第26卷，第1期
2006年1月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 26, No. 1, pp182-186
January, 2006

一种基于非线性降维求正常星系红移的新方法

许 馨[1,2]，吴福朝[1]，胡占义[1]，罗阿理[2]

1. 中国科学院自动研究化所国家模式识别实验室，北京 100080
2. 中国科学院国家天文台，北京 100012

摘 要 提出了一种确定正常星系红移的有效方法，该方法分为以下3个步骤：(1)利用四阶小波系数作为正常星系的特征表示，它能较好地反映吸收线、跳变点和吸收带的信息；(2)利用非线性降维方法LLE(locally linear embedding)将特征数据映射到三维空间中一维流形；(3)由一维流形上的红移分布数据，根据最近邻方法得到正常星系的红移值。实验表明，文中所给的方法较文献中通常使用的PCA方法对于红移的确定具有更高的精度。

主题词 正常星系；红移；LLE(locally linear embedding)；流形；PCA(principal component analysis)
中图分类号：TP29 **文献标识码**：A **文章编号**：1000-0593(2006)01-0182-05

引 言

星系是众多天体的重要组成部分之一。通过天文望远镜能够观测到大量早型星系。目前，我国正在建设“大天区面积多目标光纤光谱望远镜”(LAMOST)的目标是观测10^7的星系光谱和10^5的类星体光谱，极限星等为20.^{m}5。由此建立一个庞大的天文数据库，将为天文学家做前沿课题研究提供丰富的资源。

红移是所有河外天体的重要的参数之一，通过红移可以计算天体的距离，就可以研究天体的各种物理性质，如：质量、大小、光度、爆发规模等[1]。同时，河外天体的红移也是天文学家研究星系和宇宙大尺度结构的基础。星系红移自动测量的传统方法是用观测得到的光谱和已有的光谱模板进行交叉相关[2,3]。Glazebrook等[4]用PCA方法(主分量分析方法)构造模板后，与观测光谱作交叉相关获得红移值，它需要准确的连续谱和光谱模板。文献[5]利用光谱400 nm跳变点进行谱线证认，通过已证认的谱线计算出红移。文献[6]通过估计红移候选的密度，选取具有最大密度估计值的红移候选为光谱的红移。

由于正常星系吸收线特征在低信噪比下容易与噪声混淆，再加上低分辨率下的谱线混合使正常星系红移的确定相对于活动星系来说更加困难。针对在将来的LAMOST星系巡天中存在大量的低信噪比的吸收线星系，本文提出了利用一种非监督的方法，同时也是一种非线性降维的方法LLE(locally linear embedding)来求光谱的红移。为了提取光谱中含有的谱线、正常星系的跳变点及吸收带的信息，我们利用小波变换的小波系数来表示一条光谱，即用小波变换来提取光谱的特征。长期以来，在光谱的处理中，线性的PCA[7]方法被广泛应用，一方面它可以起到特征提取的作用，另一方面它可以对高维数据进行降维。但是对于数据本身中包含的非线性信息，它无法以简单的形式表达出来。Roweis和Saul[8]提出了一种非线性降维的方法LLE，能够将数据从高维降到低维，并在低维空间中仍能保持数据的局部几何结构。将LLE用于一维光谱信号的处理，发现在低维空间正常星系红移的连续变化可以一维流形的简单形式表示。

1 特征提取

对于光谱数据可以用一些特定的谱线参数作为特征，如：谱线的线心深度(R_c)、极大值一半处的全宽(FWHM)、等值宽度(Equivalent width)、平均相对强度I、特征谱线最大相对强度I_h、特征谱线的特征波长λ、特征谱线辐射强度度量C等。但是，特征谱线参量的自动测量技术依赖于连续谱的精确提取等一系列问题。我们希望能找到一种不用精确提取连续谱同时又能够提取出谱线信息等特征的方法，而小波变换正适用于我们的目的。

正常星系的光谱数据[图1(a)]标出了部分谱线和分子吸收带等特征。可以看出，光谱信号是具有突变性的信号。针对这种特征，我们使用具有紧支撑的二次样条函数作为基函数的小波变换。小波函数相应的滤波器的传递函数为[9]

$$H(\omega) = e^{i\omega/2}[\cos(\omega/2)]^3,\ G(\omega) = 4ie^{i\omega/2}\sin(\omega/2)$$

收稿日期：2004-06-30，**修订日期**：2004-11-15
基金项目：国家“863”项目(2003AA133060)资助
作者简介：许 馨，女，1974年生，中国科学院自动化研究所博士研究生

