Statistics and Application 统计学与应用, 2017, 6(1), 7-16
Published Online March 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.61002

Hans汉斯

Research on the Problem of Selecting the Parameter Values for LLE Algorithm

Fang Li, Xiang Gao*

School of Mathematical Sciences, Ocean University of China, Qingdao Shandong
Email: oucfLi@163.com, *gaoxiangshuli@126.com

Received: Feb. 21st, 2017; accepted: Mar. 6th, 2017; published: Mar. 9th, 2017

Abstract

Aiming at the problem of how to select two parameter values, the number of nearest neighbor points k and the output dimension d, locally linear embedding (LLE) algorithm is improved. Firstly, we describe dimensionality reduction and why reduce dimension of high-dimensional data. Secondly, we discuss the basic idea and computational procedure of the LLE algorithm. Finally, the problems existing in the LLE algorithm are analyzed.

Keywords

LLE Algorithm, Correlation Coefficient, Number of Nearest Neighbor Points k, Maximum Likelihood Estimation, Output Dimension d

LLE算法中有关参数选取问题的研究

李　芳，高　翔*

中国海洋大学数学科学学院，山东　青岛
Email: oucfLi@163.com, *gaoxiangshuli@126.com

收稿日期：2017年2月21日；录用日期：2017年3月6日；发布日期：2017年3月9日

摘　要

本文针对locally linear embedding (LLE)算法中的两个参数：近邻点的个数k和降维后输出的维数d如何选取的问题，对LLE算法进行了改进。首先对降维的相关知识进行了描述，并具体介绍了对高维数据进行降维的目的。其次，讨论了LLE算法的基本思想和计算步骤。最后，针对LLE算法中存在的问题进行了分析。

*通讯作者。

文章引用: 李芳, 高翔. LLE 算法中有关参数选取问题的研究[J]. 统计学与应用, 2017, 6(1): 7-16.
https://doi.org/10.12677/sa.2017.61002

关键词

LLE算法，相关系数，近邻点的个数 k ，极大似然估计，降维后输出的维数 d

1. 研究背景

1.1. 问题的提出

为了提高统计模式识别的正确识别率，人们通常需要采集数量巨大的数据特征，使得原始空间或输入空间的维数可能高达几千维或者上万维。如果在输入空间上直接进行分类器训练，那么就很可能会带来如下两个问题[1]：

1) 很多在低维空间具有良好性能的分类算法在计算上变得不可行；

2) 在训练样本容量一定的前提下，特征维数的增加将使得样本统计特性的估计变得更加困难，从而降低分类器的推广能力或泛化能力，呈现所谓的“过学习”或“过训练”的现象。

要避免出现“过学习”的情况，用于统计分类器训练的训练样本个数必须随着维数的增长而呈指数增长，从而造成人们所说的“维数灾难”，即在给定精度下，准确地对某些变量的函数进行估计，所需样本量会随着样本维数的增加而呈指数形式增长。

为了解决“维数灾难”的问题，且在涉及的维数较少的情况下得到原始高维空间或输入空间较多的信息。这个时候，人们就希望通过降维算法从高维数据中提取有效的、紧致的描述，即在保持数据信息损失最小的情况下，寻找原始高维空间中数据的内在规律与本质特征，减少冗余信息所带来的误差，提高问题解决的效率和精度。

1.2. 降维的含义与目的

所谓的降维就是指采用某种映射方法[2]，将原高维空间中的数据点映射到低维的空间中，从而找到隐藏在高位观测数据中有意义的低维结构。降维的本质是学习一个映射函数 $f:x\to y$，其中 x 是原始高维空间中数据点的表达，目前最多使用向量表达形式。 y 是数据点映射后的低维向量表达，通常 y 的维度小于 x 的维度。 f 可能是显式的或隐式的、线性的或非线性的。

对原始空间或输入空间的高维数据降维的目的主要有以下四个方面：

1) 压缩数据到低维空间，可以解决“维数灾难”的问题，降低存储要求，并简化计算复杂度。

2) 在剔除冗余信息的同时，也降低了噪声对原始数据的影响。

3) 从非结构化数据集中提取出某种结构化成分，有利于寻找原始高维空间中数据的内在规律与本质特征，以便更好地认识和理解数据。

4) 把数据投影到低维空间，特别是人眼可观测的二维空间或三维空间，可以实现高维数据可视化。

2. LLE 算法介绍

流形学习的目的[3]，就是找出原始高维空间中数据样本点隐藏在高维空间中的低维结构。针对高维空间中的非线性流行，2000 年，Roweis 和 Saul 在 Science 上提出了一种非线性降维方法——LLE 算法[4]。

