的特征空间用以展开流形的全局嵌入，因此邻域选取时偏向大的邻域。但是，对于更大的邻域，流形的局部线性逼近的精度也许会遭到破坏，即选取出的邻域也许不能反映出流形的局部几何性质。而且，流形曲率的变化以及数据点的样本密度的变化，此外还有噪音都会影响邻域的大小和局部线性逼近的精度，并且使问题进一步的复杂化。在本节中，我们会主要讨论这些问题。

我们首先给出两个一维流形的例子。一个例子的样本点的密度高度不一致，并且流形的所有点的曲率都是 1。在另一例子中，流形的曲率高度变化且样本点的密度也是高度的不一致。对于这两个数据集，采用 k-邻域选取策略的 LTSA 算法都失败了，这主要是由于以下两个原因：（1）对于高密度区域的样本点，所构造的邻域同邻近的邻域之间并没有足够的交叠；（2）大曲率的样本点导致了很差的局部线性逼近。

例 4.1： 我们考虑一个有着一致的曲率但密度高度不一致的 2 维数据集作为例子。数据点 $x_i = [\cos(t_i), \sin(t_i)]^T, i = 1\ldots, N$，它采自一个生成参数为 t_i 的半圆，t_i 选择方式为

$$t_{i+1} = t_i + 0.1(0.001 + |\cos(t_i)|),$$

起点 $t_1 = 0$ 而终点 t_N 略小于 π，这样共有 $N = 152$ 个样本点。$t_1, \ldots, t_N$ 表示了这个半圆的弧长坐标*。显然，这个流形在每个点的曲率都为 1。

例 4.2： 数据点生成如下，

$$x_i = [t_i, 10e^{-t_i^2}]^T, \quad i = 1\ldots, N,$$

* 若 $x_i = [\phi(t_i), \psi(t_i)]^T$ 且 $t_1 < \cdots < t_N$，弧长坐标可通过 $s_i = \int_{t_1}^{t_i} \sqrt{(\dot{\phi}(t))^2 + (\dot{\psi}(t))^2}dt$ 计算。在稍后的图中，我们采用中心化的弧长 $\hat{s}_i = s_i - \frac{1}{2}s_N$。

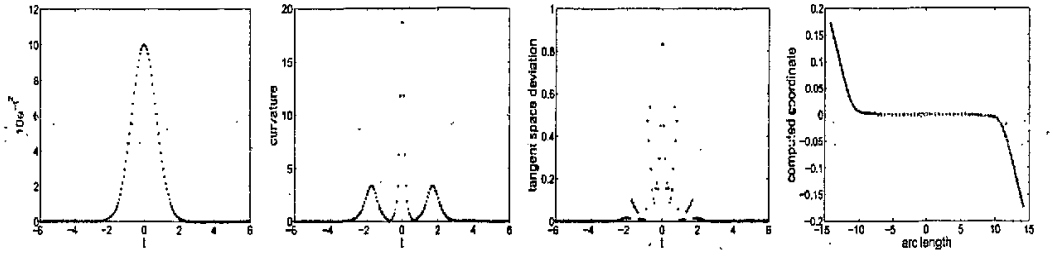

图 4.2: 例 4.2: (从左到右) 数据集，曲率，计算切空间的偏离，以及 LTSA 在邻域 $k = 8$ 时的结果。

采自于曲线 $f(t) = [t, 10e^{-t^2}]^T$，其中 t_i 在区间 $[-6, 6]$ 中均匀分布。我们采集 $N = 180$ 个样本点。由于流形的曲率函数

$$c(t) = \frac{20|1 - 2t^2|e^{-t^2}}{(1 + 40t^2e^{-2t^2})^{3/2}}$$

在 $[-6, 6]$ 内从 $c_{\min} = 0$ 变化到 $c_{\max} = 20$，样本点的密度也是高度不一致的。

上面两个例子中的样本点分布画在图 4.1和图 4.2的左边。理想的情况下，计算的嵌入结果 τ_i 应该是关于弧长 τ_i^* 的一个线性函数，即 2 维点 $\{(\tau_i^*, \tau_i), i = 1, \ldots, N\}$ 应该是一条直线。然而，对这两个例子 LTSA 都不能产生好的嵌入结果，即计算的坐标并不是线性依赖于弧长。

对于第一个例子，那些有着高密度的样本点的邻域太小以至于相邻的邻域之间不能有足够的交叠来得到一个好的排列。我们可以用 $r_i = \max_j \|x_{i_j} - \bar{x}_i\|$ 来度量邻域半径的变化，我们将 r_i 画在图 4.1的中间。在图 4.1的右边，我们画出 2 维点 $\{(\tau_i^*, \tau_i)\}$，其中 τ_i 是 LTSA 采用邻域大小 $k = 8$ 的计算结果，这条线很明显的在那些有着最大密度的样本点处折段。这种现象对于其它邻域大小 k 也会出现。对于第二个例子，LTSA 之所以又失败是因为在大曲率的点 x_i 处，通过局部最佳线性拟合于邻域集 $\mathcal{N}_i$ 而得到的线性子空间 $\mathrm{span}(q_i)$ 很大的偏离了流形上这些点的切空间 $\mathrm{span}(u_i)$。在图 4.2的第二幅中，我们画出了这条 1 维曲线的曲率以及这种偏离度 $\mathrm{dev}(\mathcal{N}_i)$，其中

$$\mathrm{dev}(\mathcal{N}_i) = \sqrt{1 - (q_i^T u_i)^2},$$

是两个单位向量 q_i 和 u_i 之间夹角的 sine 值*。

即使流形有着几乎一致的曲率并且样本密度的变化并不大，当采用 k-邻域或 ϵ-邻域策略的时候，要选择一个合适的值 k 或 ϵ 也并不容易。为了说明这种情况，我们来考虑下面一个例子。

*在下一节中，我们会给出更高维的子空间之间偏离的定义。

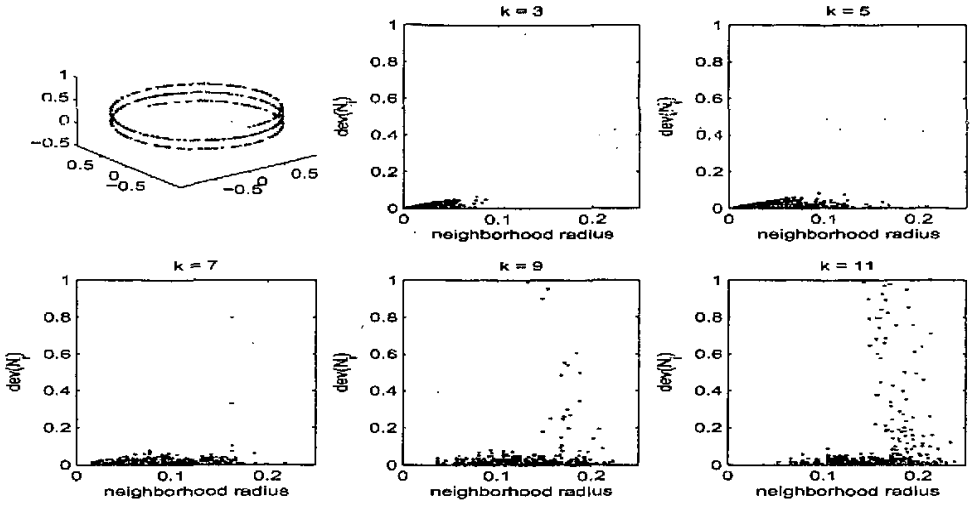

图 4.3: 例 4.3: 局部线性拟合偏离 vs. 采用不同 k 的 k-邻域策略选取的邻域的半径。

例 4.3: 我们构造一个有 $N = 500$ 个点的 3 维数据集如下，

$$x_i = [\sin(t_i), \cos(t_i), 0.02t_i]^T, \quad i = 1, \ldots, N,$$

其中 t_i 一致分布在区别 $[0, 4\pi]$。对于小的 k，局部最佳线性拟合同切空间之间的偏离往往也比较小。可是，当增大 k 的值，由于参数 t_i 的变化对样本点的第三个分量的影响很小，k-邻域策略会产生坏的邻域（局部线性拟合会很大的偏离切空间）。这种偏离现象很明显的体现在图 4.3: 当 k 增大时，邻域大小和最小邻域半径都随之增加，而坏邻域数量也随之增加。

上面三个例子很明显的说明有必要有更好的策略来选择邻域的大小以便在保持邻域之间交叠的同时能匹配流形的曲率。在本章接下来的内容中，我们将解决这个问题。

§ 4.2 自适应邻域选取方法

调整邻域大小以匹配流形的局部几何结构的想法是很自然的，这在 [57] 和 [67]中都提及。可是，并没有提及如何解决这个问题。在这节中，我们给出一种方法以自适应的选择满足下面两个要求的邻域：

- 对每个样本点，选择的邻域点应该能反映流形的局部几何结构，这样邻域集的最佳线性拟合所决定的线性子空间能高精度的逼近流行的切空间。
- 在相邻的邻域之间应该保持足够大的交叠以便加强局部信息的传播效率 [69]。

首先，我们给出一个标准来判断是否一个邻域集能满足上面的第一个要求，即理想的切空间能被邻域集的最佳线性拟合在一个给定精度内逼近。接下来，我们提出一种压缩算法来决定一个（相对小的）邻域集来满足这个标准。最后，我们将讨论如何尽可能大的扩张这个几何以满足第二项要求，同时最终的邻域集仍然满足提出的标准。

§4.2.1 切空间逼近标准

让数据点 $\{x_i\}$ 采自下列的参数化流形 $\mathcal{M} = f(\Omega)$，其中 $f : \Omega \subset \mathcal{R}^d \to \mathcal{R}^m$ 是一个定义在开连通集 Ω 上的光滑映射。考虑流形上一个选定点附近的局部线性结构，即 $x = f(\tau)$，流形 $\mathcal{M}$ 在 x 附近的局部线性结构能被流形在点 x 处的切空间所描述。对 f 在 x 点作一阶泰勒展开，x 的邻域点 $\hat{x} = f(\hat{\tau})$ 可以表示成

$$\hat{x} = x + J_\tau \cdot (\hat{\tau} - \tau) + \varepsilon(\tau, \hat{\tau}). \tag{4.1}$$

这里 $J_\tau \in \mathcal{R}^{m\times d}$ 是 f 在 τ 点的雅科比矩阵，它的列张成切空间*，而 $\varepsilon(\tau, \hat{\tau})$ 表示主要被 f 的海赛张量 $H_\tau = [\frac{\partial^2 f}{\partial\tau_i \partial\tau_j}]$ 所逼近的高阶项。显然的，$\|\varepsilon(\tau, \hat{\tau})\| = O(\|\hat{\tau} - \tau\|^2)$度量了$\hat{x}$同切空间的逼近误差。如果我们对海赛张量 H_τ 给出一个上界 h，那么误差项的上界为 $\|\varepsilon(\tau, \hat{\tau})\| \le h\|\hat{\tau} - \tau\|^2$。在这种情况下，

$$\|\hat{x} - x - J_\tau(\hat{\tau} - \tau)\| \le h\|\hat{\tau} - \tau\|^2.$$

海赛张量 H_τ 以及它的上界 h 依赖于流形的局部曲率。对于一个选定的参数 $\eta > 0$，如果我们用 $\Omega_\tau = \{\hat{\tau} : h\|\hat{\tau} - \tau\| \le \eta\}$ 定义 τ 在参数空间上的邻域，那么 $\mathcal{N}_x = \{\hat{x} = f(\hat{\tau}) : \hat{\tau} \in \Omega_\tau\}$ 是 x 在流形上的邻域。这样，$\hat{x} \in \mathcal{N}_x$ 满足

$$\|\hat{x} - x - J_\tau(\hat{\tau} - \tau)\| \le \eta\|\hat{\tau} - \tau\|. \tag{4.2}$$

我们可以用（4.2）作为一个选择邻域的标准。邻域

$$\{\hat{x} = f(\hat{\tau}) : \|\hat{x} - x - J_\tau(\hat{\tau} - \tau)\| \le \eta\|\hat{\tau} - \tau\|\} \supset \mathcal{N}_x$$

显然依赖于流形在 x 附近的局部曲率。如果选定 η，x 附近越小的曲率会导致越大的邻域，而大的曲率会趋向于收缩邻域。

在实际计算中，由于雅科比阵 J_τ 和 τ 是未知的，我们有必要对（4.2）做一些小的改动。考虑 $x_i = f(\tau_i)$ 的邻域集 $\mathcal{N}_i = \{x_{i_1}, \ldots, x_{i_{k_i}}\}$，由（4.2）有

$$\|x_{i_j} - x_i - J_{\tau_i}(\tau_{i_j} - \tau_i)\| \le \eta\|\tau_{i_j} - \tau_i\|. \tag{4.3}$$

*对一个等距流形，雅克比矩阵是正交的，即 $J_\tau^T J_\tau = I$。

记 $\mathcal{N}_i$ 的邻域点 x_{i_j} 的最佳线性拟合为 $\bar{x}_i + Q_i\theta_j^{(i)}$，则由 PCA 的有关知识可知，$\bar{x}_i$ 为邻域集 $\mathcal{N}_i$ 的中心点，Q_i 为矩阵 $[x_{i_1} - \bar{x}_i, \ldots, x_{i_{k_i}} - \bar{x}_i]$ 的最大 d 个奇异值对应的左奇异向量组成的矩阵，$\theta_j^{(i)} = Q_i^T(x_{i_j} - \bar{x}_i)$。我们将邻域点 x_{i_j} 的未知的线性逼近 $x_i - J_{i_i}(\hat{\tau}_{i_j} + \tau_i)$ 近似的用 $\mathcal{N}_i$ 的最佳线性拟合 $\bar{x}_i + Q_i\theta_j^{(i)}$ 来代替，并且局部距离 $\|\hat{\tau}_{i_j} - \tau_i\|$ 近似等于 $\|\theta_j^{(i)}\|$，这样标准（4.3）变成

$$\|x_{i_j} - \bar{x}_i - Q_i\theta_j^{(i)}\| \le \eta\|\theta_j^{(i)}\|, \quad j = 1, \ldots, k_i,$$

或者用更为简洁的矩阵形式

$$\|X_i - (\bar{x}_i\mathbf{1}_{k_i}^T + Q_i\Theta_i)\|_F \le \eta\|\Theta_i\|_F, \tag{4.4}$$

其中 $X_i = [x_{i_1}, \ldots, x_{i_{k_i}}]$ 及 $\Theta_i = [\theta_1^{(i)}, \cdots, \theta_{k_i}^{(i)}]$。

很显然，（4.4）中的范数 $\|\Theta_i\|_F$ 和 $\|X_i - (\bar{x}_i\mathbf{1}_{k_i}^T + Q_i\Theta_i)\|_F$ 能用 $X_i - \bar{x}_i\mathbf{1}_{k_i}^T$ 的奇异值来表示。记 $\sigma_1^{(i)} \ge \cdots \ge \sigma_d^{(i)} \ge \cdots \ge \sigma_{k_i}^{(i)}$ 为矩阵 $X_i - \bar{x}_i\mathbf{1}_{k_i}^T$ 的奇异值，V_i 为最大 d 个奇异值对应的右奇异向量组成的矩阵，则有

$$\Theta_i = [\theta_1^{(i)}, \cdots, \theta_{k_i}^{(i)}] = Q_i^T(X_i - \bar{x}_i\mathbf{1}_{k_i}^T) = \mathrm{diag}(\sigma_1^{(i)}, \ldots, \sigma_d^{(i)})V_i^T.$$

从而有 $\|\Theta_i\|_F = \sqrt{\sum_{j\le d}(\sigma_j^{(i)})^2}$ 和

$$\|X_i - (\bar{x}_i\mathbf{1}_{k_i}^T + Q_i\Theta_i)\|_F = (\|X_i - \bar{x}_i\mathbf{1}_{k_i}^T\|_F^2 - \|\Theta_i\|_F^2)^{1/2} = \sqrt{\sum_{j>d}(\sigma_j^{(i)})^2}.$$

因此（4.4）等价于

$$\sqrt{\sum_{j>d}(\sigma_j^{(i)})^2} \le \eta\sqrt{\sum_{j\le d}(\sigma_j^{(i)})^2}. \tag{4.5}$$