正常星系光谱经过小波变换，在 4 个尺度上的小波系数如[图 1(b)]所示。正常星系中的吸收线、400 nm 处的跳变、发射线及吸收带都对应着小波系数相应的较大变化。考虑到噪声的影响，尺度越大对噪声的抑制越强，因此我们选择了尺度 4 的小波系数作为正常星系特征的表示。

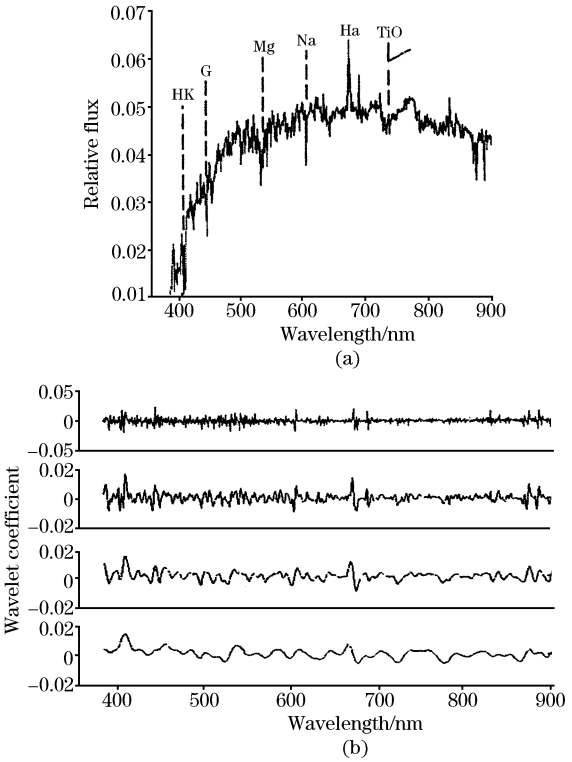

Fig. 1　Normal galaxy (a) and wavelet coefficient (b) (from top to bottom: scale 1, 2, 3, 4)

2　非线性降维

对于高维数据，为了便于处理，一般要进行数据降维，降维的同时也可以凸现数据特征。常用的线性降维方法是 PCA[5] 方法，依据特征向量的方差贡献率，提取主要特征向量构造特征空间，再将原始数据映射到特征空间得到降维后的数据。但是线性的方法在处理非线性结构的数据时效果一般。一些非线性方法如 SOM[7] 计算复杂，有较多的参数需要调整，而且结果也存在收敛性问题。Roweis 和 Saul[8] 在 2000 年提出了一种新的非线性降维方法 LLE。将 LLE 用于正常星系的降维处理，在三维空间中可以得到正常星系数据按照红移大小顺序排列成一条简单的空间曲线，而用 PCA 方法却得不到这样好的效果。

LLE 算法叙述如下：设 D 维空间中有 N 个数据属于同一流形，记做：$X_i = [x_{i1}, x_{i2}, \cdots, x_{iD}]$，$i = 1 \sim N$。假设有足够的数据点，并且认为空间中的每一个数据点可以用它的 K 个近邻线性表示，代价函数为

$$\varepsilon(W) = \sum_i \left| X_i - \sum_j W_{ij} X_j \right|^2 \qquad (1)$$

并且权值 W_{ij} 要满足两个约束条件：(1) 每一个数据点 X_i 都只能由它的邻近点来表示，若 X_j 不是近邻点，则 $W_{ij} = 0$；(2) 权值矩阵的每一行的和为 1，即：$\sum_j W_{ij} = 1$。这样，求最优的 W_{ij} 就是对于公式(1) 在 2 个约束条件下的求解最小二乘问题。权值 W_{ij} 体现了数据间内在的几何关系，并具有平移不变性，尺度不变性和旋转不变性。保持权值 W_{ij} 不变，在低维空间 $d(d \ll D)$ 中对原数据点重构。设低维空间的数据点为 Y_i，可以通过求最小的代价函数(2)来得到。

$$\Phi(Y) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|_0^2 \qquad (2)$$

公式(2)的最优解需要满足下面的约束条件：(1) $\sum_i Y_i = 0$；(2) $\frac{1}{N}\sum_i Y_i Y_i^T = \boldsymbol{I}$，$\boldsymbol{I}$ 为单位矩阵。算法中的可调参数有 2 个：近邻点 K 和维数 d。W_{ij} 和 Y_i 都可以用线性代数中的相应方法求出。对数据进行 LLE 变换时应注意以下 4 点：(1)首先数据点都是属于同一流形的；(2)寻找近邻点的规则要合适，能使最相似的点聚集；(3)近邻点的个数也要恰当，过多或过少都不能正确反映数据间的几何特征。(4)实际应用中，测试数据的分布密度也很大程度影响 LLE 的结果。

