第25卷,第11期
2005年11月

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

Vol.25,No.11,pp1895-1898
November, 2005

用于红移测量的基于密度估计的模板匹配法

段福庆[1], 吴福朝[1], 罗阿理[2], 赵永恒[2]

1. 中国科学院自动化研究所模式识别国家重点实验室, 北京 100080
2. 中国科学院国家天文台, 北京 100012

摘 要 文章给出了一种基于密度估计的模板匹配法来确定红移, 将确定红移问题转化为寻找密度最大点问题。该方法首先利用基于均值漂移的谱线自动提取方法提取出特征谱线; 再根据提取出的特征波长序列与模板的谱线表, 由红移公式构造出一个数据集 Z; 最后, 寻找数据集中的密度最大点, 对密度最大点的 ε-邻域中的点进行平均得到红移值。该方法利用了特征波长和谱线类型信息, 可以处理各种类型的天体。在构造数据集时忽略谱线类型不匹配及特征波长明显不匹配的情况, 这就去除了很大的干扰并且加快了运行速度。试验结果表明: 该方法的稳定性较好, 正确率也较高。

主题词 光谱分析; 红移测量; 模板匹配; 特征谱线
中图分类号: TN911.7 **文献标识码:** A **文章编号:** 1000-0593(2005)11-1895-04

引 言

天体光谱中蕴含着许多关于天体本身的重要信息如化学元素的丰度、天体的温度等, 因此光谱分析技术在天文学中有很广泛的应用。红移值是河外天体的一个重要物理参量, 由于天体以很高的速度背离地球运动使得观测到的谱线波长比静止的谱线波长要大, 因此就产生了光谱学中的波长向红端移动的所谓红移现象。对于河外天体, 红移测量是光谱分析的首要任务, 通常是由专家根据经验知识通过人机交互的方式完成的。常用的天文软件包如 MIDAS, FIGARO 和 IRAF 均是如此。当光谱的信噪比比较低时, 天文学家也无能为力。在大规模星系光谱巡天中(如 SDSS, 2dF 和 LAMOST),光谱数据数以亿计, 以人工为主的传统的光谱分析方法显然不能满足实际需要, 因此寻求自动的光谱分析方法迫在睫眉。

最早出现的红移自动测量技术是 Tonry 和 Davis 的交叉相关法[1]。这种方法目前被认为是最成功的红移自动测量方法, 其核心是用实测光谱和一系列模板光谱做交叉相关, 在所有交叉相关函数中寻找最大峰, 这个峰的位置和宽度分别决定了红移值和置信度。澳大利亚天文台的 Glazebrook 曾将这种方法进行了推广, 提出了一种基于主分量分析(PCA)的 PCAZ 方法, 它是用 PCA 提取出一组正交模板, 然后用正交模板的线性组合与实测光谱做交叉相关。这两种方法只适合红移较小的天体。国内相关的工作有吴永东最早针对类星体提出的结合形态滤波[2], 样条逼近, 弹性匹配, 以及证据组合的求红移方法; 周虹等基于 Hough 变换和神经网络求红移的方法[3]; 邱波的伪三角法[4]。这些方法中大多数速度相对较慢, 并且只能处理发射线天体。本文从实用的角度出发, 尝试用一种全新的方法来简单有效地处理各种类型天体的红移测量问题。