2.1. LLE 算法的基本思想

LLE 是一种无监督的降维方法。其核心主要是将流形上的近邻点映射到低维空间的近邻点，保存原流行中的局部几何特性，以达到高维数据映射到低维全局坐标系中的目的。该算法的前提假设是采样数据所在的低维流形在局部是线性的，即每个采样点可以用它的近邻点线性表出。

LLE 算法基于用局部的线性来逼近全局的非线性，通过保持高维数据与低维数据间的局部领域几何结构不变的几何思想，使在高维空间中相邻或相关的两个点映射到低维空间中也同样相邻或相关。LLE 算法是依赖于局部线性的的算法。它认为在局部意义下，数据的结构是线性的，或者说，局部意义下的点在一个超平面上。再通过互相重叠的局部邻域来提供整体的信息，从而保证整体的几何性质，得到一个全局的坐标系统。

如图 1 所示，LLE 算法能成功地将三维非线性数据映射到二维空间中[5]。(A)中不同的颜色分别代表原始三维空间中流行的不同结构；(B)是通过随机采样从原始三维空间(A)中提取的数据样本点；通过 LLE 算法降维后，我们看到原始三维空间的数据样本点(B)映射到了二维空间中；通过观察(C)，我们知道，通过 LLE 算法降维后的数据样本点在二维空间中仍能保持相对独立的状态，即红色的点互相接近，黄色的点互相接近，蓝色的也互相接近。这说明在将原始高维空间中的数据样本点映射到低维全局坐标系的过程中，LLE 算法确实能有效地保持原有数据的邻域特性和流形结构。而线性方法，如 PCA 和 MDS，都不能与 LLE 算法相比拟。

当原始高维空间中的数据分布在缺少北极面的球形面时，如最后一行图所示，在保持原有数据流行的局部领域几何结构不变的意义下，应用 LLE 算法仍能很好地将其映射到二维空间中。但是，在有些情况下 LLE 算法也并不适用，即如果原始高维空间中的数据分布在整个封闭的球面上，LLE 算法则不能通过降维将它映射到二维空间，且不能保持原有的数据流形。所以在我们应用 LLE 算法处理原始高维空间中的数据的时候，首先要假设原始高维空间中的数据不是分布在闭合的球面或者椭球面上。

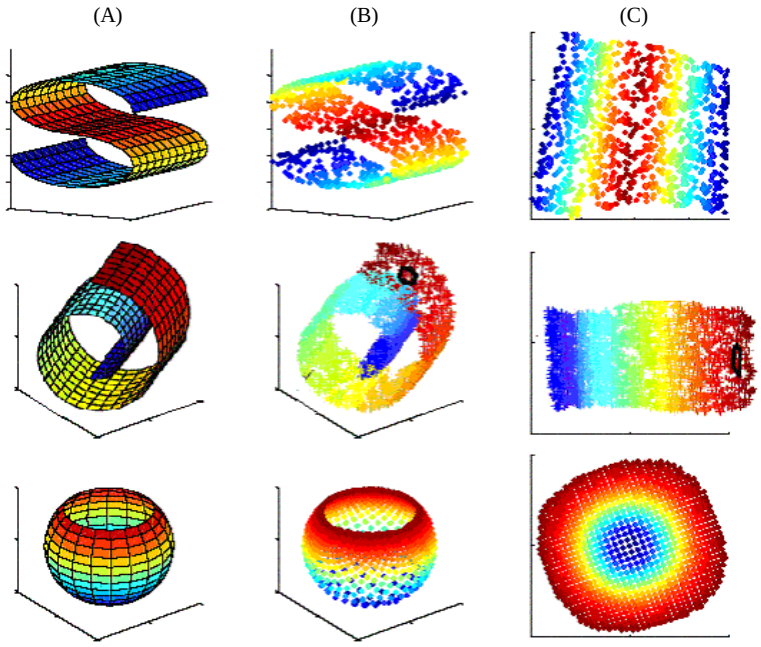

Figure 1. Reduced-Dimension Map: map three-dimensional data into a two-dimensional system by LLE
图 1. LLE 算法将三维非线性数据映射到二维空间的降维图