3　求红移值

利用 LLE 将高维空间的正常星系的数据降到三维空间，仍可以保持数据间的几何关系。在三维空间中数据形成简单的一维流形，流形上的点按照红移从大到小的顺序进行排列。设在低维空间中，任意一个测试点 Y，与训练集 $T_i(i = 1, 2, \cdots, M)$ 属于同一个流形，Y 的红移值由下述方法确定：(1) 求 Y 的 2 个最近邻 T_j，T_k，即

$$\| Y - T_j \| = \min_{1 \leqslant i \leqslant M} \| Y - T_i \|,$$

$$\| Y - T_k \| = \min_{1 \leqslant i \leqslant M,\ i \neq j} \| Y - T_i \|$$

(2)令 $V_1 = \| Y - T_j \|$，$V_2 = \| Y - T_k \|$，Y 的红移值 Z_y 由下面的公式求出

$$Z_y = \frac{V_1}{V_1 + V_2} Z_j \frac{V_2}{V_1 + V_2} Z_k \qquad (3)$$

其中，Z_j 为 T_j 的红移值，Z_k 为 T_k 的红移值。

4　实验结果

训练样本来自 Kinney[10] 在其文章中构造的星系的模板，选取其中的静止模板 Ellipticals，s0，sa，sb 作主分量分析，取方差贡献最大的第一个主分量作为正常星系模板，它代表了正常星系的 4 个模板中的最主要的特点。由红移公式：$z = \frac{\lambda - \lambda_0}{\lambda_0}$，得到

$$\lambda = \lambda_0 (1 + z) \qquad (4)$$

其中，z 为红移值，λ_0 为静止波长，λ 为观测波长。

给定红移的范围为 0～0.5，红移模拟的步长为 0.001，利用公式(4)对模板进行红移模拟，得到各个红移值下的模拟光谱共计 501 条。待测红移样本来自于 SDSS 数据库中 0266～0280 天区的正常星系的观测数据共 4 782 个。

将模板数据进行小波变换后再用 LLE 方法进行降维处理，所得结果如图 2(a)所示。横轴代表 501 条光谱的编号，第一条光谱红移为 0，第二条光谱红移增加 0.001，依次递增，到 501 条光谱红移为 0.5；纵轴为 LLE 第一个分量。可以看到，曲线是单调的，这意味着在 LLE 变换后正常星系的红移在一维流形上呈单调变化趋势。图 2(b)所示的是模板数据小波变换后用 PCA 降维提取第一个主分量的结果，有多个不同红移的样本对应同一个 PCA 主分量系数，这说明对于红移的变化表达，PCA 并没有 LLE 那样的良好特性。

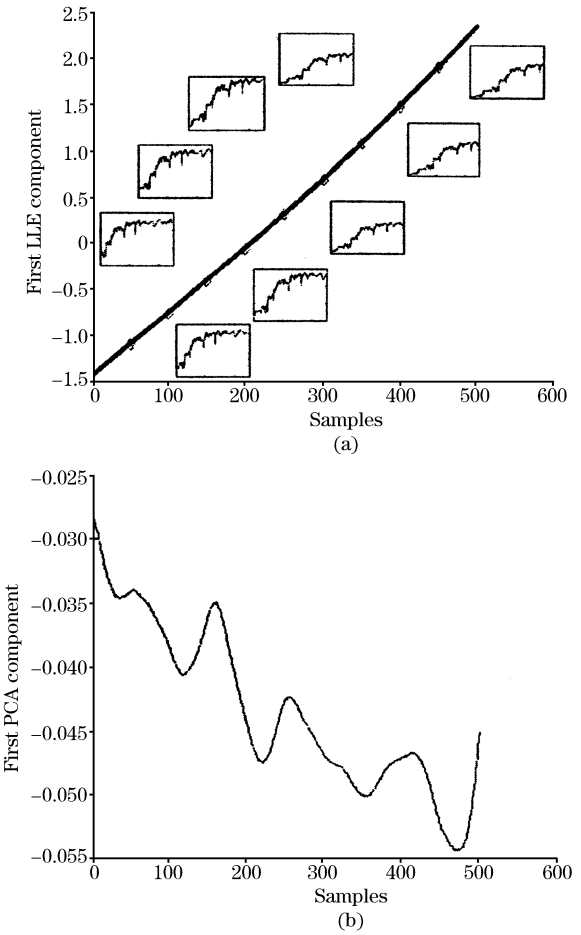

Fig. 2 501 samples *vs.* corresponding first LLE component (a); 501 samples vs. corresponding first PCA component (b)