这就是我们用来自适应选择邻域的标准。在下面两节中，我们将给出算法来选择满足这个标准的集合。

§ 4.2.2 邻域压缩

对于样本点 x_i，假设我们有相对大的一个邻域集 $\mathcal{N}_i = \{x_{i_1}, \ldots, x_{i_{K_{\max}}}\}$。我们可以用 k-邻域策略确定这样的邻域集。我们预先选择参数 $\eta < 1$。如果（4.5）不被满足，我们通过移出离邻域均值 $\bar{x}_i$ 最远的邻域点来压缩邻域集。这种压缩步骤可以不断的重复直到（4.5）成立，或者直到邻域被压缩到一个预先设定的最小邻域大小 $k_{\min}$。如果没有找到能满足（4.5）的邻域，我们选一个在压缩过程中使比值 $r = \sqrt{\sum_{j>d}\sigma_j^2 / \sum_{j\le d}\sigma_j^2}$ 最小的 k-邻域作为压缩的结果。我们在下列算法中实现这个压缩过程。

邻域压缩算法

C0 决定初使的 $k = K_{\max}$ 和 x_i 的 k-邻域 $X_i^{(k)} = [x_{i_1}, \ldots, x_{i_k}]$，邻域按离 x_i 的距离升序排列，即 $\|x_{i_1} - x_i\| \le \|x_{i_2} - x_i\| \le \cdots \le \|x_{i_k} - x_i\|$。

C1. 计算 $X_i^{(k)} - \bar{x}_i^{(k)} 1_k^T$ 的奇异值 $\{\sigma_j^{(k,i)}\}$ 以及比值

$$r_i^{(k)} = \sqrt{\sum_{j>d} (\sigma_j^{(k,i)})^2 / \sum_{j \le d} (\sigma_j^{(k,i)})^2}.$$

C2. 如果 $r_i^{(k)} < \eta$，那么设 $X_i = X_i^{(k)}$ 并且算法结束，否则进行下一步。

C3. 如果 $k > k_{\min}$，去除 $X_i^{(k)}$ 的最后一列并得到 $X_i^{(k-1)}$，设 $k = k-1$ 并且返回步骤C1；否则，进入步骤C4。

C4. 计算 $k_i = \arg\min_{k_{\min} \le k \le K_{\max}} r_i^{(k)}$ 且设 $X_i = X_i^{(k_i)}$。

从算法上看，每当我们从当前的邻域中删除一个样本点的时候，似乎需要计算 d 个最大的奇异值。这里有两种方法来加快这个步骤：（1）我们能在每步中删除几个样本点而不是一个样本点，这样能减少步骤 C1 所需要的执行次数；（2）$\sigma_j^{(k+1,i)}$ 能被更新来得到 $\sigma_j^{(k,i)}$，而且我们不需要从头开始。这是著名的 SVD downdating problem，在[28]中有很详细的描述。在我们的数值实验中，对于 $d = 2$，我们设 $k_{\min} = 5$ 或 6。目前，还不清楚应该选多大的 $K_{\max}$ 最好，但在数值实验中，算法对于初使的 $K_{\max}$ 并不是很敏感。

子空间偏离. 线性拟合所决定的正交矩阵 Q_i 张成一个 d 维的线性空间 $\mathrm{span}(Q_i)$。它同理想的切空间 $\mathrm{span}(J_{\tau_i})$ 之间的偏离度 $\mathrm{dev}(\mathcal{N}_i)$ 可以用 $\mathrm{span}(Q_i)$ 和 $\mathrm{span}(J_{\tau_i})$ 之间的距离来度量，定义如下，

$$\mathrm{dev}(\mathcal{N}_i) = \mathrm{dist}(\mathrm{span}(Q_i), \mathrm{span}(J_{\tau_i})) \equiv \|P_{\mathrm{span}(Q_i)} - P_{\mathrm{span}(J_{\tau_i})}\|,$$

其中 $P_{\mathcal{X}}$ 表示到线性子空间 $\mathcal{X}$ 上的正交投影 [28, p.76]。如果 G_i 是切空间 $\mathrm{span}(J_{\tau_i})$ *的一组正交基所构成的矩阵，那么

$$\|P_{\mathrm{span}(Q_i)} - P_{\mathrm{span}(J_{\tau_i})}\| = \|Q_i Q_i^T - G_i G_i^T\| = \sqrt{1 - \sigma_{\min}^2(Q_i^T G_i)}, \tag{4.6}$$

其中 $\sigma_{\min}(\cdot)$ 是矩阵的最小奇异值。（见 [28]中定理2.6.1证明中的最后一个等式。）所以我们可以很简单就得到偏离度量

$$\mathrm{dev}(\mathcal{N}_i) = \sqrt{1 - \sigma_{\min}^2(Q_i^T G_i)}. \tag{4.7}$$

*如果 f 是等距的，J_τ 是 $\mathrm{span}(J_{\tau_i})$ 的一组正交基。

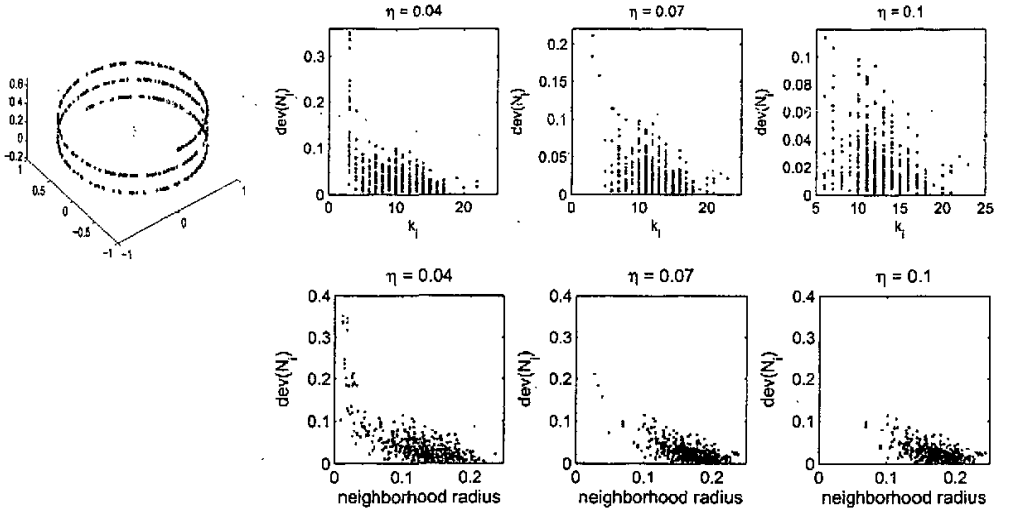

图 4.4: 对例 4.3中的数据加入噪音后，采用不同参数 η（$k_{\min}=3$，$K_{\max}=30$）的邻域压缩算法所得到的邻域的偏离度。左边: 数据集。右边三列: 偏离度 $\mathrm{dev}(\mathcal{N}_i)$ vs. 邻域大小 k_i（顶行）或邻域半径（底行）。

特别的，如果 $\mathrm{span}(Q_i)$ 和 $\mathrm{span}(J_{\tau_i})$ 都是 1 维空间，那么有 $\sigma_{\min}(Q_i^T G_i)=\cos(\theta_{Q_i,G_i})$ 即向量 Q_i 和 G_i 之间夹角的 cosine 值，以及 $\mathrm{dev}(\mathcal{N}_i)=\sin(\theta_{Q_i,G_i})$*。

例4.3（续）：甚至对于有噪音的数据集，压缩方法也能很大程度的改善 k-邻域策略所选出的邻域集。我们往例 4.3中的数据集中增加一些噪音，

$$x_i=[\sin(t_i),\cos(t_i),0.02t_i]^T+0.003\epsilon_i,\quad i=1,\ldots,N,$$

其中 ϵ_i 有着标准的正态分布，见图 4.4的左边。对于每个样本点，我们采用邻域压缩算法并以 $k_{\min}=3$、$K_{\max}=30$ 以及 η 分别为 0.04、0.07和0.1 作为参数来得到它们压缩后的邻域。在图 4.4的顶行，我们画出了最佳线性拟合的偏离度 $\mathrm{dev}(\mathcal{N}_i)$ 同邻域点的个数 k_i 的对比；而在底行，我们画出它同邻域半径的对比。对于小的 η，一些邻域集也许会达到最小的 $k_{\min}$，但并不满足标准（4.5）。对那些邻域集，如图 4.4的顶行所示，它们的偏离度也许会相对的大一些。

§ 4.2.3 邻域扩张

在一个维数 $d>1$ 的流形上一个点的不同方向的曲率通常是不同的。用压缩算法计算的邻域集趋向于匹配这个点的最大曲率。这样，即使存在一个相对大的邻域同流形的

*两个子空间的距离也可以认为是两个不同子空间中两个向量的最大夹角的 sine 值。

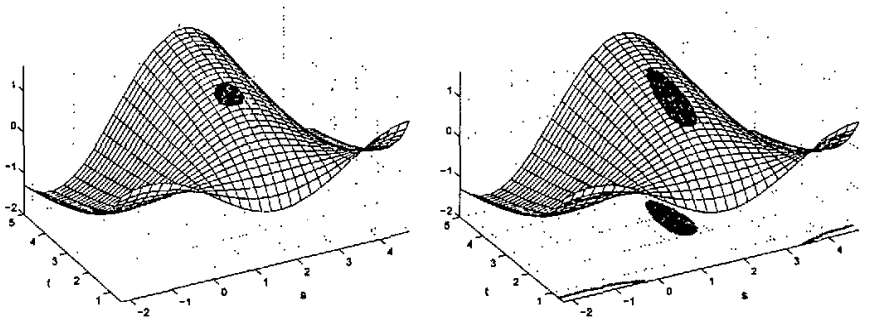

图 4.5: 例 4.4中的 sin-log 流形。流形上逼近 x_0 处的切空间到一个预定精度的邻域集。左边显示的是 ϵ-邻域，右边是逼近切空间的最大（非连通）区域。

曲率匹配，压缩算法计算出的邻域也可能会过小。这个现象在一个曲率高度变化的样本点处尤其明显。下面用一个例子来说明这种现象。

例 4.4: 我们考虑一个 3 维空间中的 2 维流形，参数化如下

$$f(s,t) = [s,\, t,\, \sin(s)\ln(t)]^T, \quad s,t \in [-0.7\pi,\, \pi] \times [0.5,\, 5]$$

并且将流形画在图 4.5。考虑对应 $s = \pi/2$ 和 $t = 2.8$ 的样本点 x_0。样本点 x_0 处的两个主曲率是 $c_{\min} = 0.33850$ 和 $c_{\max} = 0.96964$。预先给定 $\mu = 0.05$，我们可以计算出流形上到 x_0 处的切平面 $\mathcal{T}$ 的距离小于 μ 的 x_0 的最大 ϵ-邻域。在图 4.5中的左边，流形上的阴影区域就是这样的 ϵ-邻域。由于流形在 x_0 处各向相异的形状，这个邻域集合能扩张到一个相对更大的区域而仍然维持到切空间的距离小于 $\mu = 0.05$。在图 4.5的右边，我们用流形上的深色阴影区域表示扩张后的邻域集。这个最大的邻域集合 $\mathcal{D} = \{x \in \mathcal{M},\, \mathrm{dist}(x, \mathcal{T}) \le \mu\}$ 是非连通的。为了清晰的表明它的非连通性，我们画出了 $\mathcal{D}$ 到下面 st-平面的投影。注意到我们要找的是尽可能大的连通区域，就像在图中右边流形中间所示的深色区域。

上面的例子表明利用流形的各向相异的形状，我们能够得到更大的邻域。一个可能的方法是一旦压缩步骤完成，我们试图再加回一些在 x_i 初始邻域中未被选中的点，而与此同时，我们要尽可能的满足条件（4.4）。下面定理 4.1 表明如果我们加回那些满足条件 $\|x_{i_j} - \bar{x}_i - Q_i\theta_j\| \le \eta\|\theta_j\|$ 的邻域点 x_{i_j}，那么（4.4）仍成立，只是其中 η 稍微的增大一些。

引理 4.1： 记 $\{\bar{x}^* + Q^*\theta_j^*\}$ 为邻域集 $\mathcal{N} = \{x_1, \ldots, x_k\}$ 的最佳线性拟合。记 $X = [x_1, \ldots, x_k]$，以及 $\Theta^* = [\theta_1^*, \ldots, \theta_k^*]$。那么有

$$\|X - \bar{x}^* \mathbf{1}_k^T - Q^* \Theta^*\|_F = \min_{x, Q, \Theta: Q^T Q = I_d} \|X - x\mathbf{1}_k^T - Q\Theta\|_F, \tag{4.8}$$

$$\|\Theta^*\|_F = \max_{Q: Q^T Q = I_d} \|Q^T (X - \bar{x}^* \mathbf{1}_k^T)\|_F. \tag{4.9}$$

证明： 通过 PCA，性质（4.8）显然成立。由于

$$\|Q^T (X - \bar{x}^* \mathbf{1}_k^T)\|_F^2 = \|X - \bar{x}^* \mathbf{1}_k^T\|_F^2 - \|X - \bar{x}^* \mathbf{1}_k^T - Q\Theta\|_F^2$$

以及 $\Theta^* = (Q^*)^T (X - \bar{x}^* \mathbf{1}_k^T)$，（4.9）直接可由（4.8）得出。 ■

定理 4.1： 记 $\mathcal{N}_i = \{x_{i_1}, \ldots, x_{i_k}\}$ 为满足（4.4）的 x_i 的邻域，且它的最佳线性拟合为 $\{\bar{x}_i + Q_i \theta_j^{(i)}\}$。假设我们通过增加其它 p 个邻域点扩张 $\mathcal{N}_i$，且每个邻域点满足

$$\|x_{i_j} - \bar{x}_i - Q_i \theta_j^{(i)}\| \le \eta \|\theta_j^{(i)}\|, \theta_j^{(i)} = Q_i^T (x_{i_j} - \bar{x}_i), \quad j = k_i + 1, \ldots, k_i + p. \tag{4.10}$$

那么扩张后的邻域集 $\tilde{\mathcal{N}}_i$ 的最佳线性拟合 $\{\tilde{x}_i + \tilde{Q}_i \tilde{\theta}_j^{(i)}\}$ 满足

$$\|\tilde{X}_i - \tilde{x}_i \mathbf{1}_{k_i+p}^T - \tilde{Q}_i \tilde{\Theta}_i\|_F \le \eta \alpha_i \|\tilde{\Theta}_i\|_F, \tag{4.11}$$

其中 $\tilde{\Theta}_i = [\tilde{\theta}_1^{(i)}, \ldots, \tilde{\theta}_{k_i+p}^{(i)}]$ 且

$$\alpha_i = \left(1 + \frac{\|\sum_{j=k_i+1}^{k_i+p} \theta_j^{(i)}\|^2}{(k_i + p)\|\tilde{\Theta}\|^2}\right)^{1/2}. \tag{4.12}$$

证明： 我们用引理 4.1中的 PCA 的最优性质来证明这个定理。出于记号上的方便，我们在证明中记 $k = k_i$。在性质（4.8）中 $\tilde{X}_i$ 的最佳线性拟合 $\tilde{x}_i \mathbf{1}_{k+p}^T + \tilde{Q}_i \tilde{\Theta}_i$，我们设 $x = \bar{x}_i$，$Q = Q_i$ 以及 $\Theta = \hat{\Theta}_i = [\theta_1^{(i)}, \ldots, \theta_{k+p}^{(i)}]$，可以得到

$$\|\tilde{X}_i - \tilde{x}_i \mathbf{1}_{k+p}^T - \tilde{Q}_i \tilde{\Theta}_i\|_F^2 \le \|\tilde{X}_i - \bar{x}_i \mathbf{1}_{k+p}^T - Q_i \hat{\Theta}_i\|_F^2 = \sum_{j=1}^{k+p} \|x_{i_j} - \bar{x}_i - Q_i \theta_j^{(i)}\|^2.$$