2.2. LLE 算法的计算步骤

设给定数据集合 $X=\{x_1,x_2,\cdots,x_N\}$ 采样于某个潜在的光滑的流行，且这些数据包含 N 个实值向量，向量的维数为 D。采样的数据点要求足够多，每个采样数据点及其近邻点都落在该潜在流行的一个局部线性块上或该块附近。从而采样数据点所在的低维流形在局部是线性的，每个采样点都可以用它的近邻点线性表出。进一步我们就可以得到数据样本点的局部重建权值矩阵 W。由于 LLE 算法是基于通过保持高维数据与低维数据间的局部领域几何结构不变的几何思想，用局部的线性来逼近全局的非线性。而局部重建权值矩阵 W 又代表着局部信息，可以用于刻画流行的局部几何性质，因而保持 W 固定不变，在使得重构的误差最小的条件下，优化输出原始高维空间中所有的数据样本点 $X=\{x_1,x_2,\cdots,x_N\}\in R^D$ 映射嵌入到低维空间中的数据 $Y=\{y_1,y_2,\cdots,y_N\}\in R^d$。

1) 对于高维空间的每个样本点 x_i $i=1,2,\cdots N$，计算和其他 $N-1$ 个样本点之间的距离。根据距离的远近，找出与 x_i 最近的 k 个点作为其近邻点(k 是一个预先给定值)。

距离公式通常采用 $d_{ij}=\left(\sum_k\left|x_{ik}-x_{jk}\right|^p\right)^{\frac{1}{p}}$ 来表示。当 $p=2$ 时表示欧氏距离；当 $p=2$ 时表示 City－Block 距离；当 $p=\infty$ 时表示 Doninance 距离。原始 LLE 算法是采用欧式距离来确定每个数据点的 k 个近邻点。

2) 由每个样本点 x_i $i=1,2,\cdots N$ 的 k 个近邻点计算出该样本点的局部重建权值矩阵 W：

$$W=\arg\min\varepsilon(W)=\sum_{i=1}^{N}\left|x_i-\sum_{j=1}^{k}w_{ij}x_{ij}\right|^2$$

其中，x_{ij} 表示第 i 样本点个 x_i 的 k 个近邻点。w_{ij} 表示第 j 个近邻点对第 i 样本点 x_i 的权值(即贡献)，且满足条件：$\sum_{j=1}^{k}w_{ij}=1$。

3) 将原始高维空间中的所有数据样本点映射嵌入到低维空间中，映射嵌入满足如下条件：

$$\min\Phi(Y)=\sum_{i=1}^{N}\left|y_i-\sum_{j=1}^{k}w_{ij}y_{ij}\right|^2$$

其中，$\Phi(Y)$ 为损失函数。此时固定局部重建权值矩阵 W，优化输出向量 Y。且向量 Y 应满足以下两个条件：$\sum_{i=1}^{N}y_i=0$ 和 $\frac{1}{N}\sum_{i=1}^{N}y_iy_i^{\mathrm{T}}=I$。

根据高维空间中的样本点 $X=\{x_1,x_2,\cdots,x_N\}\in R^D$ 和它的近邻点 x_{ij} 之间的权重 w_{ij} 来计算映射嵌入到低维空间中的数据 $Y=\{y_1,y_2,\cdots,y_N\}\in R^d$。由于 LLE 算法要求映射嵌入低维空间中的数据尽量保持高维空间中的局部线性结构。而局部重建权值矩阵 W 又代表着局部信息，因而在将原始高维空间中的所有数据样本点映射嵌入到低维空间中的过程中，要保持局部重建权值矩阵 W 固定不变。另外，为了让重构的误差最小，局部重建权值矩阵 W 必须服从一种重要的对称性。即对所有特定样本点来说，他们与自己的近邻点之间经过旋转、重新排列、转换等各种变换后，彼此之间的拓扑结构必须保持不变，已达到让局部重建权值矩阵 W 准确描述每个近邻的基本几何特性的要求。所以可以认为，应用 LLE 算法映射嵌入后的低维流形上的局部拓扑结构，和原始高维空间内的数据的局部几何特性是完全等效的(图 2)。