取 0266 天区的数据和 501 个训练样本作 LLE 变换(参数 $K=4$，$d=3$)所得到的三维空间中的一维流形如图 3(a)所示，点 a 为红移值最小点，沿着曲线的方向红移呈增大的趋势，到点 b 为红移值最大。相同的数据作 PCA 变换所得结果如图 3(b)所示，可以看出，LLE 这种非线性变换确实可以在低维空间中以简单的形式凸现正常星系光谱数据的红移变化特征，而 PCA 线性变换难以清楚简单地表达红移的变化特征。

经 LLE 和 PCA 变换后的数据按照公式(3)求得的红移值与 SDSS 给出的红移参考值的对比分别如图 4(a)，(b)所示，可以看出 LLE 降维的低维空间中包含的信息要高于 PCA，且得到的红移值准确度大大高于 PCA 降维。图 3(a)中有个别点偏离了曲线，称为离群点。这些离群点的红移值在图 4(a)中表现为红移值偏差较大的点。这些离群点的产生是由于在作 LLE 变换时，首先要找每一个数据点的 K 个近邻点，使用的决策是欧氏距离最近。而根据这个决策，离群点找到的 K 个近邻有可能红移值相差很大，或者是正常星系数据集中混入了其他类型天体的数据。

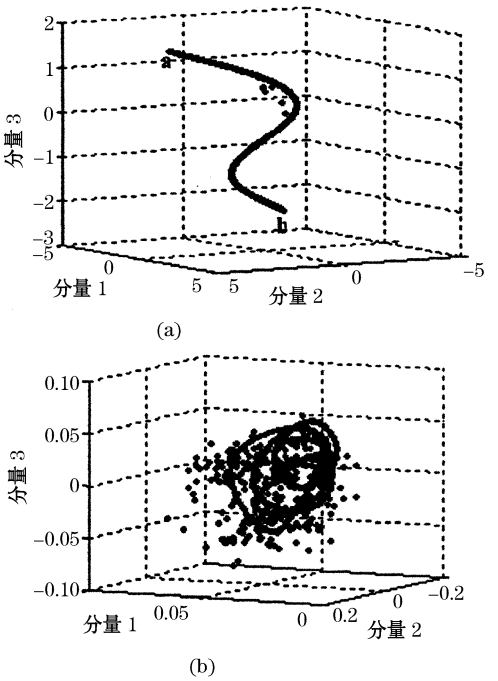

Fig. 3 3-D space after LLE transformation (a) and 3-D space after PCA transformation (b)

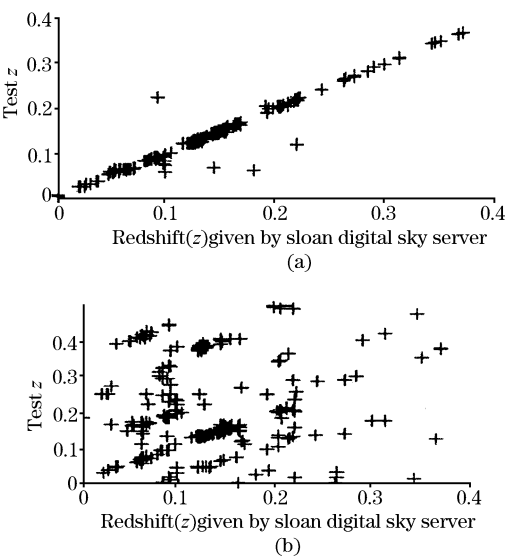

Fig. 4 Test redshifts *vs.* SDSS reference redshifts on 0266 square samples after LLE transformation (a) and PCA transformation (b)

15 个天区的 4 782 条光谱特征提取后用上述方法求出的

红移值与 SDSS 给出的红移参考值的对比如图 5 所示。4 782 条光谱的误差绝对平均值为 0.002 8，方差为 0.013 1，这说明本文所给出的求红移的方法具有非常好的数值精度。