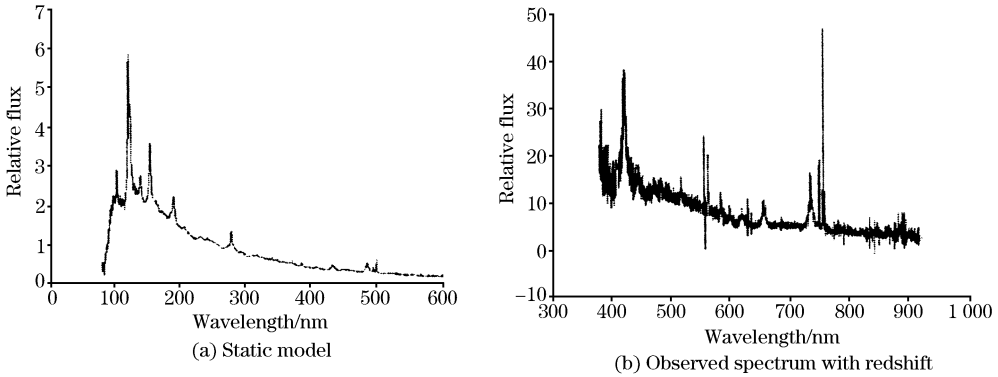
1 原理和方法

设天体静止谱线波长为 λ', 该谱线的观测波长为 λ, 则 $\lambda = (1+z)\lambda'$, 其中 z 就是天体的红移值。图 1 给出了一个静止模板和实测光谱的示例(横坐标为波长, 纵坐标为相对光强)。基于模板匹配求红移的一般过程为: 先提取特征谱线, 再进行谱线证认并确定红移值。

1.1 特征谱线的提取

天体光谱是由连续谱、谱线和各种噪声组成的, 谱线是由天体中的各种原子、分子等在连续谱基础上吸收或辐射能量所体现出的特征, 噪声是叠加在连续谱和谱线之上的。自动提取谱线的过程如下: 首先提取出连续谱, 然后用原始光谱除去连续谱使连续谱归一化, 最后进行光谱去噪处理得到谱线。本文采用基于均值漂移的谱线自动提取方法[7]来提取特征谱线。均值漂移是模式识别中的一种经典方法, 它的主要作用是在特征空间求取模式点也即局部密度最大点。该方法提取特征谱线的基本原理如下: 首先, 利用均值漂移总是指向局部密度最大点这一性质, 通过使用均值漂移过程迭代

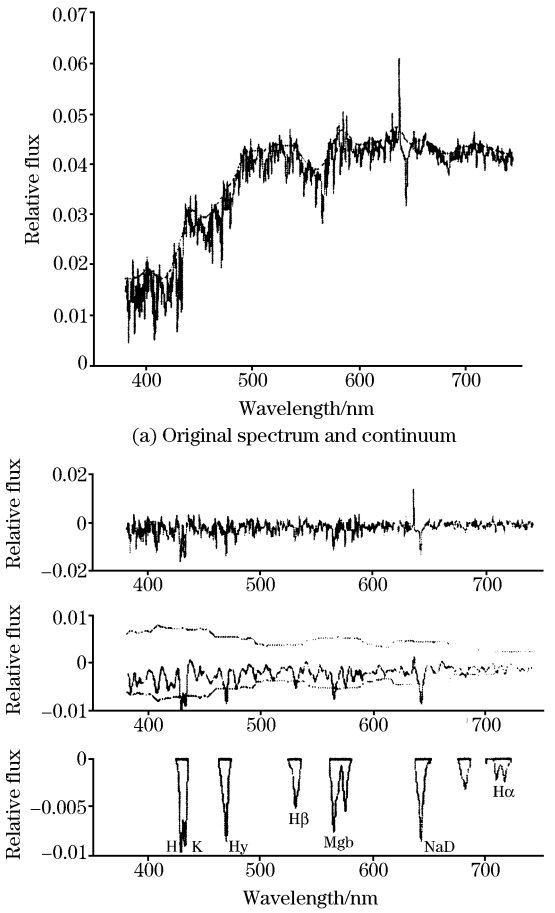
收稿日期: 2004-05-08, **修订日期:** 2004-08-16
基金项目: "863"计划(2003AA133060)和国家重大科学工程 LAMOST 计划资助
作者简介: 段福庆, 1973 年生, 中国科学院自动化研究所博士研究生



(a) Static model (b) Observed spectrum with redshift

Fig.1 Model and observed spectrum

逼近连续谱，它比较有效地解决了目前较常采用的中值滤波、小波变换等方法在光谱发生跳变处和较宽的谱线处不能得到有效处理的问题；其次，在光谱归一化后，采用均值漂移去噪得到谱线光谱；最后，对谱线光谱设置局部阈值提取出特征谱线。图 2 给出了该方法提取谱线的一个实例。图 2(a)为原始光谱和连续谱(红色曲线)，图 2(b)上部为归一化后的光谱，中部为谱线光谱和局部阈值，下部为提取出的特征谱线。



(a) Original spectrum and continuum

(b) Normalized spectrum(up),line spectrum and local thresholds(middle),feature lines(low).