3. LLE 算法改进

3.1. LLE 算法的问题分析

通过介绍 LLE 算法，我们了解到，在 LLE 算法中有两个参数需要设置[6]。其一是 LLE 算法的第一

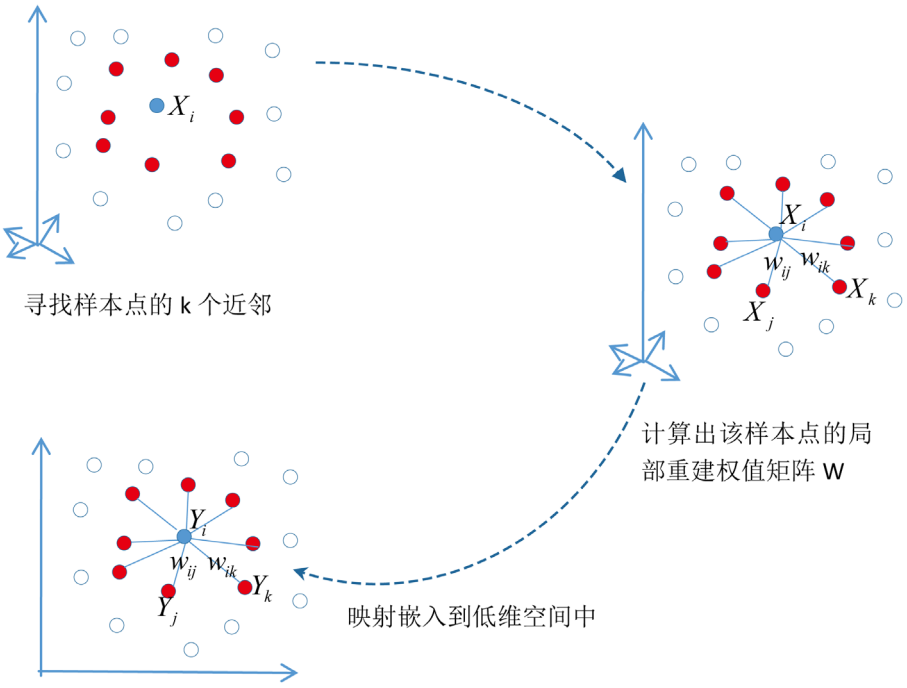

Figure 2. Flowchart: calculation steps of LLE algorithm
图 2. LLE 算法的计算步骤图

步中近邻点的个数 k；其二是降维后输出的维数 d，即对观测数据集进行建模所需的最少独立变量的个数，通常称之为最优嵌入维数，也称为本征维数。

(一) 我们先来分析 LLE 算法中的第一个参数，即原始高维空间中的数据样本点 x_i $i=1,2,\cdots N$ 的 k 个近邻点个数的有关问题。

在使用 LLE 对数据进行处理的过程中，我们发现，近邻点个数 k 的取值对实验结果的影响是比较大的，在大部分情况下是根据经验选择的一个活动性很大的值，需要尝试且不方便控制。

1) k 的取值[7]过大就可能会影响整个流行的平滑性，就不能体现其局部特性，这样会导致 LLE 算法趋向于 PCA 算法，甚至丢失流行的一些小规模的结构。而 k 的取值过小，LLE 就很难保证样本点在低维空间的拓扑结构，则又可能会把连续的流行脱节的子流行。

2) 如果对所有样本集区域内的样本点选取相同个数的近邻点，对于所含结构信息很重要的样本集区域，也许就会丢失很多我们所需要和寻找的内容，相反，对于所含结构信息不重要的样本集区域，就会额外加大许多不必要的计算量，浪费了计算机的效率和时间，甚至可能会因为多选取了一些错误的近邻点，破坏了其真实的局部结构特征，使最后的低维流行与实际不符，从而误导我们的分析，也就是所谓的噪声干扰和冗余数据的影响。

3) 对于弯曲弧度非常大的不光滑流行，基本 LLE 算法所采用的一致的值也会使高维数据流行在低维空间映射的结果与数据集本身的实际不符。

另外，LLE 算法假设原始高维空间中的数据样本点在流行上的分布是比较均匀的，即数据样本点是均匀采样于原始高维空间的。但通常情况下，这种理想状态的假设是不满足的，我们很难做到数据样本点均匀采样于原始高维空间，特别是对于那些分布不均匀的数据样本点来说[8]。k 的取值对 LLE 算法结果的影响就更为明显了，因为与样本分布稀疏的那部分区域相比，由 k 个近邻点所组成的局部邻域显然要比，在样本分布比较密集的那部分区域内，由 k 个近邻点所组成的局部邻域要大得多。而且就不同样