由条件（4.4）和（4.10）可得

$$\|\tilde{X}_i - \tilde{x}_i \mathbf{1}_{k+p}^T - \tilde{Q}_i \tilde{\Theta}_i\|_F^2 \le \eta^2 \|[\theta_1^{(i)}, \ldots, \theta_{k+p}^{(i)}]\|_F^2. \tag{4.13}$$

现在我们用 $\|\tilde{\Theta}_i\|_F^2 = \|[\tilde{\theta}_1^{(i)}, \ldots, \tilde{\theta}_{k+p}^{(i)}]\|_F^2$ 给出 $\|[\theta_1^{(i)}, \ldots, \theta_{k+p}^{(i)}]\|_F^2$ 的界。由于 $Q_i^T (\tilde{X}_i - \tilde{x}_i \mathbf{1}_{k+p}^T)$ 和 $Q_i^T (\tilde{x}_i - \bar{x}_i) \mathbf{1}_{k+p}^T$ 行正交，易得

$$\|[\theta_1^{(i)}, \ldots, \theta_{k+p}^{(i)}]\|_F^2 = \|Q_i^T (\tilde{X}_i - \bar{x}_i \mathbf{1}_{k+p}^T)\|_F^2$$

$$\begin{aligned} &= \|Q_i^T(\tilde{X}_i - \tilde{x}_i 1_{k+p}^T) + Q_i^T(\tilde{x}_i - \bar{x}_i)1_{k+p}^T\|_F^2 \\ &= \|Q_i^T(\tilde{X}_i - \tilde{x}_i 1_{k+p}^T)\|_F^2 + (k+p)\|Q_i^T(\tilde{x}_i - \bar{x}_i)\|^2. \end{aligned}$$

利用性质（4.9），有 $\|Q_i^T(\tilde{X}_i - \tilde{x}_i 1_{k+p}^T)\|_F \le \|\tilde{\Theta}_i\|_F$，从而有

$$\|[\theta_1^{(i)}, \ldots, \theta_{k+p}^{(i)}]\|_F^2 \le \|\tilde{\Theta}_i\|_F^2 + (k+p)\|Q_i^T(\tilde{x}_i - \bar{x}_i)\|^2. \tag{4.14}$$

注意到

$$\tilde{x}_i - \bar{x}_i = \frac{1}{k+p}(k\bar{x}_i + \sum_{j=k+1}^{k+p} x_{i_j}) - \bar{x}_i = \frac{1}{k+p}\sum_{j=k+1}^{k+p}(x_{i_j} - \bar{x}_i),$$

我们可以得到

$$Q_i^T(\tilde{x}_i - \bar{x}_i) = \frac{1}{k+p}\sum_{j=k+1}^{k+p} Q_i^T(x_{i_j} - \bar{x}_i) = \frac{1}{k+p}\sum_{j=k+1}^{k+p}\theta_j^{(i)}.$$

记 $\hat{\theta}_i = \sum_{j=k+1}^{k+p}\theta_j^{(i)}$ 并且将它和（4.14）代入（4.13），我们可以得到

$$\|\tilde{X}_i - \tilde{x}_i 1_{k+p}^T - \tilde{Q}_i\tilde{\Theta}_i\|_F^2 \le \eta^2\Big(\|\tilde{\Theta}_i\|_F^2 + \frac{\|\hat{\theta}_i\|^2}{k+p}\Big) = \eta^2\Big(1 + \frac{\|\hat{\theta}_i\|^2}{(k+p)\|\tilde{\Theta}_i\|_F^2}\Big)\|\tilde{\Theta}_i\|_F^2.$$

这样立刻可证得（4.11）成立。 ■

通常情况下，$\|\sum_{j=k_i+1}^{k_i+p}\theta_j^{(i)}\|^2 < \sum_{j=k_i+1}^{k_i+p}\|\theta_j^{(i)}\|^2$。（4.12）中所定义的因子只是稍大于而且很接近 1。定理 4.1表明了扩张后的邻域也满足有着稍大一些的 η 的（4.4）。基于以上的分析，我们给出了下列的邻域扩张算法。

邻域扩张算法

E0. 采用邻域压缩方法计算得到邻域集 $\mathcal{N}_i$ 以及它的最佳线性拟合 $\{\bar{x}_i + Q_i\theta_j^{(i)}\}$。

E1. 计算初始邻域处 $\mathcal{N}_i$ 外的所有邻域点 x_{i_j} 的坐标 $\theta_j^{(i)} = Q_i^T(x_{i_j} - \bar{x}_i)$，$j = k_i+1, \ldots, K_{\max}$。

E2. 将那些满足条件 $\|x_{i_j} - \bar{x}_i - Q_i\theta_j^{(i)}\| \le \eta\|\theta_j^{(i)}\|$ 的邻域点 x_{i_j} 加到邻域集 $\mathcal{N}_i$ 中。

表 4.1: 采用参数为 $k_{\min}=3$，$k_{\max}=30$ 的邻域扩张方法得到的改善结果。

		$\mathrm{dev}(\mathcal{N}_i)$			Radius		
η	N.S.	min	mean	max	min	mean	max
0.1	C	4.3e-5	1.3e-2	6.3e-2	3.2e-2	1.1e-1	1.6e-1
	CE	4.3e-5	1.5e-2	8.3e-2	5.9e-2	1.2e-1	2.0e-1
0.2	C	4.3e-5	1.3e-2	6.3e-2	3.2e-2	1.1e-1	1.6e-1
	CE	4.3e-5	1.5e-2	8.3e-2	5.9e-2	1.2e-1	2.0e-1
0.3	C	4.3e-5	1.7e-2	7.7e-1	3.2e-2	1.1e-1	1.6e-1
	CE	4.3e-5	1.8e-2	7.5e-1	6.1e-2	1.2e-1	2.0e-1
0.4	C	4.3e-5	1.8e-2	7.6e-1	3.2e-2	1.1e-1	1.6e-1
	CE	4.3e-5	1.9e-2	7.6e-1	6.1e-2	1.2e-1	2.0e-1
0.5	C	4.3e-5	2.9e-2	8.5e-1	3.2e-2	1.1e-1	2.0e-1
	CE	4.3e-5	3.8e-2	8.2e-1	6.1e-2	1.3e-1	2.0e-1

例4.4（续）： 我们对采自例 4.4中的流形的样本点加上一些噪音，即

$$x_{i,j}=f(s_i,t_j)+\epsilon_{i,j}$$

其中 $s_i, i=1,\ldots,M$ 和 $t_j,\ j=1,\ldots,M$ 分别是区间 $[-0.7\pi,\ \pi]$ 和 $[0.5,\ 5]$ 上的 M 个等分点，且 $M=30$，ϵ_{ij} 是区间 $[-0.001,\ 0.001]$ 上的一致分布。我们考虑邻域扩张算法对这些样本点的应用。邻域压缩算法在每个样本点都能很好的选出它们的邻域点，即邻域集的局部线性拟合同样本点处的切空间只有小的偏离度；而邻域扩张算法能够在不太丢失切空间逼近精度的情况下扩大这些邻域。在表 4.1中，我们列出最佳线性拟合偏离度的最小值、均值、最大值以及采用不同精度参数 η 的邻域选取算法所得到的邻域半径。这里"C"表示只采用压缩算法而"CE" 表示压缩和扩张算法的组合。这在邻域大小上的改善是非常明显的。

邻域选取策略可以适用于包括 LTSA 在内的所有基于邻域的流形学习方法，在第六章里我们将给出自适应邻域选取策略和几种流形学习方法相结合的实验结果。

§4.3 本章小结

对于我们前面提到的流形学习的算法有着共同的特征：首先构造流形上样本点的局部邻域结构，然后用这些局部邻域结构来将样本点全局的映射到一个低维空间。因此，流形学习的首要步骤是选取合适的邻域。邻域越小可以认为邻域的线性结构越明显，但是邻域之间需要有足够的交叠以保证较远的点之间有足够的联系，这又使得邻域不能过小。从直观上想像，流形上曲率大的样本点的邻域应该小一些，而流形上曲率小的样本

点处的邻域可以大一些。因此，关于邻域选取我们需要考虑的问题是：在加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质？在本章中，我们提出算法来自适应的选取邻域以匹配流形的局部几何性质。首先，我们给出切空间标准来判断是否一个邻域集能匹配流形的局部几何性质，即理想的切空间能被邻域集的最佳线性拟合在一个给定精度内逼近。然后，我们提出邻域压缩算法来决定一个（相对小的）邻域集来满足这个标准。最后，我们给出邻域扩张策略来尽可能的扩大这个集合以加强邻域之间的交叠，同时最终的邻域集仍然满足提出的标准。我们的方法从理论上保证了所选出的邻域在匹配流形的局部几何性质的前提下，尽可能的扩张邻域以加强样本点之间的关联性。自适应邻域选取方法能适用于所有基于邻域的流形学习方法。

第 5 章　自适应局部切空间排列方法

本章要点：

- 理想嵌入的局部误差估计
- LTSA 的修正模型
- 流形曲率的估计
- 自适应局部切空间排列方法

在第二章中我们提到过，有两个共同的因素决定着流形学习算法的效果。一个是当加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质。我们在上一章中提出了自适应邻域选取策略来解决这个问题。另一个是流形上的曲率以及样本点密度的变化，会使得所寻找的局部邻域结构产生偏差。应该如何估计流形曲率的变化同数据集的样本点密度变化的影响，并且减少由这些局部线性结构的偏差而产生的全局低维嵌入的偏差？在本章中，我们会提出一种解决这个问题的思路和方法，这种想法专门为 LTSA 而设计。我们通过将流形的局部曲率引入 LTSA 的极小化模型中来解决这个问题。这种修正的模型能减少 LTSA 在构造全局嵌入时产生的偏差，而且我们相信这样的想法也能使用于其它的流形方法。另外，同自适应的邻域选取策略相结合，我们还提出了自适应局部切空间排列方法。

§ 5.1　自适应减少偏差

除非我们是在处理一个线性流形，否则采用一个线性逼近来获取流形的局部几何结构不可避免会导致局部投影的误差 $\|x_{i_j} - (\bar{x}_i + Q_i\theta_j^{(i)})\|$ 以及相应的局部坐标 $\{\theta_j^{(i)}\}$ 会产生偏差。误差或偏差的大小依赖于曲率的大小；大的曲率会导致大的投影误差或偏差。图 5.1归纳了这种现象：当曲率小的时候，样本点（$*$）到拟合点（$\circ$）的投影距离比较小；当曲率大的时候，这样的投影距离会比较大。另外，当曲率小的时候，拟合坐标之间能反映出曲线的弧长关系；而当曲率大的时候，拟合坐标之间的关系同样本点之间的弧长关系会有较大的偏差，如果用这样的拟合坐标来反映局部邻域结构不可避免的会对最终的嵌入结果带来影响。在上一章中讨论的自适应的邻域选取方法能够使得在精度参数 η 设置的比较小的时候，局部线性拟合产生的投影误差和局部坐标的偏差相对的比较小。可是，为了维持邻近的邻域之间所必须的交叠，η 并不能设置的过小，而且在邻域压缩步骤中也采用了邻域大小的极小值 $k_{\min}$。所以对于一个有着不同曲率的流形，仍然有必要处理这种投影误差和坐标偏差。

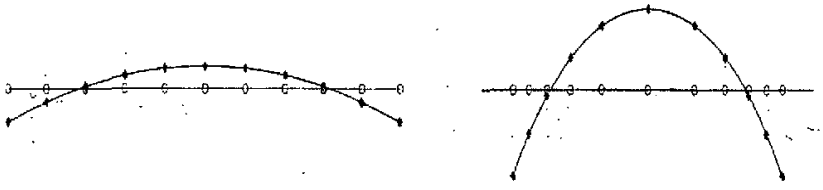

图 5.1: 等分样本点（$*$）的线性拟合的坐标（o）产生的偏差：当曲率小时偏差是不明显的（左）；而当曲率大的时候偏差则相对的比较大（右）。

这章的重点在于通过修正 LTSA 的极小化模型来自适应的减小嵌入坐标 T 的偏差。LTSA 通过求解下面的最优化问题来得到低维嵌入 T：

$$\min_{T} \sum_{i=1}^{N} \frac{1}{k_i} \min_{c_i, L_i} \|T_i - (c_i \mathbf{1}_{k_i}^T + L_i \Theta_i)\|_F^2,$$

其中局部坐标矩阵 $\Theta_i = [\theta_1^{(i)}, \ldots, \theta_{k_i}^{(i)}]$ 可以通过对 x_i 的邻域矩阵 X_i 做 PCA 来得到。注意到这个最优化问题同第二章介绍过 LTSA 的全局最优化问题

$$\min_{T} \sum_{i=1}^{N} \min_{c_i, L_i} \|T_i - (c_i \mathbf{1}_{k_i}^T + L_i \Theta_i)\|_F^2$$

有一点不同。这是因为在第二章里，我们假设 LTSA 采用 k-邻域选取策略，而实际上每个邻域样本点个数可以是不同的。事实上如果假设 $k_i = k$，则很明显两个最优化问题是等价的。注意到

$$\min_{c_i, L_i} \|T_i - (c_i \mathbf{1}_{k_i}^T + L_i \Theta_i)\|_F^2 = \sum_{j=1}^{k_i} \|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|^2, \tag{5.1}$$

其中 $L_i^* = (T_i - \bar{\tau}_i \mathbf{1}_{k_i}^T)\Theta_i^+$。这里 $\|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|$ 定义了一个度量局部坐标 $\theta_j^{(i)}$ 和参数向量 τ_{i_j} 之间误差的测度。LTSA 平等的对待每个局部误差项：

$$\min_{T} \sum_{i=1}^{N} \frac{1}{k_i} \sum_{j=1}^{k_i} \|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|^2. \tag{5.2}$$

这也许能使得总体的误差比较小，但是嵌入结果 T 也可能会有大的偏差。为了归纳这种现象，我们给出了下面的例子。

例 5.1： 记 $\mathcal{M} = \{(t, s) : s = g(t)\}$ 为平面上的曲线，

$$g(t) = \frac{1}{5}(t - 1.5)(t - 4)(t - 4.5)\sin(2t), \quad t \in [0, 2\pi].$$

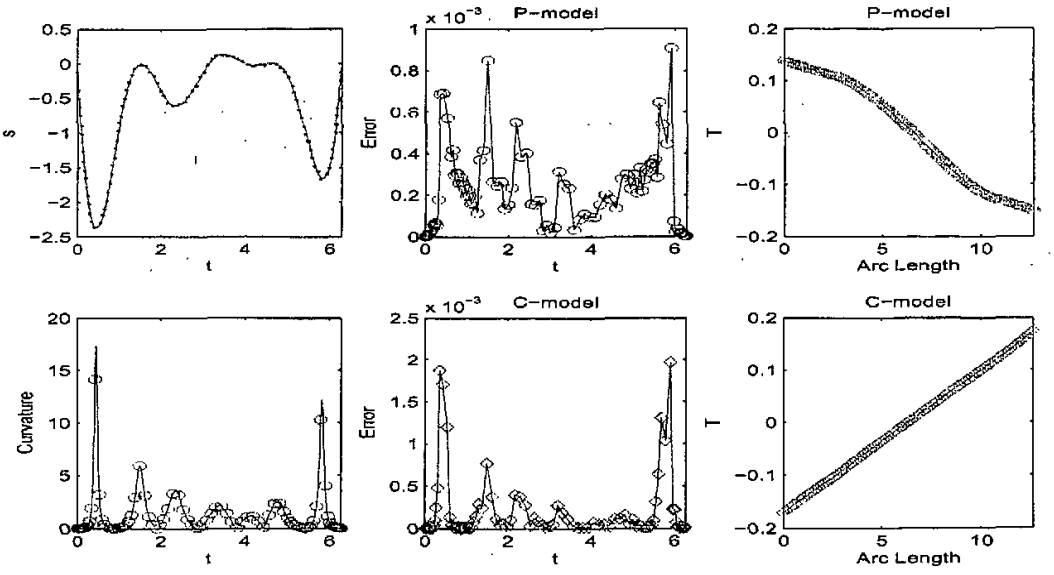

图 5.2: 顶行（从左到右）：数据点，小的局部误差以及（5.2）的最优解得到的坏的嵌入结果。底行：曲率，同曲率一致的局部误差以及修正 LTSA 得到的好的嵌入结果。