Fig.2 Spectral line extraction

1.2 红移确定

1.2.1 问题描述

在天体光谱中，谱线分为吸收线和发射线两种，它们分别是由于天体中的原子和分子在发生能级跃迁时吸收或辐射能量所体现出的特征。不同的原子和分子有其特定的谱线，特征谱线的线中心所对应的波长就是特征波长。每一类天体的静止光谱模板都有其特定的特征波长序列，用数列 $\{\lambda'_i, i = 1, 2, \cdots, M\}$ 表示。实际观测光谱经过谱线的自动提取处理后，也得到一个特征波长序列，用 $\{\lambda_i, i = 1, 2, \cdots, N\}$ 表示。如果单纯考虑红移的因素，$\{(1+z)\lambda'_i, i = 1, 2, \cdots, M\}$ 应该等同于 $\{\lambda_i, i = 1, 2, \cdots, N\}$，但在实际情况中，两者有很大的差异，这主要由以下两个原因造成：首先由于各种噪声的影响或者谱线自动提取方法本身的缺陷造成某些特征谱线未能提取出来，而且往往提取出许多假谱线；其次，红移引起某些特征谱线移到红外，同时紫外的某些谱线移入可见光区。因此，确定红移的问题转化为寻找 $\{\lambda'_i, i = 1, 2, \cdots, M\}$ 与 $\{\lambda_i, i = 1, 2, \cdots, N\}$ 中特征波长的匹配问题(也即谱线证认)。

1.2.2 方法描述

对于每个观测光谱来说，红移 z 是不变的未知量，所以在 $\{\lambda'_i, i = 1, 2, \cdots, M\}$ 和 $\{\lambda_i, i = 1, 2, \cdots, N\}$ 中，总存在某些匹配对 (λ, λ') 满足红移关系式 $\lambda = (1 + z)\lambda'$。为了便于处理各种类型的天体，我们在上述特征波长序列中加入谱线类型信息，构造二维序列 $\{\lambda'_i, t'_i, i = 1, 2, \cdots, M\}$ 和 $\{\lambda_i, t'_i, i = 1, 2, \cdots, N\}$，其中 $t_i, t'_i = \pm$，(1代表该谱线是吸收线，-1代表发射线)。按照红移公式 $z = \frac{\lambda}{\lambda'} - 1$，我们得到一个数列 $\{z_i = \frac{t_k\lambda_k}{t'_l\lambda'_l} - 1, k = 1, 2, \cdots, N, l = 1, 2, \cdots, M, i = 1, 2, \cdots, M \times N\}$。舍弃 $\{Z_i, i = 1, 2, \cdots, M \times N\}$ 中的负数(负数表明两个谱线的类型不匹配或者谱线类型相同但特征波长明显不匹配)，得到数据集 $Z = \{z_i, i = 1, 2, \cdots, P\}$。很显然，在这个集合中，红移值附近的点一般来说是最密集的，也就是说，红移值处的密度是最大的。我们将数列 Z 进行排序后在二维平面上显示出来，如图3所示，图中线段所示的即为密度最大区。因此，只要找到数据集 Z 中密度为最大的点并对其附近的点进行平均即可得到红移值

的估计。本文采用 *Parzen* 窗法[6] 进行密度估计。定义窗函数 $\phi(u)=\begin{cases}1, & |u|\leqslant\varepsilon\\ 0, & \text{others}\end{cases}$，因为我们的红移误差上界为 0.001，所以 ε 选 0.001，因此密度估计函数为 $\hat{f}(z)=\frac{1}{P}\sum_{j=1}^{P}\phi(z-z'_j)$。令 $\hat{f}(z)=\max\{\hat{f}(z'_i), i=1,2,\cdots,P\}$，$z'$ 即为数据集 Z 中密度最大点，对满足 $|z'-z'_i|\leqslant\varepsilon$ 的所有点进行平均即为红移估计值。