本点 x_i $i=1,2,\cdots N$ 的 k 个近邻点所组成的不同的局部邻域而言，样本分布的密集稀疏程度对所提供信息的重要程度也存在明显差异[9]。

(二) 我们再来分析 LLE 算法中的第二个参数，即降维后输出的维数 d 的有关问题。

在应用 LLE 算法将原始高维空间中的所有数据样本点映射嵌入到低维空间后，映射嵌入后输出的维数 d 的取值也是一个重要因素[10]，它决定了应用 LLE 算法映射嵌入后的低维流形上的局部拓扑结构能否充分描述原始高维空间内的数据的局部几何特征。即在保持原始高维空间中的数据信息损失最小的情况下，寻找原始高维空间中数据的内在规律与本质特征，进而寻求一个降维后输出的最低维数 d 对原始高维空间中的数据进行合理的、有效的低维可视表示。最低维数 d 的取值过大将会使降维结果中含有过多的噪声， d 的取值过小，致使本来不同的点在低维空间可能会彼此交叠。维数 d 的取值是一个需要为们去估计的未知量，准确地估计出高维数据的本征维数，对接下来的一系列降维处理问题都有着重要的指导意义。

3.2. LLE 算法的改进

(一) 近邻点个数 k 值的选择

LLE 算法基于用局部的线性来逼近全局的非线性，通过保持高维数据与低维数据间的局部领域几何结构不变的几何思想，使在高维空间中相邻或相关的两个点映射到低维空间中也同样相邻或相关。我们知道，相关系数是用以反映两变量之间相关关系密切程度的统计指标。针对问题(一)，本文选择相关系数的绝对值作为高维空间的每个样本点 x_i $i=1,2,\cdots N$ 选择其 k 个近邻点的衡量标准，

给定两个变量 $x=(x_1,x_2,\cdots,x_n)$ $y=(y_1,y_2,\cdots,y_n)$，则两变量 x 和 y 的相关系数的定义如下：

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$|r_{xy}|$ 的取值在 0 和 1 之间，$|r_{xy}|$ 取值越大，说明两变量之间相关关系密切程度也就越高，$|r_{xy}|$ 取值越小，说明两变量之间相关关系密切程度也就越低。我们知道当 $|r_{xy}|<0.3$ 时，表示变量 x 与变量 y 不相关；当 $0.3\le|r_{xy}|<0.5$ 时，表示变量 x 与变量 y 低度相关；当 $0.5\le|r_{xy}|<0.8$ 时，表示变量 x 与变量 y 中度相关；当 $|r_{xy}|\ge0.8$ 时，表示变量 x 与变量 y 呈现高度相关的关系。

设给定原始高维空间中所有的数据样本点 $X\in R^D$ 包含 N 个实值向量。首先根据公式

$$r_{x_ix_j}=\left|\frac{\sum_{m=1}^{D}\left(x_i^{(m)}-\bar{x}_i\right)\left(x_j^{(m)}-\bar{x}_j\right)}{\sqrt{\sum_{m=1}^{D}\left(x_i^{(m)}-\bar{x}_i\right)^2}\sqrt{\sum_{m=1}^{D}\left(x_j^{(m)}-\bar{x}_j\right)^2}}\right|$$

计算样本点 $x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(D)}\right)$ 与样本点 $x_j=\left(x_j^{(1)},x_j^{(2)},\cdots,x_j^{(D)}\right)$ 的相关系数的绝对值大小。依据上述 $|r_{x_ix_j}|$ 的取值大小对两变量之间相关关系密切程度高低的描述，我们可以认为当 $|r_{x_ix_j}|\ge0.5$ 时，样本点 x_j 是样本点 x_i 的近邻点，即样本点 x_i 可由样本点 x_j 重构。选取这样的近邻点才能够使得在高维空间中的重构误差 $\varepsilon(w)$ 较小。

对于高维空间中的样本点 x_i，我们已经知道，当 $|r_{x_ix_j}|\ge0.5$ 时，样本点 x_j 是样本点 x_i 的近邻点。前

文我们已经提到，在通常情况下很难做到数据样本点均匀采样于原始高维空间。从而我们不妨假设，在样本点 x_i 领域内一共有 $k_i\ i=1,2,\cdots,N$ 个样本点 x_j 是其近邻点，进一步有：