我们在曲线上的两点（$t_1 = 0, t_N = 2\pi$）之间等弧长的收集了 $N = 100$ 个样本点 $\{(t_i, g(t_i))\}$，并且设

$$x_i = [t_i, g(t_i)]^T + \epsilon_i, \quad i = 1, \ldots, N,$$

其中 ϵ_i 一致分布在区间 $[-0.02, 0.02]$ 中。曲线和样本点画在图 5.2顶行的左边。我们用 $k = 5$ 的 k-邻域策略来选取邻域。极小化误差项（5.2）的嵌入坐标 T 同弧长坐标有着很大的偏差。在图 5.2 的顶行的右边，我们画出了计算的嵌入结果 T 同弧长的对比。

我们关于减少嵌入偏差的一个主要的观察所得是好的低维嵌入 T 应该有着同局部曲率相匹配的局部误差 $\|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|$。为了验证这种观点，我们在图 5.2的底行的中间画出了一个低维嵌入的误差项 $\sum_{j=1}^{k_i} \|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|^2$。这个低维嵌入是用我们关于 LTSA 的新模型计算所得，同（5.2）的最优解相比，这个低维嵌入有着更大的局部误差。可是，它同弧长相比，有着非常小的偏差，具体见图 5.2底行的右边。注意到此时它的局部误差项同画在图底行左边的局部曲率相一致。

在下一节中，我们将对采自参数化流形 $\mathcal{M} = f(\Omega)$（f 是光滑映射）的数据点 $\{x_i = f(\tau_i)\}$ 的理想嵌入 $\{\tau_i\}$ 的局部误差项 $\|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|$ 给出一个界，并进行相应的分析。它将引导出我们关于 LTSA 的修正模型。

§5.2 理想嵌入的局部误差估计

我们假设 $\mathcal{M} = f(\Omega)$ 是一个参数化的流形，其中 $f : \Omega \subset \mathcal{R}^d \to \mathcal{M} \subset \mathcal{R}^n$ 是一个光滑而且局部等距的映射。对于有限个样本点 $x_i = f(\tau_i), i = 1, \ldots, N$，$\{\tau_1, \ldots, \tau_N\}$ 是一个理想的低维嵌入。我们能推导出它对应邻域 $\mathcal{N}_i$ 的局部误差 $\|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|$ 的估计量。我们先进行如下的二阶泰勒展开

$$\hat{x} - x = J_\tau(\hat{\tau} - \tau) + \frac{1}{2} H_\tau(\hat{\tau} - \bar{\tau}, \hat{\tau} - \tau) + o(\|\hat{\tau} - \tau\|^2). \tag{5.3}$$

这里 H_τ 是一个有两个变量的对称双线性向量值函数

$$H_\tau(u, v) = \sum_{\alpha,\beta=1}^{d} u(\alpha) v(\beta) \frac{\partial^2 f}{\partial \tau(\alpha) \partial \tau(\beta)},$$

其中 $\tau(\alpha), \tau(\beta), u(\alpha), v(\beta)$ 分别表示 τ, u, v 的第 α, β 个分量。假设我们有一个线性子空间 $\mathrm{span}(Q_\tau)$ 逼近切空间 $\mathrm{span}(J_\tau)$。记 $\theta_x(\hat{x}) = Q_\tau^T(\hat{x} - x)$，$P_\tau = Q_\tau^T J_\tau$，$w = H_\tau(\hat{\tau} - \tau, \hat{\tau} - \tau)$，然后在等式（5.3）的左右两边都左乘上 $P_\tau^{-1} Q_\tau^T$，我们可以得到

$$\begin{aligned} P_\tau^{-1}\theta_x(\hat{x}) - \hat{\tau} - \tau &= P_\tau^{-1} Q_\tau^T(\hat{x} - x) - P_\tau^{-1} Q_\tau^T J_\tau(\hat{\tau} - \tau) \\ &= \frac{1}{2} P_\tau^{-1} Q_\tau^T w + o(\|\hat{\tau} - \tau\|^2), \end{aligned}$$

从而有

$$\|\hat{\tau} - \tau - P_\tau^{-1}\theta_x(\hat{x})\| \leq \frac{1}{2}\|P_\tau^{-1}\|\|Q_\tau w\| + o(\|\hat{\tau} - \tau\|^2). \tag{5.4}$$

接下来，我们将给出 $\|P_\tau^{-1}\|$ 和 $\|Q_\tau w\|$ 的估计量，在此之前我们先给出定理 5.1。

定理 5.1：*如果 f 是局部等距的，那么对任何向量 u、v和w，点积*

$$H_\tau(u, v) \cdot J_\tau w = 0.$$

证明：用 $D(u, v, w)$ 表示向量 $H_\tau(u, v)$ 和 $J_\tau w$ 的点积，即 $D(u, v, w) = H_\tau(u, v) \cdot J_\tau w$。不失一般性，我们可以假设 $\|u\| = \|v\| = \|w\| = 1$。

注意到 $J_\tau u = \sum_{\alpha=1}^{d} u(\alpha) \frac{\partial f}{\partial \tau(\alpha)} = \frac{\partial}{\partial u} f(\tau)$，为 f 在 τ 沿着向量 u 的方向导数。$H_\tau(u, v)$ 能用下面的形式简单的表示，

$$H_\tau(u, v) = \frac{\partial^2}{\partial u \partial v} f(\tau) = \frac{\partial}{\partial v}(J_\tau u) = \frac{\partial J_\tau}{\partial v} u.$$

这就证明了

$$D(u, v, w) = \frac{\partial J_\tau}{\partial v} u \cdot J_\tau w.$$

在另一方面，由等距性可知，J_τ 是正交的，即 $J_\tau^T J_\tau = I$。所以 $J_\tau u \cdot J_\tau w = u \cdot w$。对等式的两边关于 v 作偏导可以得到

$$\frac{\partial J_\tau}{\partial v} u \cdot J_\tau w + J_\tau u \cdot \frac{\partial J_\tau}{\partial v} w = 0.$$

从而

$$D(u,v,w) = -J_\tau u \cdot \frac{\partial J_\tau}{\partial v} w = -D(w,v,u),$$

即 $D(u,v,w)$ 关于 u 和 w 是反对称的。由 $H_\tau(u,v)$ 关于 u,v 的对称性可知 $D(u,v,w)$ 关于 u 和 v 是对称的，我们可以推出 $D(u,v,w)$ 关于 v 和 w 也是反对称的。因此

$$D(u,v,w) = -D(w,v,u) = D(w,u,v) = D(u,w,v) = -D(u,v,w),$$

从而有 $D(u,v,w) = 0$。 ■

现在，我们来给出因子 $\|P_\tau^{-1}\|$ 和 $\|Q_\tau w\|$ 的估计量。首先有 $\|P_\tau^{-1}\| \leq 1/\sigma_{\min}(P_\tau)$，其中 $\sigma_{\min}(P_\tau)$ 表示矩阵 P_τ 的最小奇异值。其次，利用定理 5.1所证明的正交性，我们有 $J_\tau^T w = 0$。所以，

$$\|Q_\tau^T w\| = \|(Q_\tau Q_\tau^T - J_\tau J_\tau^T) w\| \leq \|Q_\tau Q_\tau^T - J_\tau J_\tau^T\| \|w\|.$$

由（4.6）有

$$\|Q_\tau Q_\tau^T - J_\tau J_\tau^T\| = \sqrt{1 - \sigma_{min}^2(Q_\tau^T J_\tau)} = \sqrt{1 - \sigma_{min}^2(P_\tau)},$$

从而有

$$\|Q_\tau^T w\| \leq \sqrt{1 - \sigma_{min}^2(P_\tau)} \|w\|.$$

由 H_τ 的定义我们有 $\|w\| \leq c_\tau \|\hat{\tau} - \tau\|^2$，其中 c_τ 是在 $x = f(\tau)$ 处的最大主方向曲率。这样，由（5.4）我们可以给出 $\|\hat{\tau} - \tau - P_\tau^{-1}\theta_x(\hat{x})\|$ 的一个上界估计

$$\|\hat{\tau} - \tau - P_\tau^{-1}\theta_x(\hat{x})\| \leq c_\tau \|\hat{\tau} - \tau\|^2 + o(\|\hat{\tau} - \tau\|^2), \tag{5.5}$$

其中 $\gamma_\tau = \frac{\sqrt{1-\sigma_{\min}^2(P_\tau)}}{2\sigma_{\min}(P_\tau)}$。

对于有限的样本点，我们设 $\hat{\tau} = \tau_{i_j}$、$\tau = \bar{\tau}_i$、$Q_\tau = Q_i$，那么 $\theta_{x_i}(x_{i_j}) = \theta_j^{(i)}$ 以及由（5.5）得到

$$\|\tau_{i_j} - \bar{\tau}_i - P_{\tau_i}^{-1}\theta_j^{(i)}\| \leq \gamma_{\tau_i} c_{\tau_i} \|\tau_{i_j} - \bar{\tau}_i\|^2 + o(\|\tau_{i_j} - \bar{\tau}_i\|^2).$$

我们由（5.1）的最优性可以得出

$$\|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\| \leq \gamma_{\tau_i} c_{\tau_i} \|\tau_{i_j} - \bar{\tau}_i\|^2 + o(\|\tau_{i_j} - \bar{\tau}_i\|^2). \tag{5.6}$$

§5.3 LTSA 的修正模型

记 $\phi_j^{(i)} \equiv \gamma_{\tau_i} c_{\tau_i} \|\theta_j^{(i)}\|^2$（在下一节中，我们将说明如何决定 $\{\phi_j^{(i)}\}$），由上面的分析我们可以得到如下的修正的最小问题

$$\begin{aligned}&\min_T \sum_i \frac{1}{k_i} \sum_{j=1}^{k_i} \left(\frac{\|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|}{\phi_j^{(i)}} \right)^2 \\ =&\min_T \sum_i \frac{1}{k_i} \|T_i (I - \frac{1}{k_i} \mathbf{1}_{k_i} \mathbf{1}_{k_i}^T)(I - \Theta_i^+ \Theta_i) D_i^{-1}\|_F^2, \end{aligned} \tag{5.7}$$

其中 $\bar{\tau}_i = \sum_{j=1}^{k_i} \tau_{i_j}/k_i = T_i \mathbf{1}_{k_i}/k_i$，$\Theta_i = [\theta_1^{(i)}, \ldots, \theta_{k_i}^{(i)}]$，$L_i^* = (T_i - \tau_i \mathbf{1}_{k_i}^T)\Theta_i^+$，$D_i = \mathrm{diag}(\phi_1^{(i)}, \ldots, \phi_{k_i}^{(i)})$。同（2.10）类似，最优化问题（5.7）也能用下面的特征值问题表示，

$$\min_T \sum_i \frac{1}{k_i} \|T_i (I - \frac{1}{k_i} \mathbf{1}_{k_i} \mathbf{1}_{k_i}^T)(I - \Theta_i^+ \Theta_i) D_i^{-1}\|_F^2 = \min_T \mathrm{trace}(T \Phi T^T) \tag{5.8}$$

其中 $\Phi = \sum_i \frac{1}{k_i} (S_i W_i D_i^{-1})(S_i W_i D_i^{-1})^T$，$S_i \in R^{N \times k_i}$ 是满足 $[T_1, \ldots, T_N] S_i = [T_{i_1}, \ldots, T_{i_{k_i}}]$ 的选择矩阵，且

$$W_i = (I - \frac{1}{k_i} \mathbf{1}_{k_i} \mathbf{1}_{k_i}^T)(I - \Theta_i^+ \Theta_i).$$

再加上正则化条件 $TT^T = I$，（5.8）的唯一解可以由矩阵 Φ 的从第 2 到第 $(d+1)$ 小的特征值所对应的 d 个特征向量所组成的矩阵得到。

§5.4 估计曲率和 $\{\phi_j^{(i)}\}$

现在我们考虑如何决定 $\phi_j^{(i)}$，同时也考虑噪音对估计 $\phi_j^{(i)}$ 的影响。估计 $\phi_j^{(i)}$ 的关键步骤在于估计曲率 c_{τ_i}，这里我们估计的是 x_i 处的平均曲率 $\bar{c}_{\tau_i}$ 而不是最大曲率 c_{τ_i}。这是因为如果采用最大曲率，关于 $\|w\|$ 的界 $\|w\| \le \gamma_\tau \|\hat{\tau} - \tau\|^2$ 也许会太过于宽松，而平均曲率的计算能更加的稳定，对有噪音的数据集更是如此。最后，我们要说明怎样计算在样本点 x_i 处的方向曲率 $c_{\tau_i}(\tau_{i_j})$。$c_{\tau_i}(\tau_{i_j})$ 定义了 x_i 处一条通过 x_i 和 x_{i_j} 的测地线的曲率。

由曲线曲率的定义可知，曲线的曲率是切空间的夹角相对于弧长的变化。由于 $\sigma_{\min}(J_{\tau_i}^T J_{\tau_{i_j}})$ 定义了点 $x_i = f(\tau_i)$ 和点 $x_{i_j} = f(\tau_{i_j})$ 处两个切空间的最大夹角的 cosine 值 [28, §12.4.3]，这样切空间夹角的变化就可以由 $\arccos\left(\sigma_{\min}(J_{\tau_i}^T J_{\tau_{i_j}})\right)$ 表示。因此连接 x_i 和 x_{i_j} 的测地线的曲率 $c_{\tau_i}(\tau_{i_j})$ 能近似的估计为

$$c_{\tau_i}(\tau_{i_j}) \approx \frac{\arccos\left(\sigma_{\min}(J_{\tau_i}^T J_{\tau_{i_j}})\right)}{\|\tau_{i_j} - \tau_i\|}. \tag{5.9}$$

由于每个同曲率匹配的邻域集的局部最佳线性拟合 $\bar{x}_i + Q_i\theta_j^{(i)}$，$j = 1, \ldots, k_i$ 以及线性空间 $\mathrm{span}(Q_i)$ 是对样本点 x_i 处的切空间的很好的逼近，那么有 $\|\tau_{i_j} - \tau_i\| \approx \|\theta_j^{(i)}\|$。这样，夹角 $\arccos\left(\sigma_{\min}(J_{\tau_i}^T J_{\tau_{i_j}})\right)$ 和 $\|\tau_{i_j} - \tau_i\|$ 能分别被很容易计算的 $\arccos\left(\sigma_{\min}(Q_{i_j}^T Q_i)\right)$ 和 $\|\theta_j^{(i)}\|$ 所逼近。当 $\|\theta_j^{(i)}\|$ 不是太小的时候，就能得到 $c_{\tau_i}(\tau_{i_j})$ 的一个近似

$$c_j^{(i)} = \frac{\arccos\left(\sigma_{\min}(Q_{i_j}^T Q_i)\right)}{\|\theta_j^{(i)}\|}.$$

由于流形可能会存在噪音，对于离样本点 x_i 太近的邻域点 x_{i_j}，我们用上面给出的近似值来估计 x_i 和 x_{i_j} 的方向曲率可能会不准确。因此，在计算平均曲率的时候，我们先给定一个阈值 $\delta_c > 0$，并忽略那些满足 $\|\theta_j^{(i)}\| \le \delta_c \max_{j \le k_i} \|\theta_j^{(i)}\|$ 的方向曲率 $c_j^{(i)}$。实验中，我们设 $\delta_c = 0.1$。这样，我们可以用那些未被忽略的 $c_j^{(i)}$ 来估计平均曲率，即

$$\bar{c}_i = \text{mean of } \left\{ c_j^{(i)} \;\middle|\; x_{i_j} \in \mathcal{N}_i,\ \|\theta_j^{(i)}\| > \delta_c \max_{j \le k_i} \|\theta_j^{(i)}\| \right\}. \tag{5.10}$$

这样可以得到 $\phi_j^{(i)}$ 的一个估计，$\phi_j^{(i)} \approx \bar{c}_i \|\theta_j^{(i)}\|^2$。