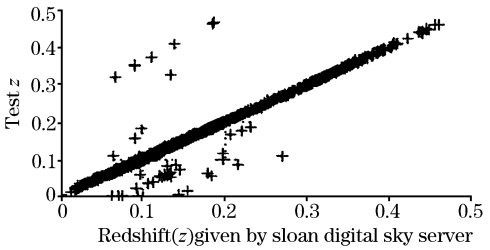

Fig. 5　Test redshift *z vs*. SDSS redshift *z* on all test samples

图 6 所示为红移的绝对误差分布的直方图。可以看到，大多数的数据点的误差分布在 0～0.005 之间。

5　总　结

本文首先利用小波变换提取正常星系的吸收型的特征，并消除了连续谱的影响，然后利用 LLE 非线性降维的方法将正常星系的特征数据表示为三维空间中简单形式的一维流形；最后由训练样本的红移在这个一维流形上的分布根据最近邻方法获得实测光谱的红移值。本文所给出的求红移方法可应用于 LAMOST 连续谱未定标的光谱数据。

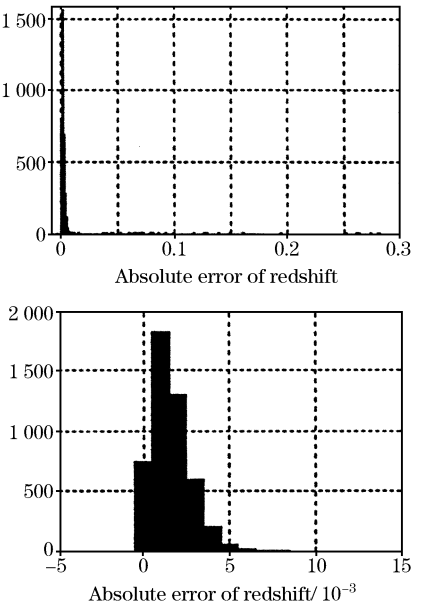

Fig. 6　Histogram of absolute error of redshift(left) and zoom in between the 0-0.01 redshift(right)

参 考 文 献

[1] ZHU Ci-sheng(朱慈盛). Astronomy Tutorial(天文学教程：下册). Beijing：Higher Education Press(北京：高等教育出版社)，1987. 280.
[2] Tonry J，Davis M. Astronomical Journal，1979，84：1511.
[3] LIU Rong，DUAN Fu-qing，LUO A-li(刘 蓉，段福庆，罗阿理). Spectroscopy and Spectral Analysis(光谱学与光谱分析)，2005，25(7)：1155.
[4] Glazebrook Karl，et al. Astrophysical Journal，1998，492：98.
[5] LUO A-li，ZHAO Yong-heng(罗阿理，赵永恒). Spectroscopy and Spectral Analysis(光谱学与光谱分析)，2001，21(1)：19.
[6] DUAN Fu-qing，WU Fu-chao，LUO A-li，et al(段福庆，吴福朝，罗阿理，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析)，2005，25(11)：1895.
[7] YAN Ping-fan，ZHANG Chang-shui(阎平凡，张长水). Artificial Neural Networks and Evolutionary Computation(人工神经网络与模拟进化计算). Beijing：Tsinghua University Press(北京：清华大学出版社)，2000. 212.
[8] Roweis S，Saul L. Science，2000，290：2323.
[9] FENG Xiang-chu，GAN Xiao-bing(冯象初，甘小冰等编著). Numerical Functional and Wavelets Theory(数值泛函与小波理论). Xi'an：Xidian University Press(西安：西安电子科技大学出版社)，2003. 61.
[10] Kinney A L，Calzetti E，Bohlin R C，et al. Astrophysical Journal，1996，467：38.

A Novel Method for the Determination of Redshifts of Normal Galaxies by Non-Linear Dimensionality Reduction

XU Xin[1,2]，WU Fu-chao[1]，HU Zhan-yi[1]，LUO A-li[2]
1. Robot Vision Group，National Laboratory of Pattern Recognition，Institute of Automation，Chinese Academy of Sciences，Beijing 100080，China
2. National Astronomical Observatories，Chinese Academy of Sciences，Beijing 100012，China

Abstract It is difficult to determine the redshifts of normal galaxies (NG) from their spectra because of their common weak absorption property. In the present work，a novel method is proposed to effectively deal with this issue. The proposed method is composed of the following three parts：At first，the wavelet transform coefficients at the fourth scaling are experimentally found to be appropriate and used as our features to represent the absorption information from NG absorption lines，break points，and absorption bands. Then，the features are mapped by a non-linear method，LLE(locally linear embedding)，onto an one-dimensional manifold in the 3D space；Finally，the NG redshifts are obtained by the nearest neighborhood technique from the redshift distribution on the manifold. Besides，the proposed method is compared with widely used PCA method in the literature with SDSS database，and is shown to be more accurate for the redshifts determination.

Keywords Normal galaxies；Redshift；LLE；Manifold；PCA

(Received Jun. 30，2004；accepted Nov. 15，2004)