在样本点 x_1 邻域内有 k_1 个样本点 x_j 使得 $\left|r_{x_1x_j}\right|\geq 0.5$

在样本点 x_2 邻域内有 k_2 个样本点 x_j 使得 $\left|r_{x_2x_j}\right|\geq 0.5$

…

在样本点 x_N 邻域内有 k_N 个样本点 x_j 使得 $\left|r_{x_Nx_j}\right|\geq 0.5$

转化为数学表达式为：

$$x_1=w_{11}x_{11}+w_{12}x_{12}+\cdots+w_{1k_1}x_{1k_1}=\sum_{j=1}^{k_1}w_{1j}x_{1j}$$

$$x_2=w_{21}x_{21}+w_{22}x_{22}+\cdots+w_{2k_2}x_{2k_2}=\sum_{j=1}^{k_2}w_{2j}x_{2j}$$

$$\cdots$$

$$x_N=w_{N1}x_{N1}+w_{N2}x_{N2}+\cdots+w_{Nk_N}x_{Nk_N}=\sum_{j=1}^{k_N}w_{Nj}x_{Nj}$$

即 $x_i=\sum_{j=1}^{k_I}w_{ij}x_{ij}$ 。

下面运用最下二乘法使得 $\min\varepsilon(W)=\sum_{i=1}^{N}\left|x_i-\sum_{j=1}^{k_i}w_{ij}x_{ij}\right|^2$，计算出局部重建权值矩阵 w_{ij}。我们不妨以 $x_1=w_{11}x_{11}+w_{12}x_{12}+\cdots+w_{1k_1}x_{1k_1}=\sum_{j=1}^{k_1}w_{1j}x_{1j}$ 为例，将方程写成矩阵的形式：

$$\begin{bmatrix}x_1^{(1)}\\x_1^{(2)}\\\vdots\\x_1^{D}\end{bmatrix}=\begin{bmatrix}x_{11}^{(1)}&x_{12}^{(1)}&\cdots&x_{1k_1}^{(1)}\\x_{11}^{(2)}&x_{12}^{(2)}&\cdots&x_{1k_1}^{(2)}\\\vdots&\vdots&&\vdots\\x_{11}^{(D)}&x_{12}^{(D)}&\cdots&x_{1k_1}^{(D)}\end{bmatrix}\begin{bmatrix}w_{11}\\w_{12}\\\vdots\\w_{1k_1}\end{bmatrix}$$

我们令 $X_1=\begin{bmatrix}x_1^{(1)}\\x_1^{(2)}\\\vdots\\x_1^{(D)}\end{bmatrix}$；$Z_1=\begin{bmatrix}x_{11}^{(1)}&x_{12}^{(1)}&\cdots&x_{1k_1}^{(1)}\\x_{11}^{(2)}&x_{12}^{(2)}&\cdots&x_{1k_1}^{(2)}\\\vdots&\vdots&&\vdots\\x_{11}^{(D)}&x_{12}^{(D)}&\cdots&x_{1k_1}^{(D)}\end{bmatrix}$；$W_1=\begin{bmatrix}w_{11}\\w_{11}\\\vdots\\w_{1k_1}\end{bmatrix}$。从而 $X_1=Z_1W_1$。于是根据最小二乘法可计算出：

$$\hat{W}_1=\left(Z_1^{\mathrm{T}}Z_1\right)^{-1}Z_1^{\mathrm{T}}X_1=\begin{bmatrix}\hat{w}_{11}\\\hat{w}_{12}\\\vdots\\\hat{w}_{1k_1}\end{bmatrix}$$

同理，我们可以得到 $\hat{W}_i=\left(Z_i^{\mathrm{T}}Z_i\right)^{-1}Z_i^{\mathrm{T}}X_i=\left(\hat{w}_{i1},\hat{w}_{i2},\cdots,\hat{w}_{ik_i}\right)^{\mathrm{T}}\ i=1,2,\cdots,N$ 。

又因为对于高维空间的每个样本点的 $k\in\{k_1,k_2,\cdots k_N\}$ 个近邻点，其局部重建权值矩阵 w_{ij} 满足条件：

$\sum_{j=1}^{k} w_{ij} = 1$，从而对 $\hat{W}_i$ 进行单位化处理，为了表达简便，我们记

$$\hat{W}_i = \left(\frac{\hat{w}_{i1}}{\|\hat{w}_i\|}, \frac{\hat{w}_{i2}}{\|\hat{w}_i\|}, \cdots, \frac{\hat{w}_{ik_i}}{\|\hat{w}_i\|} \right)^{\mathrm{T}} \quad i = 1, 2, \cdots, N$$