可是，由于数据集中噪音的存在以及局部线性拟合时的误差，一个很小的 $\phi_j^{(i)} = \bar{c}_i \|\theta_j^{(i)}\|^2$ 会导致在（5.7）中的 $\frac{\|\tau_{i_j} - \bar{\tau}_i - L_i^* \theta_j^{(i)}\|}{\phi_j^{(i)}}$ 有很大的误差。在这种情况下，最小化（5.7）也许会得到一个有很大误差的嵌入结果。所以为了保持数值计算的稳定性，有必要加入ϕ 的一个小的阈子 δ_ϕ。我们设

$$\phi_j^{(i)} = \delta_\phi + \bar{c}_i \|\theta_j^{(i)}\|^2. \tag{5.11}$$

例 5.1（续）： 在图 5.2的底行，我们画出了曲率vs.弧长（左边）和用曲率模型（5.7）得到的嵌入结果的局部误差（中间），其中模型（5.7）中的 $\{\phi_j^{(i)}\}$ 由（5.11）定义且参数 $\delta_c = 0.1, \delta_\phi = 10^{-4}$。值得注意的是，计算的嵌入 T 的局部误差能很好的同曲率相匹配，并且同理想的嵌入坐标（弧长）对比几乎没有什么偏差。关于 δ_c 和 δ_ϕ 的选择对我们的实验结果并没有什么影响。

§ 5.5 自适应局部切空间排列

很显然，曲率的估计依赖于邻域的选取，而上一章我们提出的自适应邻选取策略能将选出的邻域同流行的曲率相匹配。因此，自适应邻域选取策略（压缩和扩张）能够和引入曲率的减少偏差策略一起使用。我们用这些策略来改善局部切空间排列（LTSA）的嵌入结果，并称之为自适应局部切空间排列方法（Adaptive Local Tangent Space Alignment）。我们将这种算法归纳如下。

自适应局部切空间排列 (Adaptive Local Tangent Space Alignment)

1. **自适应邻域选取.** 采用邻域压缩/扩张方法来决定每个样本点 x_i 的邻域 $\mathcal{N}_i = \{x_{i_1}, \ldots, x_{i_{k_i}}\}$。
2. **局部线性投影.** 对每个邻域集 $\mathcal{N}_i$，计算 $X_i - \bar{x}_i \mathbf{1}_{k_i}^T$ 对应它的 d 个最大奇异值的奇异值分解 $Q_i \Sigma_i V_i^T$，并且设 $\theta_j^{(i)} = Q_i^T(x_{i_j} - \bar{x}_i)$。
3. **曲率的估计.** 利用（5.9）和（5.10）估计曲率，并用（5.11）计算 $\phi_j^{(i)}$。
4. **局部坐标的排列.** 计算排列阵 $\Phi = \sum_{i=1}^{N} \frac{1}{k_i}(S_i W_i D_i^{-1})(S_i W_i D_i^{-1})^T$ 的 $d+1$ 个最小特征向量，其中 $D_i = \mathrm{diag}(\phi_1^{(i)}, \ldots, \phi_{k_i}^{(i)})$ 以及 $W_i = I_{k_i} - \left[\frac{1}{\sqrt{k_i}}\mathbf{1}_{k_i}, V_i\right]\left[\frac{1}{\sqrt{k_i}}\mathbf{1}_{k_i}, V_i\right]^T$。选出对应从第 2 小到第 $d+1$ 小的特阵值的特阵向量组成矩阵 $[u_2, \ldots, u_{d+1}]$，则整体的嵌入结果就是 $T = [u_2, \ldots, u_{d+1}]^T$。

在下一章中，我们将给出实验例子来说明自适应局部切空间排列算法的效果。

§5.6 本章小结

流形上的曲率以及样本点密度的变化，不可避免的会使得所寻找的局部邻域结构产生偏差。在利用这些局部邻域结构来构造全局的低维嵌入时，需要将这些偏差计入考虑范围。因此，流形学习面临着这些问题：如何估计流形上的曲率？如何估计流形曲率和样本点密度的变化对寻找局部邻域结构的影响？在利用局部邻域结构来构造全局的低维嵌入时，应该如何计入这种影响以减少低维嵌入的偏差？在本章中，我们给出估计流形局部曲率的方法，并通过引入流形的局部曲率来修正 LTSA 中的极小化模型。这个改善能减少 LTSA 造构造全局嵌入的偏差。虽然曲率模型是针对 LTSA 而设计，但我们相信所提出的基本思想也能适用于其它的流形学习方法。将曲率模型同自适应的邻域选取方法结合，我们提出了一种自适应的流形学习方法－自适应局部切空间排列方法（ALTSA）。虽然给出的理论分析是关于理想情形下光滑和局部等距的流形，我们所给出的修正的 LTSA 的曲率模型也可以适用于有噪音的数据集，在第六章中，我们将给出相应的数值实验。

第 6 章　数值实验

在本章中，我们将给出一些实验例子来说明 MLLE 、自适应邻域选取策略和自适应局部切空间排列算法的效果。我们将这些算法同其它的一些常用的流形学习方法进行比较，实验数据包括模拟数据和现实世界中的实际数据。

首先，我们先给出两个例子来说明 MLLE 的效果。

例 6.1：　我们将 MLLE 和 LLE 应用到采自无噪音的 S-曲面的样本点上，样本点的生成方法如下：

$$\begin{aligned}&\mathrm{t} = (3 * \mathrm{rand}(1, \mathrm{N}) - 1) * \mathrm{pi};\\&\mathrm{s} = 5 * \mathrm{rand}(1, \mathrm{N});\\&\mathrm{X} = [\cos(\mathrm{t}); \mathrm{s}; (\sin(\mathrm{t}) - 1). * \mathrm{sign}(\mathrm{pi}/2 - \mathrm{t})];\end{aligned}$$

其中 $N = 2000$。数据点和 2 维的生成坐标画在图 6.1的第一列上。我们可以看出，LLE 的嵌入结果中的变形（拉伸和压缩）相当明显，这在邻域大小 k 小的时候尤其明显。这是因为当邻域比较小的时候，邻域之间的交叠不够充分，样本点之间的关联也比较弱。采用 MLLE 算法可以很明显的改善嵌入结果。在图 6.1的右三列中，我们画出了 LLE 以及采用 MLLE 的嵌入结果。测试中采用不同的邻域大小 $k = 8, 12, 16$。值得注意的是 MLLE 对不同的邻域大小 k 的结果都很稳定。

例 6.2：　同样的，MLLE 对于采自带“空洞”的 Swiss-roll 的样本点也能给出很好的嵌入结果。我们以 Swiss-roll 曲面作为例子，曲面函数定义如下

$$f(s,t) = [s\cos(s),\, t,\, s\sin(s)]^T. \tag{6.1}$$

Swiss-roll 的例子在 [56]中也被用过。注意到这个流形并不等距于上面所指出的生成参数 (s,t) [68]。可是，采用弧长函数

$$r(s) = \int_0^s \sqrt{1+\alpha^2}\, d\alpha = \frac{1}{2}\Big(s\sqrt{1+s^2} + \ln(s+\sqrt{1+s^2})\Big),$$

并且用 $s = s(r)$ 表示 $r = r(s)$ 的逆变换，Swiss-roll 曲面可以重新参数化如下

$$\hat{f}(r,t) = \Big[s(r)\cos\big(s(r)\big),\, t,\, s(r)\sin\big(s(r)\big)\Big]^T$$

并且 $\hat{f}$ 等距于 (r,t)-空间。

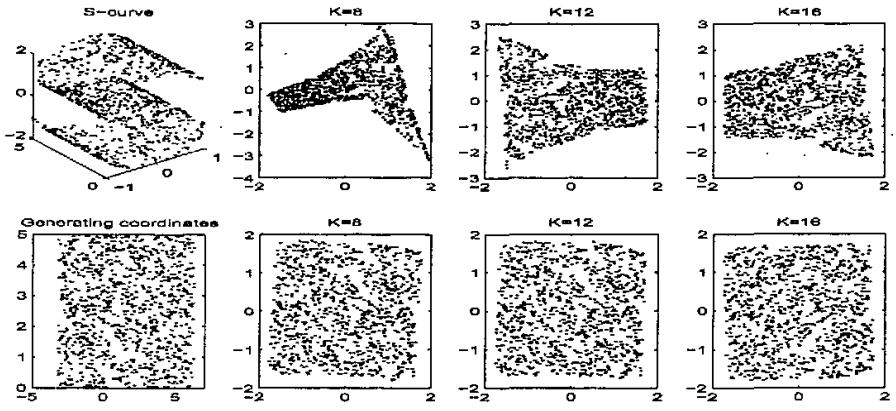

图 6.1: 第一列: S-曲面样本点和生成坐标。右三列:采用 LLE 计算的嵌入结果（顶行）以及 MLLE 计算的嵌入结果（底行），从左到右的邻域大小分别为 $k = 8, 12, 16$。

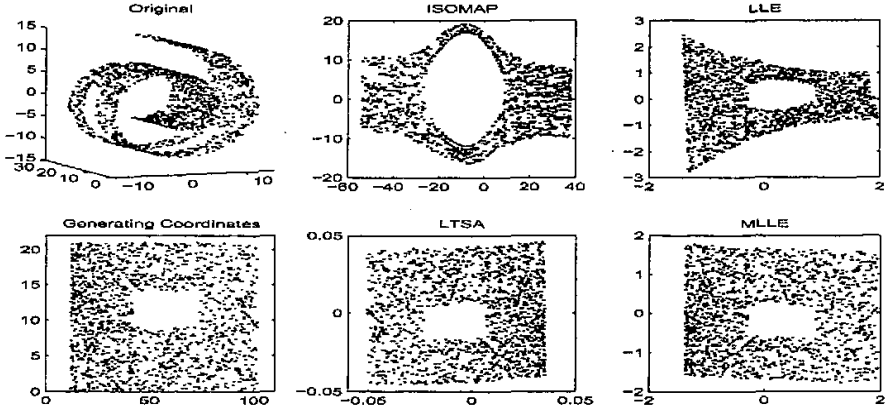

图 6.2: 左列: 带“空洞”的 Swiss-roll 的样本点以及相应的生成坐标。中列: Isomap 和 LTSA 的计算结果。右列: LLE 和 MLLE 的计算结果。

我们生成 (s,t) 区域如下:

$$\mathrm{t} = (3 * \mathrm{pi}/2) * (1 + 2 * \mathrm{rand}(1, \mathrm{N}));$$
$$\mathrm{s} = 21 * \mathrm{rand}(1, \mathrm{N});$$

其中 $N = 2000$。然后，我们挖去 (s,t) 区域中 $[9,14] \times [9,12]$ 的方形区域，并由曲面函数（6.1）生成一个带“空洞”的 Swiss-roll，见图 6.2 第一行第一列。由上面提到的弧长函数，我们可以得到这个带“空洞”的 Swiss-roll 的等距的生成坐标 (r,t)。在图 6.2的第二行的左列，我们画出了生成坐标 (r,t)。值得注意的是，生成这样带“空洞”的 Swiss-roll 的生成坐标的区域并不是凸的。因此，Isomap 并不能恢复出这样的 2 维生成坐标。见图 6.2的第一行的中列，Isomap 的嵌入结果同生成坐标 (r,t) 相比，中间的“空洞”被明显的扩大而且“空洞”上下两方的低维嵌入被挤到一个狭窄的区域。

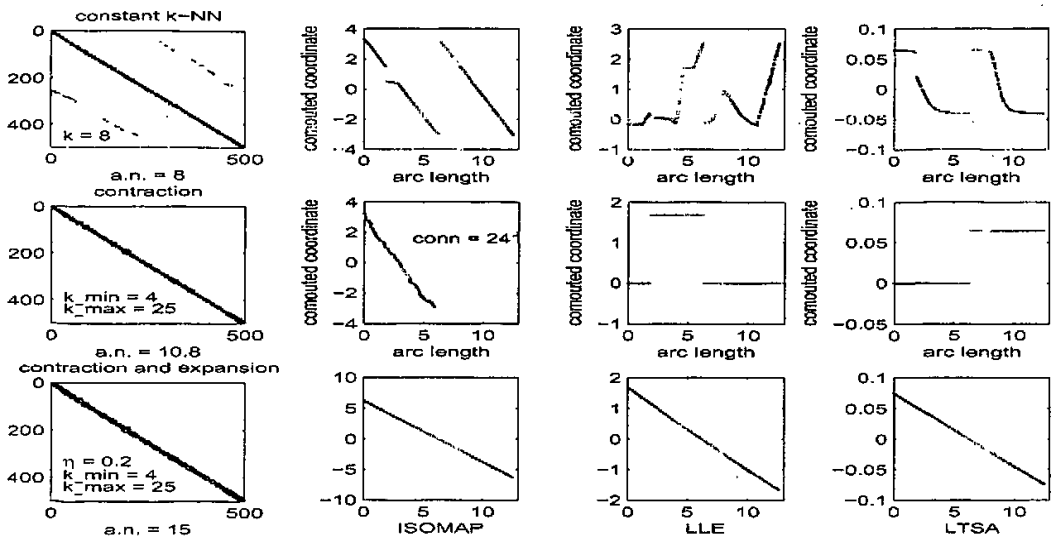

图 6.3: Isomap、LLE和LTSA应用到数据集（6.2）并采用三种邻域选取方法的比较。从上到下：k-邻域选取策略（constant k-NN），仅采用邻域压缩策略（contraction）以及同时采用邻域压缩和扩张策略（contraction and expansion）。从左到右：邻域图（画出的是邻域图的邻接矩阵，且a.n.表示邻域点的平均数量），用 Isomap、LLE和LTSA计算的嵌入结果 vs. 弧长

而对于 LLE，无论怎样的邻域大小 k，LLE 的嵌入结果中都有着很明显的变形。见图6.2的第一行的右列，LLE 的嵌入结果从左到右的压缩变形越来越明显。而对于 LTSA 和 MLLE，从图 6.2第二行的第二列和第三列中，我们可以看出它们有着同样理想的嵌入结果。

接下来，我们要举两个例子表明自适应邻域选取策略（邻域压缩和扩张方法）对流形学习方法的改善。

例 6.3: 我们将数据点采自例 4.3，并且增加了一些噪音，即

$$x_i = [\sin(t_i), \cos(t_i), 0.02t_i]^T + \epsilon_i, \quad i = 1, \ldots, N, \tag{6.2}$$

其中 ϵ_i 是分量一致分布在区间 $[-0.01, 0.01]$ 的噪音向量，并且设 $N = 500$。

对于这个例子，采用 k-邻域选取策略的 LTSA 对任何的 k 都不能恢复出弧长参数，即使不增加噪音也是如此。对于 Isomap 和 LLE 也会有同样的结果，这是由于 k-邻域不能保持样本点上的正确的连接结构。在图 6.3 第一行的左边，我们画出了采用 $k = 8$的k-邻域策略得到的邻域的邻接矩阵 G，其中 $G(i, I_i) = 1$ 和 $G(i, j) = 0, j \notin I_i$。其余三幅分别是采用 Isomap、LLE 和 LTSA 得到的嵌入结果。很显然，三种算法都不能

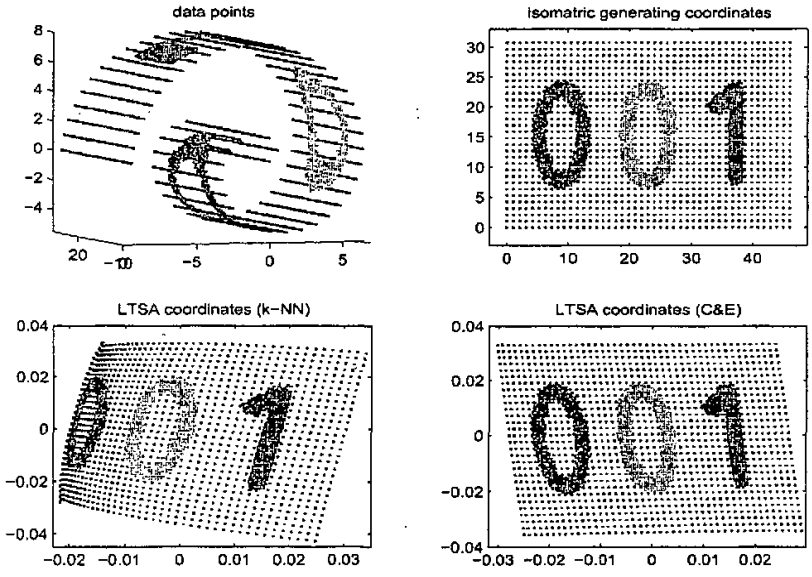

图 6.4: 对有着不同密度的3维 Swiss-roll 的样本点采用自适应邻域选取策略的效率。顶行:数据点和等距的生成坐标 (r_i, t_i)（右边）。底行: 采用 $k = 15$ 的 k-邻域策略（k-NN）的 LTSA 得到的计算结果（左边）或者采用 $k_{\min} = 5$、$K_{\max} = 30$ 和 $\eta = 0.1$ 为参数的自适应邻域选取（C&E）策略得到的结果（右边）。