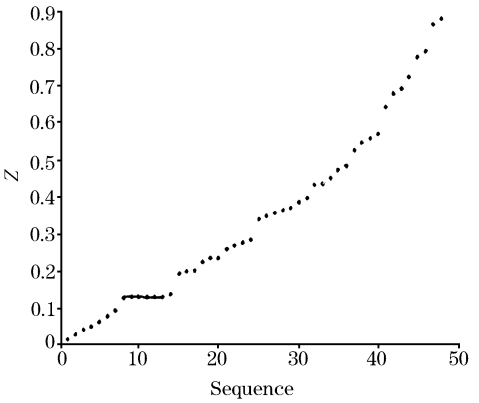


Fig.3　Redshift analysis

2　实验结果

在本节中我们给出了两组实验，第一组为模拟光谱实验，第二组是 SDSS 实测光谱的实验。

2.1　活动星系的模拟实验

图 4 为活动星系的一个标准模板，活动星系为发射线天体，其特征谱线大部分为发射线。首先，随机产生 0.001～1.0 的红移值，对模板进行红移模拟并截取波长段为 380～742 nm 的部分作为测试光谱；然后，对测试光谱叠加不同信噪比的高斯白噪声，并利用本文方法确定红移值，在每个信噪比下进行了 200 次独立试验，得到在红移精度小于 0.001 下的正确率—信噪比曲线如图 5 所示。从图中可明显看出，正确率随着信噪比的增加越来越高，说明本文方法具有很好的稳定性；另外，当信噪比为 8 时，正确率大于 96%，这说明本文方法也具有很好的鲁棒性。

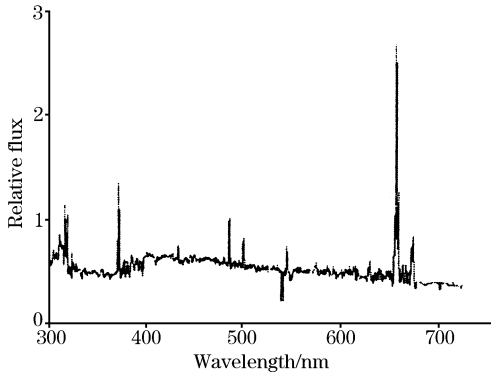


Fig.4　A standard model of active galaxy

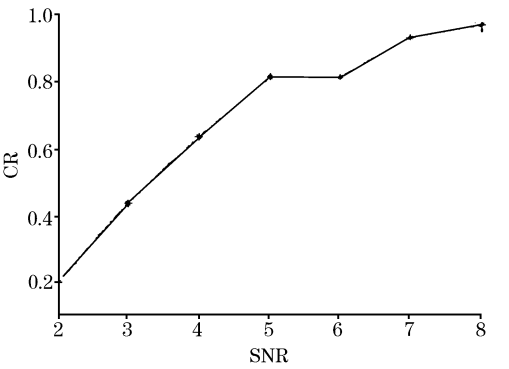


Fig.5　Correct rate curve

2.2　实测光谱的实验

在实验中，我们收集了来自 SDSS 的五个不同天区的所有正常星系的光谱 1 574 个，正常星系中的特征谱线大部分为吸收线，吸收线对噪声比较敏感。这些光谱的信噪比平均为 10 左右，Sloan 给出了它们的红移值。采用本文方法对这组数据进行红移计算，结果如图 6 所示，其中有 105 个与 Sloan 给出的红移值不匹配，匹配率为 93.5%。通过对不匹配的光谱进行分析，发现其中包含了少量的 Quasar 和星爆星系的光谱，其他不匹配光谱的信噪比都特别低，谱线几乎都被噪声淹没。