从而 $W = \left(\hat{W}_1, \hat{W}_2, \cdots, \hat{W}_N \right) = \arg\min \varepsilon(W) = \sum_{i=1}^{N} \left| x_i - \sum_{j=1}^{k_i} w_{ij} x_{ij} \right|^2$ 即为所求。

(二) 降维后输出的维数 d 的估计

本文的估计算法是在极大似然估计方法的基础上实现的。通过对映射嵌入后的低维空间的 N 个局部邻域分别运用极大似然估计方法[11]，得到 N 个局部本征维数的估计值 $\hat{d}_k(y)\ k \in \{k_1, k_2, \cdots, k_N\}$。然后再对 $\hat{d}_k(y)$ 进行加权平均，最后得到整个映射嵌入后的低维空间的维数 d 的极大似然估计值 $\hat{d}$。具体算法如下：

假设高维空间的数据集合 $X = \{x_1, x_2, \cdots, x_N\} \in R^D$，通过 LLE 算法，在低维空间中表示为数据集合 $Y = \{y_1, y_2, \cdots, y_N\} \in R^d$，即数据集合 $Y = \{y_1, y_2, \cdots, y_N\} \in R^d$ 是来自高维空间的数据集合 $X = \{x_1, x_2, \cdots, x_N\} \in R^D$ 的嵌入结果，且维数 $d \ll D$。

极大似然估计的基本思想是[12]，对于映射嵌入后的低维空间中给定的一个点 $y \in \{y_1, y_2, \cdots y_N\}$，我们构造以 y 为中心，r 为半径的球面 $S_y(t)$，这样的球面共有 N 个。当球面 $S_y(t)$ 的半径 r 足够小时，可认为球面 $S_y(t)$ 的密度 $f(y) \approx$ 常数。假设观测样本 $Y = \{y_1, y_2, \cdots, y_N\} \in R^d$ 呈齐次的泊松分布，考察非平稳过程 $\{N(t, y), 0 \le t \le r\}$，显然 $N(t, y) = \sum_{i=1}^{N} I\{y \in S_y(t)\}$ 表示数据样本点 $\{y_1, y_2, \cdots y_N\}$ 落入以 y 为中心，r 为半径的球面 $S_y(t)$ 中的样本点个数。固定 N，$N(t, y) = \sum_{i=1}^{N} I\{y \in S_y(t)\}$ 是一个二项随机过程，下面我们用平稳的泊松过程来逼近这一过程。

对于映射嵌入后的低维空间中给定的一个点 $y \in \{y_1, y_2, \cdots y_N\}$，记 $T_k(y)\ k \in \{k_1, k_2, \cdots, k_N\}$ 为 $y \in \{y_1, y_2, \cdots y_N\}$ 的第 k 个近邻点(从近到远)到 $y \in \{y_1, y_2, \cdots y_N\}$ 的距离，则有：

$$\frac{k}{N} \approx f(y) V(d) \left[T_k(y) \right]^d$$

其中 $V(d) = \pi^{d/2} \left[\Gamma(d/2 + 1) \right]^{-1}$ 表示 d 维单位球体 $S_y(t)$ 的体积。则对于固定的 t，可以得到该泊松过程的参数：

$$\lambda(t) = f(y) V(d) dt^{t-1}.$$

如果不再考虑 $N(t, y) = \sum_{i=1}^{N} I\{y \in S_y(t)\}$ 对 $y \in \{y_1, y_2, \cdots y_N\}$ 的依赖，则 $\lambda(t)$ 是 $N(t)$ 相对于 t 的变化率。进一步，令 $\theta = \log f(y)$，我们可以对该泊松过程[13]建立对数似然函数：

$$L(d, \theta) = \int_0^r \log \lambda(t) \mathrm{d}N(t) - \int_0^r \lambda(t) \mathrm{d}t .$$

则 $L(d, \theta)$ 它满足以下似然方程：

$$\frac{\partial L}{\partial \theta} = \int_0^r \mathrm{d}N(t) - \int_0^r \lambda(t) \mathrm{d}t = N(r) - e^{\theta} V(d) R^d = 0$$

$$\frac{\partial L}{\partial d}=\left(\frac{1}{d}+\frac{V'(d)}{V(d)}\right)N(r)+\int_0^r \log t dN(t)-e^{\theta}V(d)r^d\left(\log r+\frac{V'(d)}{V(d)}\right)=0$$