得到正确的嵌入结果，这是由于采用 k-邻域策略会使得一些样本点选出错误的邻域点。（实验中， Isomap 、 LLE 和 LTSA 对所有 $k \leq 20$ 的结果都是失败的。）如图像第二行的左边所示，邻域压缩方法可以改善这些邻域的结果。但是，由于一些邻域并没有足够的交叠， Isomap 、 LLE 和 LTSA 仍不能给出可以接受的嵌入结果。（ Isomap 在总共 $N = 500$ 个点中只连接了 241 个点。见图 6.3的第二行第二列。）扩张方法改善了邻域之间的交叠部分。最后一行表明了邻域间交叠部分的改善情况，并且当采用参数为 $k_{\min} = 4$、$K_{\max} = 25$ 以及 $\eta = 0.2$ 的自适应邻域选取策略时，计算的嵌入结果是很理想的。

例 6.4： 我们以样本点密度高度变化的 Swiss-roll 曲面作为例子。在图 6.4的左上，我们画出了采自 3 维 Swiss-roll 曲面的样本点。用符号“001”标出的样本点是流形上高密度的样本点。在图的右上方画出了等距的生成参数向量 $(r_i, t_i)^T$。它的设置方法如下。首先我们计算 $r_{\max} = \int_0^{3\pi} \sqrt{1+\alpha^2}\, d\alpha$，并且在 (r,t)-空间的矩形区域 $[0, r_{\max}] \times [0,\ \frac{2}{3} r_{\max}]$ 上设置一个等分的网格，其中网格大小为 $r_{\max}/44$。这样产生 $N_0 = 45 \times 30 = 1350$ 个点，并且这些点被选作生成参数向量 $(r_i,\ t_i)$。然后我们增加 $N_1 = 2000$ 个点作为两个“0”以及 $N_2 = 400$ 个点作为“1”。这样，样本点的总数为 $N = 3750$。

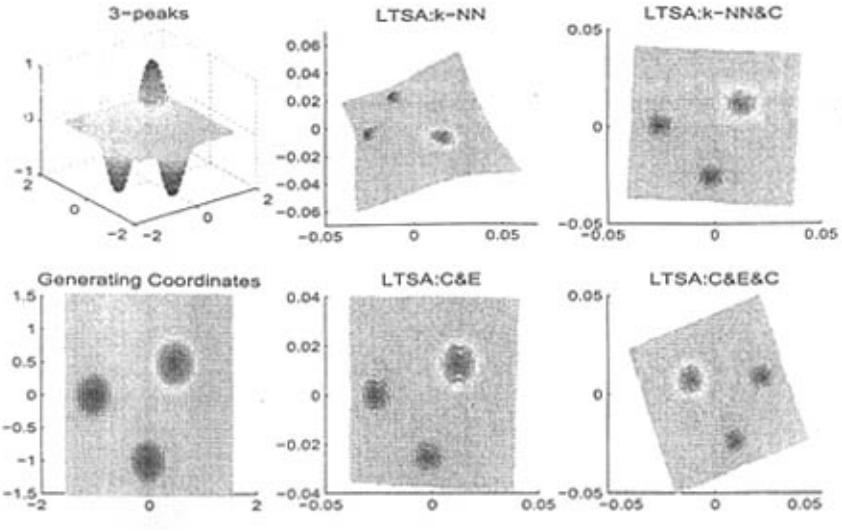

图 6.5: 左列: 三峰流形以及生成坐标。中列: 采用初始极小化模型的 LTSA 的嵌入结果（第一行: 采用k-邻域策略（k-NN）; 第二行: 采用自适应邻域选取策略（C&E））。右列: 采用曲率模型（C）的 LTSA 的嵌入结果（第一行: 采用k-邻域策略（k-NN）; 第二行: 采用自适应邻域选取策略（C&E），即自适应 LTSA）。

采用 k-邻域的 LTSA 能恢复出类似矩形的参数向量网格，只是在沿着符号“0”的边缘有着一些变形，而正是在这些地方有着更高的样本点密度。在图 6.4 的左下方给出了 $k = 15$ 时的结果。通过采用自适应邻域选取策略，这些变形能明显的减少。对比等距坐标 $(r_i, t_i)^T$，采用自适应邻域选取策略的 LTSA 所得到的嵌入结果能很好的恢复出等距坐标，它们之间只相差一个仿射变换。在图的右下方，我们画出了以 $k_{\min} = 5$、$K_{\max} = 30$ 以及 $\eta = 0.1$ 为参数的自适应邻域选取策略的嵌入结果。

现在我们要举一个流形中曲率高度变化的例子来说明自适应 LTSA 中曲率模型的效率和 MLLE 对曲率高度变化的流形的适用性。我们将采用 k-邻域选取策略、自适应邻域选取策略的 LTSA 以及自适应 LTSA 分别应用到一个嵌入到 3 维空间的有三个峰的 2 维流形，并以此说明曲率模型对 LTSA 的作用。我们还往这个三峰流形加上一些噪音，以说明自适应的 LTSA 也能适用于有噪音的数据。我们将 MLLE 也用到这个例子上以说明 MLLE 对曲率高度变化的流形的效率。

例 6.5: 我们生成 $N = 2000$ 个 3 维样本点 $x_i = [t_i, s_i, h(t_i, s_i)]^T$，其中 t_i 和 s_i 一致分布在区间 $[-1.5, 1.5]$ 并且 $h(t, s)$ 定义如下

$$h(t,s) = e^{-10\left((t-0.5)^2+(s-0.5)^2\right)} - e^{-10\left(t^2+(s+1)^2\right)} - e^{-10\left((1+t)^2+s^2\right)}.$$

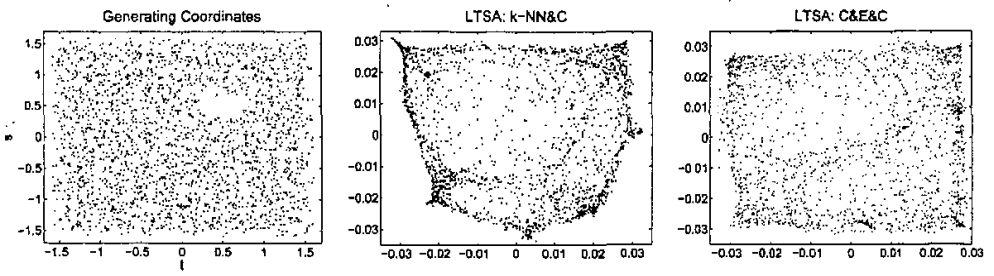

图 6.6: LTSA 的曲率模型对于有噪音的三峰流形所体现的效率。左列：生成坐标；中间：LTSA 采用k-邻域策略（k-NN）和曲率模型（C）的嵌入结果；右列：自适应 LTSA 的嵌入结果。

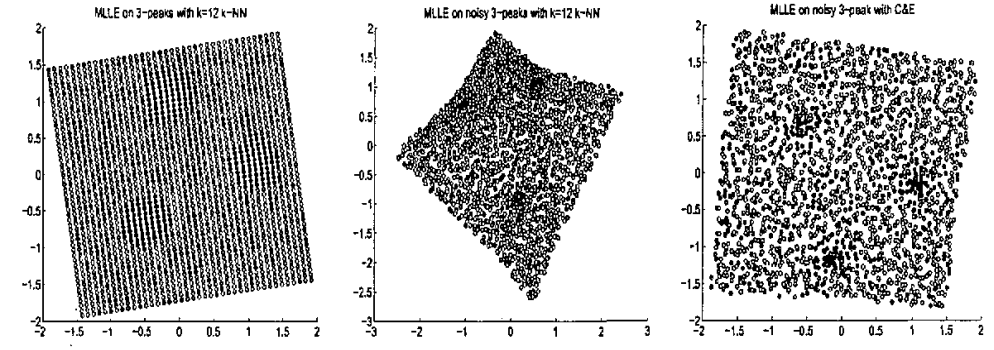

图 6.7: 左列: 对三峰流形 MLLE 采用 $k = 12$ 的 k-邻域策略的嵌入结果。中列：对有噪音的三峰流形， MLLE 采用 $k = 12$ 的 k-邻域策略的嵌入结果。右列: 对有噪音的三峰流形， MLLE 采用自适应邻域选取（C&E）策略得到的结果。

参数化的流形 $f(t,s) = [t, s, h(t,s)]^T$ 并不等距于生成参数空间。可是，由于 f 处的雅科比矩阵 $J_f(t,s)$ 是近似正交的，因而这个映射是近似等距的。为了说明这个，让我们考虑 $J_f(t,s)$ 的两个奇异值。通过简单的计算，可知 $J_f(t,s)$ 的两个奇异值为 $\sigma_1(t,s) = \sqrt{1 + \|\mathrm{grad}\big(h(t,s)\big)\|^2}$ 和 $\sigma_2(t,s) = 1$，其中 grad 表示一个函数的梯度。对于我们在这个例子中所采用的生成参数，最大的奇异值 $\sigma_1(t,s)$ 是 2.98，即雅科比矩阵的条件数的阶数是 $O(1)$。因此，生成 3 峰流形的映射是近似等距的。在图 6.5的左边，我们画出了这个三峰流形和样本点的生成坐标。我们分别采用 $k = 12$ 的 k-邻域策略和以 $k_{\min} = 5, K_{\max} = 30, \eta = 0.1$ 为参数的自适应邻域选取策略，并采用初使 LTSA 的极小化模型得到两组嵌入结果。在图 6.5的中间，我们画出这两组嵌入结果。对于这个例子，同 k-邻域选取策略相比较，用自适应邻域选取策略和初使的 LTSA 极小化模型能改善嵌入结果，但全局嵌入中仍存在有小的偏差。 LTSA 的曲率模型无论采用 k-邻域或自适应邻域选取策略都能很好的减少这种偏差。在图 6.5的右边，我们画出了分别采用k-邻域策

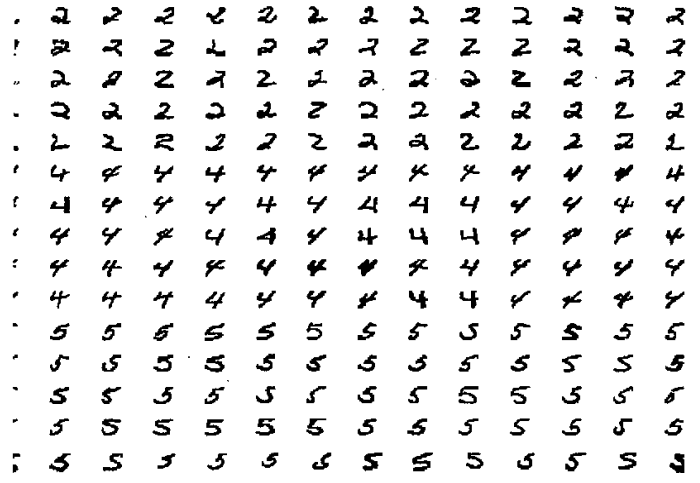

图 6.8: 部分手写数字图像。

略和自适应邻域选取策略，并采用以 $\delta_c = 0.25$ 和 $\delta_\phi = 10^{-6}$ 为参数的曲率模型的 LTSA 得到的嵌入结果。从图中我们可以看出，这两组嵌入结果同生成坐标只相差一个仿射变换，LTSA 的曲率模型对这种曲率高度变化的流形的效果是很明显的。

对于有噪音的数据，在 LTSA 的曲率模型中，自适应邻域选取策略比 k-邻域选取策略有着更好的结果。见图 6.6 中所画出的计算结果，有噪音的样本点为

$$x_i = [t_i,\ s_i,\ h(t_i, s_i)]^T + 0.1\epsilon_i,$$

其中向量 ϵ_i 的分量一致分布在 $[-1, 1]$ 中。这是因为当数据有噪音时，采用 k-邻域选取策略可能会使得曲率模型中对曲率的估计不准确，从而导致嵌入结果产生偏差。而采用自适应邻域选取策略，所选取出的邻域能很好的匹配流形的局部曲率，也能保证曲率模型中曲率的估计具有一定的准确性。因此，对于有噪音的数据，LTSA 的曲率模型更适合同自适应邻域选取策略一起使用。

对于曲率高度变化的流形，MLLE 也能有着很好的适用性。在图 6.7的左边，我们画出了 MLLE 对三峰流形采用 $k = 12$ 的 k-邻域策略的结果。很明显，MLLE 的嵌入结果很好的恢复出了三峰流形的生成坐标。对于有噪音的三峰流形，MLLE 也能得到理想的嵌入结果。在图 6.7的中间，我们画出了对于有噪音的三峰流形 MLLE 采用 $k = 12$ 的 k-邻域策略的嵌入结果。这个嵌入结果在边界上有一些变形，我们可以将自适应邻域选取策略同 MLLE 相结合来改善嵌入结果。见图 6.7的右边，将自适应邻域选取策略同

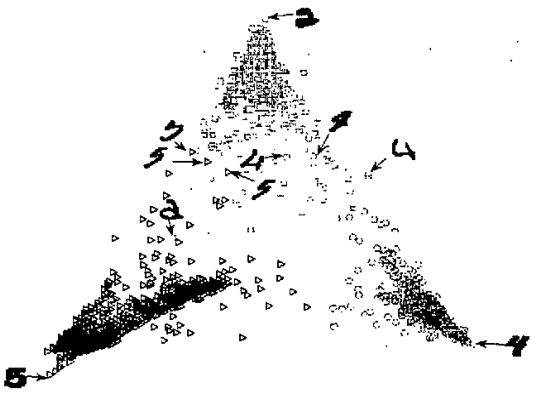

图 6.9: 对手写数字“2”、“4”、“5”采用 $k=15$ 的 k-邻域策略的MLLE得到的 2 维嵌入结果：“□”表示手写数字“2”；“○”表示手写数字“4”；“▷”表示手写数字“5”。

MLLE 结合后，得到的嵌入结果同生成坐标之间只是相差一个仿射变换。

最后我们给出两个实际例子来检验 MLLE 和自适应邻域选取策略的效果。

例 6.6： 考虑这样一组由手写数字图像组成的集合：集合由手写数字图像“2”、“4”、“5”三类组成，其中每一类有 1100 张 16×16 的灰度图像*。在图 6.8中，我们画出了部分的手写数字图像。每张数字图像对应一个 16×16 的像素矩阵，将它转化成为一个 256 维的图像向量后，我们可以得到一个维数为 $m=256$，样本点个数为 $N=3300$ 的数据集。我们将 MLLE 应用到这个数据集。在图 6.9中，我们画出了采用 MLLE 算法得到的 2 维嵌入结果，其中 MLLE 的邻域选取策略为 $k=15$ 的 k-邻域策略，标号“□”、“○”、“▷”分别表示手写数字图像“2”、“4”、“5”在低维空间中对应的嵌入坐标。从图中我们可以看出，手写数字“4”和“5”能被很好的区分开，而个别手写数字“2”和手写数字“4”、“5”混在了一起。这可能是由于这几个手写数字过于潦草而不容易被识别区分。

接下来，我们单独考虑一组手写的数字图像“1”。我们对只由手写数字图像“1”组成的数据集采用邻域选取策略为 $k=8$ 的 k-邻域策略的 MLLE 算法，在图 6.10中，我们画出了它的二维嵌入结果。在图的底行，我们画出了图中路径上“○”点所对应的数