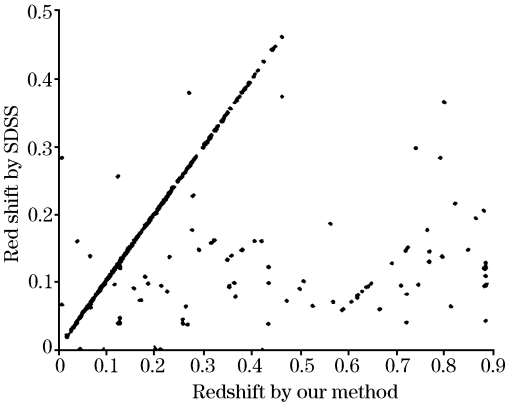


Fig.6　The result of redshift matching

3　结　论

红移是河外天体的一个重要物理参数。本文在仔细分析天体光谱特点的基础上，给出了一种基于密度估计的模板匹配法来求红移，把求红移的问题转化为寻找密度最大点的问题。该方法首先利用基于均值漂移的谱线自动提取方法提取出特征谱线，然后用提取出的特征波长序列与模板的谱线表根据红移公式构造出一个数据集 Z，通过定义窗函数 $\phi(u)$ 来寻找这个数据集中的密度最大点，对密度最大点的 ε- 邻域中的点进行平均得到红移值的估计。在计算过程中，本文利用了特征波长和谱线类型信息。在构造数据集 Z 时，我们不考虑谱线类型不匹配及特征波长明显不匹配的情况，这就去除了很大的干扰，并且加快了运行速度。另外，由于该方法加入了谱线类型信息，使得它可以处理各种类型的天体，而

已有的方法大多只针对某一类天体。实验结果表明：该方法的鲁棒性较好，正确率也较高。

目前本文方法仅使用了光谱的谱线波长信息和类型信息，而没有考虑其他谱线信息如相对强度、等值宽度、线强比等。这些信息有助于解决在出现两个或多个密度最大点时的红移确定问题。这些都将在今后的工作中进行考虑。

本文的工作有一定的创新性，类似的工作可参阅文献[6]。

参 考 文 献

[1] John Tonry, Marc Davis. The Astronomical Journal, 1979, 58(10): 1511.
[2] WU Yong-dong(吴永东). 博士论文. Institute of Antomation, the Chinese Academy of Sciences(中国科学院自动化研究所), 1997.
[3] ZHOU Hong, HUANG Ling-yun, LUO Man-li. J. of Electronics, 2000, 22(4): 529.
[4] QIU Bo, HU Zhan-yi, ZHAO Yong-heng(邱 波, 胡占义, 赵永恒). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2002, 22(4): 695.
[5] BIAN Zhao-qi, ZHANG Xue-gong, et al(边肇祺, 张学工, 等). Pattern Recognition(模式识别). Beijing: Tsinghua University Press(清华大学出版社), 2000.
[6] XU Xin, LUO A-li, WU Fu-chao, et al(许 馨, 罗阿理, 吴福朝, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(6): 996.

Density Estimation Based Model Matching Method for Redshift Determination

DUAN Fu-qing[1], WU Fu-chao[1], LUO A-li[2], ZHAO Yong-heng[2]
1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
2. National Observatory of Beijing, Chinese Academy of Sciences, Beijing 100012, China

Abstract The present paper proposes a model matching method based on density estimation for redshift determination, in which the problem of redshift determination is translated into the problem of searching for the point of maximum density within a data set. At first, the mean shift-based method for auto-extraction of spectral lines is used to get feature spectral lines. Secondly, according to the redshift formula, the authors use the feature wavelength array and the spectral template to get a data set. Finally, the authors find the point of maximum density within the data set, then the average of the data in ε-neighbor of the point is regarded as the redshift estimation. The information of feature wavelength and spectral line type is used in this method so that it can deal with every kind of spectra. Experiments show that our method is stable and the correct identification rate is high.

Keywords Spectral analysis; Redshift determination; Model matching; Feature spectral line

(Received May 8, 2004; accepted Aug. 16, 2004)