将上述两个方程联立，求解得[14]：

$$\hat{d}_r(y)=\left[\frac{1}{N(r,y)}\sum_{j=1}^{N(r,y)}\log\frac{r}{T_j(y)}\right]^{-1}$$

而在实际的计算中，我们都是通过近邻点的个数 $k\in\{k_1,k_2,\cdots k_N\}$ 取得邻域，这要比取以 y 为中心，r 为半径的球形 $S_y(t)$ 邻域方便的多[15]，从而进一步有：

$$\hat{d}_k(y)=\left[\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(y)}{T_j(y)}\right]^{-1}$$

对于我们取定的近邻点的个数 $k\in\{k_1,k_2,\cdots k_N\}$，将上式中的 y 遍历 $y_1,y_2,\cdots y_N$，便可以得到映射嵌入后的低维空间的 N 个局部邻域的本征维数[16]的估计值 $\hat{d}_k(y)$。在此基础上，我们以每个局部邻域中近邻点的个数 $k\in\{k_1,k_2,\cdots k_N\}$ 所占整个映射嵌入后的低维空间中的采样点个数 N 的百分比 $\frac{k_i}{N}$ $i=1,2,\cdots,N$ 为权重，对映射嵌入后的低维空间的 N 个局部本征维数的估计值进行加权平均，即得到整个映射嵌入后的低维空间的全局本征维数的估计值：

$$\hat{d}=\sum_{i=1}^{N}\frac{k_i}{N}\hat{d}_{k_i}(y_i)$$

基金项目

国家自然科学基金青年基金项目(项目编号：11301493，项目名称：完备 Ricci 孤立子上的几何估计与几何结构及 Ricci 孤立子分类问题的研究)。

参考文献 (References)

[1] 罗芳琼. LLE 流行学习的若干问题分析[J]. 现代计算机, 2012(3): 13-16.

[2] 张兴福. 基于流行学习的局部降维算法研究[D]: [博士学位论文]. 哈尔滨: 哈尔滨工程大学, 2012.

[3] 肖健. 局部线性嵌入的流形学习算法研究与应用[D]: [硕士学位论文]. 长沙: 国防科学技术大学, 2005.

[4] Roweis, S.T. and Saul, L.K. (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, **290**, 2323-2326. https://doi.org/10.1126/science.290.5500.2323

[5] 杨志伟, 黄秀云. 基于 LLE 的数据降维方法研究[J]. 中小企业管理与科技旬刊, 2014(25): 197-200.

[6] 高小方. 流行学习方法中的若干问题分析[J]. 计算机科学, 2009, 36(4): 25-28.

[7] 文贵华, 江丽君, 文军. 邻域参数动态变化的局部线性嵌入[J]. 软件学报, 2008, 19(7): 1666-1673.

[8] Kouropteva, O., Okun, O. and Pietikäinen, M. (2002) Selection of the Optimal Parameter Value for the Locally Linear Embedding Algorithm. *Scandinavian Conference on Image Analysis*, **3540**, 359-363.

[9] 邵超, 万春红, 赵静玉. 流形学习算法中邻域大小参数的递增式选取[J]. 计算机工程, 2014, 40(8): 194-200.

[10] 刘建. 高维数据的本征维数估计方法研究[D]: [硕士学位论文]. 长沙: 国防科学技术大学, 2005.

[11] 惠康华, 李春利, 王雪扬, 许新忠. 基于流行学习的本质维数估计[J]. 计算机科学, 2012, 39(s3): 212-214.

[12] Levina, E. and Bickel, P.J. (2004) Maximum Likelihood Estimation of Intrinsic Dimension. *Advances in Neural Information Processing Systems*, **17**, 777-784.

[13] 傅保伟. 基于 MLE 的本征维数估计方法研究[D]: [硕士学位论文]. 长春: 东北师范大学, 2010.

[14] 谷瑞军, 须文波, 刘军伟, 姚娟. 高维数据固有维数的自适应极大似然估计[J]. 计算机应用, 2008, 28(8):

2088-2090.

[15] 谭璐, 吴翊, 易东云. 基于 LLE 方法的本征维数估计[J]. 模式识别与人工智能, 2006, 19(1): 7-13.

[16] Fu, B., Wang, X., Yang, B., Han, Z. and Zhao, X. (2014) A Novel Approach for Intrinsic Dimension Estimation Based on Maximum Likelihood Estimation. *Lecture Notes in Electrical Engineering*, **163**, 91-97. https://doi.org/10.1007/978-1-4614-3872-4_12