*手写数字图像集可以从 http://www.cs.toronto.edu/ roweis/data.html 下载。

图 6.10: 对手写数字“1”采用 $k = 8$ 的k-邻域策略的MLLE得到的 2 维嵌入结果；底行：路径上“o”点所对应的数字图像。

字图像，它们很明显的反映了手写数字的一种连续变化。在图中我们还画出了“□”点所对应的数字图像，从中我们可以看出，MLLE 的嵌入结果能反映出数字“1”在写法上的一些规律性。

例 6.7： 我们考虑采用 MLLE 和自适应邻域选取策略的 LTSA 在人脸图像数据 [60] 中的应用。数据集由 698 幅 64×64 的像素图像组成，人脸图像由三个隐藏的参数（从左到右和从上到下的位置参数以及亮度参数）所决定。每个参数都是一致分布在某个区间内。每张图像转换成一个 $m = 4096$ 维的图像向量。在图 1.3中，我们画出了部分的人脸图像。我们将 MLLE应用到这个人脸图像的数据集。我们采用 $k = 14$ 的 k-邻域选取策略，在图 6.11的中间，我们画出了 MLLE 所恢复的位置坐标（即三维嵌入结果中的前两个分量）。在这个二维坐标的四条边界上，我们沿着四条路径画出了结点所对应的人脸图像。从图中我们可以看出，MLLE 的嵌入结果中的前两个分量能很好的反映了位置和亮度参数的连续变化。在图 6.12中，我们画出了 MLLE 的三维嵌入结果中的第三个分量同亮度参数的比较。值得注意的是，对这个例子，LLE 完全不能恢复出这三个隐藏的参数，这也说明了 MLLE 对 LLE 的改善。

我们考虑自适应邻域选取策略对 LTSA 的结果的改善。采用 $k = 14$ 的 k-邻域选取策略，LTSA能在一个可以接受的精度范围内恢复出两个位置参数。当采用 k-邻域策略的时候，所计算的嵌入结果同第二个位置参数（上下）相比有一些偏差。我们的自适应邻域选取算法能在某种程度上减少这种偏差。在图 6.13中，我们画出了 $k = 14$ 的 k-邻域策

图 6.11: 采用 $k = 14$ 的 k-邻域策略的MLLE得到的三维嵌入结果中的前两个分量（中间）；以及边界上连线的结点所对应的人脸图像（上、下、左、右）。

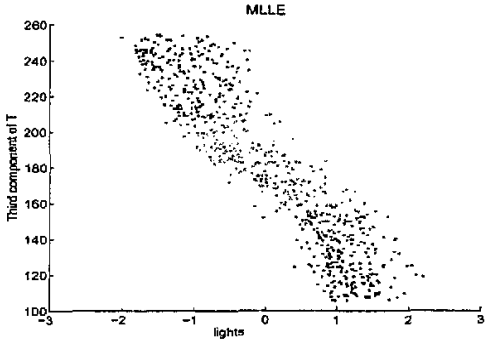

图 6.12: MLLE的嵌入结果中第三个分量同亮度参数的比较。

略得到的嵌入结果（顶行）以及以 $k_{\min} = 6$、$K_{\max} = 20$ 和 $\eta = 0.45$ 为参数的自适应邻域选取策略得到的结果（底行）。为了体现出其中的一致性，我们在每行分别以不同的位置参数为颜色映射参数画出了计算的结果。从图 6.13的左列可以看出，对两种邻域选取策略的嵌入结果，颜色的变化具有很好的渐近性和一致性。这说明采用这两种邻域选取策略， LTSA 都能很好的恢复出左右位置参数。但对于上下位置参数，采用这两种邻域选取策略的 LTSA 的嵌入结果都有一些偏差。从图 6.13 的右列可以看出，采用自适应邻域选取策略的 LTSA 的嵌入结果中颜色变化比 k-邻域策略的嵌入结果具有更好的渐近性和一致性。这说明同 k-邻域策略相比，自适应邻域选取算法能减少嵌入结果同上下位

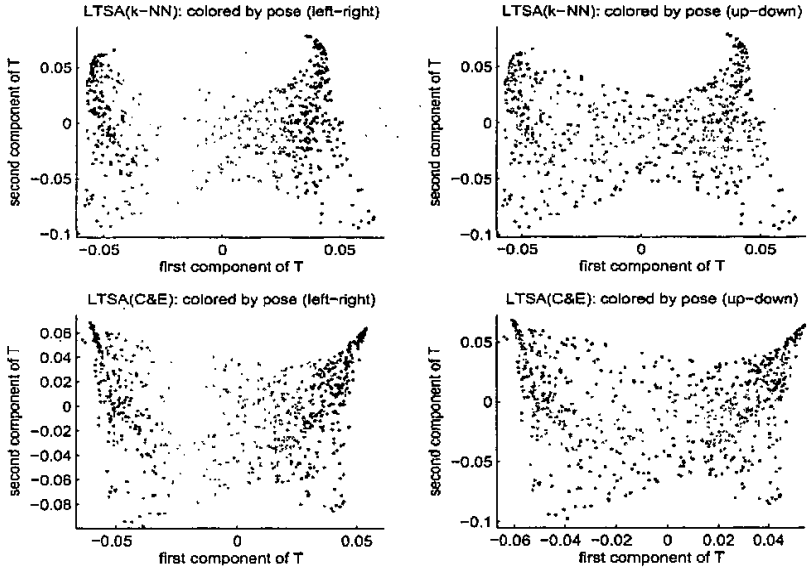

图 6.13: 采用不同邻域选取策略的 LTSA 所得到的 3 维嵌入结果中的两个分量：顶行：$k = 15$ 的 k-邻域策略（k-NN）；底行：$k_{\min} = 6$、$K_{\max} = 20$ 和 $\eta = 0.45$ 的自适应邻域选取策略（C&E）。左列：以左右位置参数作为颜色映射参数；右列：以上下位置参数作为颜色映射参数。

置参数的偏差。

§6.1 本章小节

在本章中，我们对前面给出的 MLLE 、自适应邻域选取策略和自适应LTSA几种算法给出了实验例子，实验例子包括了模拟例子和实际例子。首先，我们给出 MLLE 在 S-曲面和有“空洞”的 Swiss-roll 上的嵌入结果。实验结果表明，MLLE 能恢复出与流形等距的低维嵌入。接下来，我们用有噪音的圆柱螺线的例子来表明自适应邻域选取策略的效率。实验表明，自适应邻域选取策略能适用于Isomap、LLE、LTSA、MLLE等基于邻域选取的流形学习方法。自适应邻域选取策略还对样本点密度变化很大的 Swiss-roll 体现了很好的效果。然后，我们给出三峰流形的例子来表明 LTSA 的曲率模型（自适应 LTSA ）和 MLLE 的效率。实验结果表明，自适应 LTSA 和 MLLE 都能很好的处理曲率高度变化的流形，而且当流形有噪音的时候，这两种方法也有适用性。最后，我们给出了两个实际例子来表明 MLLE 和邻域选取策略对采自真实世界的数据的适用性。我们先给出一组有手写数字图像“2”、“4”、“5”组成的数据集。对于这个数据集，MLLE 能较好的对这三类手写数字图像进行分类。我们把 MLLE 用在一组由手写数字图像“1”组成的数据集上。MLLE 的嵌入结果能很明显的反映出手写数字

“1”的写法上的规律性。我们还对一组由两个位置参数和一个亮度参数所决定的人脸图像进行数值实验。MLLE 能很好的反映出人脸图像的这三个隐藏参数。实验结果还表明，自适应邻域选取策略能改善 LTSA 在这个实际例子上的结果。

第7章 总结和展望

数据降维是数据挖掘的一个非常重要的工具和方法。数据降维的目的是找出隐藏在高维数据中的低维结构，通常可以分为线性降维和非线性降维。线性降维是指通过降维所得到的低维数据能保持高维数据点之间的线性关系，常用的线性降维方法有主分量分析法（PCA）和多维尺度变化（MDS）等。由于线性降维方法的线性本质使其无法揭示复杂的非线性流形结构，而现实中数据的有用特征往往不是特征的线性组合，因此人们提出了非线性降维方法以处理非线性的高维空间中的数据点。非线性降维就是流形学习，常用的流形学习方法有等距映射（Isomap）、局部线性嵌入（LLE）、拉普拉斯特征映射（LE）、海赛局部线性嵌入（HLLE）、局部切空间排列（LTSA）等。流形学习的方法大致上可以分成两类：一类是全局方法（如 Isomap ），在降维时将流形上邻近的点映射到低维空间中的邻近点，同时保证将流形上距离远的点映射到低维空间中远距离的点；另一类是局部方法（如LLE、LE、HLLE、LTSA等），这些流形降维方法只是保证将流形上近距离的点映射到低维空间中的邻近点。无论全局方法还是局部方法都有着共同的特征：首先构造流形上样本点的局部邻域结构，然后用这些局部邻域结构来将样本点全局的映射到一个低维空间。它们之间的不同之处主要是在于构造的局部邻域结构不同以及利用这些局部邻域结构来构造全局的低维嵌入方式的不同。比如，Isomap 利用邻域点之间的关联和欧氏距离在数据点上构造一个有权图，然后再利用这个图来估计所有的样本点之间的测地距离，构造的全局低维坐标只是用以保持这个估计的测地距离。LLE 寻找每个样本点同它的邻域点之间的一种线性组合关系，并且使得低维空间中的嵌入坐标之间也保持这种线性组合关系。LTSA 将每个样本的所有邻域点投影到样本点在流形上的局部切空间上，并将所有的局部坐标排列以得到低维的全局坐标。

LLE 是流形学习方面经典的局部非线性方法，它有参数少、计算快、易求全局最优解等优点，并在图像分类、图像识别、谱重建、数据可视化等方面都有着广泛的应用。但它也有着一些缺点：对于等距的流形，LLE 并不能很好的恢复出同它等距的低维嵌入；用以求解重构权的有约束的最小二乘问题的最优解也许不是唯一的；而且采用正则化方法求解面临着正则因子 γ 的选择问题，不同的 γ 会得到不同的重构权，进而影响最终的嵌入结果；由于 LLE 保持邻近点的几何性质，对于有噪音、样本密度稀疏或者相互关联较弱的数据集，相隔较远的点之间的关联会减弱，这样在从高维到低维的映射过程中，很可能会将相隔较远的点映射到邻近点的位置。针对 LLE 的这些缺点，本文从 LLE 计算局部重构权的有约束的最小二乘问题入手，分析了 LLE 所计算的重构权的性质，在理论上证明了（用正则化方法）确定最优权在数值上是不稳定的，同时在给定精

度下，存在着多组线性无关的近似最优权向量。采用线性无关的权向量来建立邻域内稳定的局部线性结构，本文提出了修正局部线性嵌入方法（MLLE），降维的目的是为了在低维嵌入结果中保持这些线性无关的权向量。我们从理论上证明了 MLLE 对采自等距流形的样本点有着理想的结果，通过 MLLE 和 LTSA 之间的详细对比和理论分析，揭示了 LLE、MLLE 和 LTSA 之间的内在联系。这为进一步理解与分析建立了基础。

从流形学习的共同特征中可以看出，流形学习中有两个共同的因素决定着这些流形算法的效果。一个是当加强样本点之间的关联性的时候，应该如何自适应的选取邻域以匹配流形的局部几何性质。另一个是应该如何估计流形曲率和数据集的样本点密度变化对构造局部线性结构的影响，并且减少由此造成的在构造低维嵌入时产生的偏差。本文提出方法来解决流形学习中面临的两个主要的问题：（1）采用邻域压缩/扩张策略来自适应选取邻域的大小；（2）通过引入流形的曲率和样本点密度来自适应的减少嵌入结果的偏差。自适应邻域选取方法能用于所有基于邻域的流形学习方法，而第二项改进专门针对 LTSA 而设计。本文针对 LTSA 提出了修正的 LTSA 的曲率模型并与自适应邻域选取方法结合提出了自适应 LTSA 方法。虽然给出的理论分析是关于理想情形下光滑和局部等距的流形，本文所给出的自适应邻域选取方法和修正的 LTSA 的曲率模型也可以适用于有噪音的数据集。对于一个有着小噪音的数据集，采用k-邻域策略的修正 LTSA 同采用自适应邻域选取策略的修正 LTSA 一样也有效。但是，对于有着相对大的噪音的数据，自适应邻域选取方法要更为的稳定。

自适应邻域选取方法和自适应减少偏差的模型需要流形或数据集有着低维的几何结构。对于那些如人脸图像这样的数据集，它们的局部低维几何结构是不明确的，自适应方法的长处也许会被削弱甚至失去。另一方面，当一个有噪音的数据集采自于一个有高度变化的曲率的流形且噪音相对比较大的时候，也很难得出一个好的嵌入结果：当我们注意力集中在减少噪音时，曲率将会被忽略；而当我们忽视噪音时，曲率的估计将会不准确。这也是一个值得进一步研究的问题。关于流形学习方法，今后还有几个方面值得进一步的研究：

（1）对嵌入映射或者低维流形作出某种特定的假设，或者以保持高维数据的某种性质不变为目标，如何将问题转化成求解优化问题，并提供有效的解法。

（2）如何确定低维嵌入空间的维数。对此已经有一些研究工作[2, 16, 17, 20, 21, 27, 47, 50, 61]，但如何更加精确的确定低维空间的维数还是很值得深入的研究。

（3）现有的流形学习方法大多都假设采样数据比较稠密，而当采样数据很稀疏时或数据噪音很大时，应该如何进行有效的学习[9, 53]。

（4）本文提出的流形学习方法并不适用于自交的流形。目前只有少数算法用来考

虑处理这样的流形[1, 11, 62, 63]，还有很多问题值得深入的研究下去。

（5）本文提出的流形学习方法都是无监督的学习方法，将其作为一种监督或半监督的学习方法用于模式识别，虽然已有研究者涉足[6, 7, 42, 55, 70, 71]，但是目前在这方面的工作还很有限，仍值得进一步的深入研究。

（6）数据流是一种实时、连续、有序的数据组成的序列。它广泛存在于现实世界和工程实验中，如网络监控和通信工程、Web服务产生的日志纪录、传感器监控、视频流监控、金融市场上股票交易波动分析、天气或环境监督等都能生成大量的数据流。如何将流形学习方法应用于数据流，目前已有研究工作涉足[45]，但还有很多不足，仍值得进一步的深入研究。

（7）如何将流形学习与并行计算相结合，以处理海量的数据和提高流形学习的计算效率，目前在这方面没什么研究进展，但无疑是一个很值得深入研究的问题。

参考文献

[1] D. K. Agrafiotis and H. Xu. A self-organizing principle for learning nonlinear manifolds. Proc. Natl. Acad. Sci. USA 99: 15869-15872.

[2] B. Balázskégl. Intrinsic Dimension Estimation Using Packing Numbers. Neural Information Processing Systems 15 (NIPS'2002), 2003.

[3] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva and J. C. Langford. The Isomap Algorithm and Topological Stability . *Science*, vol. 295(5552), 7a, 2002.

[4] M. Berger and B. Gostiaux. Differential Geometry: Manifolds, Curves and Surfaces. GTM115. Springer-Verlag, 1974

[5] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, June 2003; 15 (6):1373-1396

[6] M. Belkin and P. Niyogi. Semi-supervised Learning on Manifolds. Machine Learning Journal, Special Issue on Clustering, to appear.

[7] M. Belkin and P. Niyogi. Using Manifold Structure for Partially Labelled Classification. *Neural Information Processing Systems 15 (NIPS'2002)*, pp. 929-936, 2003.

[8] M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering . *Neural Information Processing Systems 14 (NIPS'2001)*, pp. 585-591, 2002.

[9] Y. Bengio, J-F. Paiement, and P. Vincent. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. NIPS 16, 2003.

[10] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum. Graph Approximations to Geodesics on Embedded Manifolds. Technical Report, Department of Psychology, Stanford University, 2000.

[11] C. M. Bishop, M. Svensen and C. K. I. Williams. GTM: the generative topographic mapping. *Neural Computation*, vol 10, pp. 215–234, 1998.

[12] M. Brand. Charting a manifold. *Advances in Neural Information Processing Systems*, 15, MIT Press, 2003.

[13] M. Brand. Nonlinear dimensionality reduction by kernel eigenmaps. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 547-552, Acapulco, Mexico, 9-15 August 2003.

[14] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. *Proc. of 5th International Conference on Computer Vision*, pages 494– 499, 1995.

[15] A. Brun, H.-J. Park, H. Knutsson and C.-F. Westin. Coloring of DT-MRI Fiber Traces using Laplacian Eigenmaps. Eurocast 2003, Neuro Image Workshop, Las Palmas, February 2003. SPL Technical Report #369, posted May 2003.

[16] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, number 5, pp. 572 - 575, 1998.

[17] F. Camastra and A. Vinciarelli. Estimating the Intrinsic Dimension of Data with a Fractal-Based Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, Oct 2002.

[18] M. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice-Hall, Englewood Cliffs, 1976.

[19] H. Chang, D.Y. Yeung and Y. Xiong. Super-resolution through neighbor embedding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol.1, pp.275-282, Washington, DC, USA, 27 June - 2 July 2004.

[20] J. Costa and A. O. Hero. Manifold learning using Euclidean K-nearest neighbor graphs. *Proceedings of IEEE International Conference on Acoustic Speech and Signal Processing*, vol. 4, pp. 988-991, Montreal, May, 2004.

[21] J. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, vol. 25, no. 8, pp. 2210-2221, August 2004.

[22] T.Cox and M.Cox. Multidimensional Scaling. Chapman & Hall,Londom,2001.

[23] D. DeCoste. Visualizing Mercer Kernel Feature Spaces Via Kernelized Locally-Linear Embeddings. The 8th International Conference on Neural Information Processing (ICONIP2001), November 2001.

[24] D. Donoho and C. Grimes. Hessian Eigenmaps: new tools for nonlinear dimensionality reduction. *Proceedings of National Academy of Science*, 5591-5596, 2003.

[25] D. L. Donoho and C. Grimes. When Does ISOMAP Recover Natural Parameterization of Families of Articulated Images? Technical Report 2002-27, Department of Statistics, Stanford University, Aug 2002.

[26] A. Efros, V. Isler, J. Shi and M. Visontai. Seeing through water. NIPS 2004.

[27] K. Fukunaga and D.R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, C-20:176 - 183, 1971.

[28] G. H. Golub and C. F Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, Maryland, 3nd edition, 1996.

[29] A. Hadid and M. Pietikäinen. Efficient locally linear embeddings of imperfect manifolds. *Proc. Machine Learning and Data Mining in Pattern Recognition*. Lecture Notes in Computer Science 2734, Springer, 188-201.

[30] A. Hadid, O. Kouropteva and M. Pietikäinen. Unsupervised learning using locally linear embedding: experiments in face pose analysis. *Proc. 16th International Conference on Pattern Recognition*, August 11-15, Quebec City, Canada, 1:111-114.

[31] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, vol 84, pp. 502–516, 1989.

[32] J. Ham, D. D. Lee, S. Mika and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. International Conference On Machine Learning 21.

[33] J. Ham, D. D. Lee, S. Mika and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. Technical Report TR-110, Max-Planck-Institut für biologische Kybernetik, Tübingen, July 2003.

[34] X. He. Laplacian Eigenmap for Image Retrieval. Master's thesis, Computer Science Department, the University of Chicago, 2002.

[35] X. He and P. Niyogi. Locality Preserving Projections. Neural Information Processing Systems 16 (NIPS'2003).

[36] X. He, S. Yan, Y. Hu, P. Niyogi and H. Zhang. Face Recognition Using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 27, number 3, March 2005, pp 328- 340

[37] O. Jenkins and M. Mataric. A Spatio-temporal Extension to Isomap Nonlinear Dimension Reduction. *Proceedings of the Twenty-First International Conference on Machine Learning (ICML-2004)*, July 4-8, 2004, Banff, Alberta, Canada.

[38] O. C. Jenkins and M. J Mataric. Deriving Action and Behavior Primitives from Human Motion Data. In the IEEE/RSJ Internation Conference on Intelligent Robots and Systems (IROS-2002), pages 2551-2556, Lausanne, Switzerland, 2002.

[39] I. T. Jollife. Principal Component Analysis. Springer-Verlag, New York,1986.

[40] D. Kulpinski. LLE and Isomap Analysis of Spectra and Color Images. Master Thesis, School of Computer Science, Simon Fraser University.

[41] O. Kouropteva, O. Okun, A. Hadid, M. Soriano, S. Marcos and M. Pietikäinen. Beyond locally linear embedding algorithm. Technical Report MVG-01-2002,Machine Vision Group, University of Oulu.

[42] O. Kouropteva, O. Okun, and M. Pietikäinen. Classification of handwritten digits using supervised locally linear embedding algorithm and support vector machine. *Proc. of the 11th European Symposium on Artificial Neural Networks (ESANN'2003)*, April 23-25, Bruges, Belgium, 229-234.

[43] O. Kouropteva, O. Okun, and M. Pietikäinen. Selection of the optimal parameter value for the locally linear embedding algorithm. *Proc. of the 1 st International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'02)*, November 18-22, Singapore, 359-363.

[44] N. A. Laskaris, A. A. Ioannides. Semantic geodesic maps: a unifying geometrical approach for studying the structure and dynamics of single trial evoked responses. *Clinical Neurophysiology*, 113 (2002) 1209 - 1226.

[45] M. H. Law, N. Zhang and A. K. Jain. Nonlinear Manifold Learning for Data Stream. *Proceedings of SIAM Data Mining*, pp. 33-44, Orlando, Florida, 2004.

[46] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature* 401, 788-791 (1999).

[47] E. Levina and P.J. Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. Advances in Neural Information Processing Systems 17 (NIPS2004). MIT Press, 2005.

[48] F. Memoli and G. Sapiro. Distance Functions and Geodesics on Points Clouds. Institute for Mathematics and its Applications, Dec 2002.

[49] M. Niskanen and O. Silvén. Comparison of dimensionality reduction methods for wood surface inspection. Proc. 6th International Conference on Quality Control by Artificial Vision (QCAV 2003), May 19-23, Gatlinburg, Tennessee, USA.

[50] K. Pettis, T. Bailey, A. K. Jain and R. Dubes. An Intrinsic Dimensionality Estimator from Near-Neighbor Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, pp. 25-36, 1979.

[51] P. Perona and M. Polito. Grouping and dimensionality reduction by locally linear embedding. Neural Information Processing Systems 14 (NIPS'2001).

[52] D. D. Ridder and R. P. W. Duin. Locally linear embedding for classification. Technical report PH-2002-01, Pattern Recognition Group, Dept. of Imaging Science and Technology, Delft University of Technology, pp. 1-15, 2002.

[53] D. D. Ridder and V. Franc. Robust manifold learning. Technical report CTU-CMP-2003-08, Center for Machine Perception, Department of Cybernetics Faculty of Electrical Engineering, Czech Technical University, Prague, 2003, pp. 1-36.

[54] D. D. Ridder, M. Loog and M. J. T. Reinders. Local Fisher embedding. Proc. 17th International Conference on Pattern Recognition (ICPR2004), 2004.

[55] D. D. Ritter, O. Kouropteva, O. Okun, M. Pietikäinen and Duin RPW. Supervised locally linear embedding. *Artificial Neural Networks and Neural Information Processing, ICANN/ICONIP 2003 Proceedings*, Lecture Notes in Computer Science 2714, Springer, 333-341.

[56] S. Roweis and L Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290: 2323–2326, 2000.

[57] L. Saul and S. Roweis. Think globally, fit locally: unsupervised learning of nonlinear manifolds. *Journal of Machine Learning Research*, 4:119-155, 2003.

[58] F. Sha and LK Saul. Analysis and extension of spectral methods for nonlinear dimensionality reduction. Proceedings of the Twenty Second International Conference on Machine Learning. (ICML-05), Bonn, Germany, 2005.

[59] V. D. Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction . *Neural Information Processing Systems 15 (NIPS'2002)*, pp. 705-712, 2003.

[60] J. Tenenbaum, V. De Silva and J. Langford. A global geometric framework for nonlinear dimension reduction. *Science*, 290:2319–2323, 2000

[61] P. Verveer and R. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, pp. 81-86, 1995.

[62] J. J. Verbeek, N. Vlassis, and B. Kröse. The Generative Self-Organizing Map: a Probabilistic generalization of Kohonen's SOM. Technical report, Computer Science Institute, University of Amsterdam, The Netherlands, 2002. IAS-UVA-02-03.

[63] J. J. Verbeek, N. Vlassis and B. J. A. Kröse. Self-Organizing Mixture Models. *Neurocomputing* 63, pages 99-123, 2005.

[64] J. Wang, Z. Zhang and H. Zha Adaptive Manifold Learning. *Neural Information Processing Systems 17 (NIPS'2004)*, pp. 1473-1480.

[65] C. Williams. On a connection between kernel PCA and metric multidimensional scaling. *Machine Learning*,46:11-19,2002.

[66] J. Zhang, S. Z. Li and J. Wang. Manifold Learning and Applications in Recognition. Intelligent Multimedia Processing with Soft Computing. Springer-Verlag, Heidelberg. 2004.

[67] Z. Zhang and H. Zha. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. *SIAM J. Scientific Computing*, 26(1):313–338, 2004.

[68] H. Zha and Z. Zhang. Isometric Embedding and Continuum ISOMAP. *Proceedings of the Twentieth International Conference on Machine Learning* (ICML-2003), 864-871, 2003.

[69] H. Zha and Z. Zhang. Spectral Analysis of Alignment in Manifold Learning. Submitted to *Journal of Machine Learning Research*, 2004.

[70] D. Zhou and B. Schölkopf. Learning from Labeled and Unlabeled Data Using Random Walks. DAGM'04: 26th Pattern Recognition Symposium, 2004.

[71] D. Zhou, B. Schölkopf and T. Hofmann. Semi-supervised Learning on Directed Graphs. Advances in Neural Information Processing Systems 17, 2005.

[72] 陈省身, 陈维桓. 微分几何讲义. 北京大学出版社, 1983.

致　谢

首先衷心感谢我的导师张振跃教授，衷心感谢他多年来的严格要求，耐心指导和不断激励。他严谨的治学态度、渊博的学识以及敏锐的学术眼光给予了我深刻的影响，并将使我终生受益。几年来他在研究方向上不断鼓励、引导我探索新的学术方向、跟踪新的理论；同时又提供了一个宽松的研究学习环境。值此在论文完成之际，谨向导师及其家人致以诚挚和深切的谢意。

我还要感谢在浙大九年中给我关怀和帮助的老师们，感谢我本科班主任尹永成老师，感谢管志成老师、卢兴江老师、应文隆老师等浙大数学系所有的老师们，正是有了他们，才使我顺利度过了浙大的九年求学生涯。

感谢我们研究方向的每位成员：李旭东、罗政、方敏、裘渔洋、杜克勤、方腊、卢俊峰、李丽敏、陈孝瑞、何睿、赵凌潇、曹沛霖、徐丽娟，大家共同学习，互相帮助，度过了许多难忘的时光。还要感谢我的同窗陈汀、李银飞、刘魁、季敏、王成、徐佳佳、许鑫慧、严卫华、张挺等，陪我度过了美好的浙大生活。

我还要特别感谢我的家人。感谢我的父母和亲人对我的关心和照顾。没有父母和亲人的长期的理解和支持，要完成这样的研究工作几乎是难以想象的。

流形学习的理论与方法研究

作者： 王靖
学位授予单位： 浙江大学
被引用次数： 57次

本文读者也读过(10条)

1. 李波 基于流形学习的特征提取方法及其应用研究[学位论文]2008
2. 黄启宏 流形学习方法理论研究及图像中应用[学位论文]2007
3. 魏艳涛 基于流形学习的数据降维方法研究[学位论文]2008
4. 李春光 流形学习及其在模式识别中的应用[学位论文]2007
5. 王庆军 流形学习算法分析及应用研究[学位论文]2008
6. 曾宪华 流形学习的谱方法相关问题研究[学位论文]2009
7. 吴晓婷 基于流形学习的数据降维算法的研究[学位论文]2010
8. 朱韬 流形学习方法在图像处理中的应用研究[学位论文]2009
9. 陆建新 流形学习理论及其应用研究[学位论文]2007
10. 侯晓宇 基于流形学习的特征提取方法研究[学位论文]2009

引证文献(35条)

1. 林强,董平,林嘉宇 基于增量的ISOMAP算法研究[期刊论文]-数字技术与应用 2015(05)
2. 改进的有监督的局部线性嵌入算法及实验演示[期刊论文]-现代计算机（专业版） 2014(10)
3. 王宗利,刘希玉 一种基于流形的蚁群聚类算法[期刊论文]-山东大学学报（理学版） 2008(11)
4. 侯文广,丁明跃 基于流形学习的三维空间数据网格剖分方法[期刊论文]-电子学报 2009(11)
5. 周梅,刘秉瀚 基于拉普拉斯特征映射的分类器设计[期刊论文]-计算机工程 2009(16)
6. 王博,刘美玲,张学敏 两种流形学习算法的对比研究[期刊论文]-微型机与应用 2013(08)
7. 范进富,陈锻生 流形学习与非线性回归结合的头部姿态估计[期刊论文]-中国图象图形学报 2012(08)
8. 肖传乐,曹槐 基于流形学习的基因表达谱数据可视化[期刊论文]-生物信息学 2009(01)
9. 利用局部线性嵌入的模态识别[期刊论文]-西安交通大学学报 2013(01)
10. 谷瑞军 基于流形学习的高维空间分类器研究[学位论文]博士 2008
11. 赵洪杰,潘紫微,童靳于,刘燕 基于相空间重构与非线性流形的滚动轴承复合故障诊断[期刊论文]-振动与冲击 2013(11)
12. 唐晓燕,高昆,倪国强,朱振宇,程颢波 基于流形学习和空间信息的改进N-FINDR端元提取算法[期刊论文]-光谱学与光谱分析 2013(09)
13. 刘海锋 基于相对变换的非线性降维研究[学位论文]硕士 2009
14. 王宪保,陆飞,陈勇,方路平,王守觉 仿生模式识别的算法实现与应用[期刊论文]-浙江工业大学学报 2011(01)
15. 魏莱,王守觉,徐菲菲 一种自适应邻域选择算法[期刊论文]-模式识别与人工智能 2008(03)
16. 黄静,肖先勇,刘旭娜 短期负荷局部线性嵌入流形学习预测法[期刊论文]-电力系统保护与控制 2012(07)
17. 杨昭 面向物联网的海量数据降维算法研究[学位论文]硕士 2011
18. 高小方 流形学习方法中的若干问题分析[期刊论文]-计算机科学 2009(04)
19. 靳丽丽 基于子空间算法的人脸识别——流形学习算法[学位论文]硕士 2011
20. 韦佳 流形学习与基于流形假设的半监督学习研究[学位论文]博士 2009
21. 蒋莲 基于统计流形的Bag of Features降维研究与应用[学位论文]硕士 2013
22. 王常武,吕伟泽,王宝文,刘文远 黎曼流形数据类型的判别[期刊论文]-小型微型计算机系统 2013(11)
23. 井经涛 一种智能化网络安全态势评估方法[学位论文]硕士 2011
24. 屈治礼 高维数据可视化研究及在商业智能中的应用[学位论文]硕士 2013
25. 常伟 流形学习理论研究及相关改进[学位论文]硕士 2010
26. 侯晓宇 基于流形学习的特征提取方法研究[学位论文]硕士 2009

27. 李剑 掌纹识别算法研究[学位论文]硕士 2011

28. 付会欣 李群机器学习中的辛群分类器研究[学位论文]硕士 2008

29. 冉丹 基于核方法和流形学习的雷达目标识别[学位论文]硕士 2008

30. 申中华 数据降维技术的建模研究与应用——特征降维及其应用[学位论文]硕士 2008

31. 朱韬 流形学习方法在图像处理中的应用研究[学位论文]硕士 2009

32. 贺惠新 基于流形学习的高维流场数据分类研究[学位论文]硕士 2008

33. 宋涛 基于流形学习的风电机组传动系统早期故障特征提取方法研究[学位论文]博士 2013

34. 王彤 高维生物数据的分类与预测研究[学位论文]博士 2009

35. 王雷 基于全局统计与局部几何性质的数据降维算法研究[学位论文]博士 2009

引用本文格式：王靖 流形学习的理论与方法研究[学位论文]博士 2006