文章编号: 1000-8349(2008)03-0266-12

述评

测光红移算法概述

王 丹, 张彦霞, 赵永恒
(中国科学院 国家天文台, 北京 100012)

摘要: 随着大规模多色巡天项目的完成, 测光红移已被视为研究宇宙大尺度结构以及星系形成和演化的有效工具。该文介绍了测光红移的背景、现状、算法及其在天文学中的应用。综述了九种较为常用的估测测光红移的算法, 包括 HyperZ 、颜色 − 星等 − 红移关系法 (CMR) 、多项式回归、基于 Kd 树的多项式回归、贝叶斯方法、支持向量机 (SVMs) 、人工神经网络 (ANNs) 、最近邻或 K 近邻方法和核回归等。着重讨论和比较了这些算法的效果和性能, 同时也对它们的优缺点进行了阐述, 并对未来的测光红移算法研究进行了展望。

关 键 词: 天体物理学; 测光红移; 距离; 统计方法

中图分类号: P141.2 **文献标识码**: A

1 引 言

测光红移是指利用中波段和宽波段的测光数据或者图像得到的红移, 因此, 测光红移主要是由星系的颜色决定的。除了颜色外, 测光红移还可以用角大小或者聚集度指数等参数来衡量。

"测光红移"并不是一个新名词, 它最早出现在 Puschell 等人 (1982) [1] 的文章中。Puschell 利用宽带测光数据估测暗射电星系的红移, 在三个方面起到了先驱作用: (1) 使用了近红外 (JHKs) 和光学波段 (RI) ; (2) 采用了 χ^2 最小的谱能量分布; (3) 在估测红移过程中, 应用了不同类型的模板。模板包括未演化的本星系, 来自于 Bruzual 的 SED 理论模板, 以及从已知射电星系得到的 SED 观测模板。

Loh 和 Spillar(1986) [2] 第一次在文章题目 (Photometric Redshifts of Galaxies) 中使用了"测光红移" 的字样。他们工作的闪光点在于使用电荷耦合器件 (Charge Coupled Device , 简称 CCD) 可以观测到星等 $I = 21.5$ mag 的星系, 同时利用 6 个中波段的滤光片和 χ^2 最小的模

收稿日期: 2007-06-12 ; **修回日期**: 2007-09-20

基金项目: 国家自然科学基金资助项目 (10473013 , 10778724 , 90412016)

板匹配方法得到了测光红移。但是他们所用方法的不足之处在于只使用了 3 个本星系的 SED 模板来代表所有红移值下的星系类型。

在 20 世纪 30 年代末， Vandererkoven [3] 就已经从理论上证明可以根据星系光谱在 5 000 Å 附近的倾斜度估算红移。但真正将多波段测光方法应用到红移估测工作中的是 Baum 。 1957 年， Baum [4] 提出利用测光数据研究红移，并于 1962 年开发了一种估计测光红移的算法 [5] ，即使用光电光度计和 9 个滤光片，这 9 个滤光片覆盖了从 3 730 Å 到 9 875 Å 的光谱范围。利用这个系统，他获得了 6 个比较亮的位于室女星系团 (Virgo) 的椭圆星系的谱能量分布，随后又得到了 C1095+2044 星系团 (又称为 Abell 0801) 中的 3 个椭圆星系的谱能量分布。利用波长的对数尺度，他将平均的室女星系团的谱能量分布与平均的 C10925 星系的谱能量分布画在同一张图上进行比较，算出两个能量分布之间的位移，从而可以获得 C1095+2044 星系团的红移。利用这种方法估测的 C1095+2044 星系团的红移 $z = 0.19$ ，这与利用光谱方法测得的红移 $z = 0.192$ 十分接近， Baum 继续扩展这种方法到那些未知红移的星系上去，星系的红移范围可以达到 $z = 0.46$ 。 Baum 的这种方法用于测量红移时比较精确，但是由于其依靠在 4 000 Å处的光谱截断特征来估测红移，所以只适合用于椭圆星系。

Koo [6] 在 1985 年采用了一种新的不同于 Baum 的方法来测量红移。首先，他利用照相底片代替光度计，这种方法可以在同一时间内得到大批星系的红移。其次，他用 4 个滤光片 (UJFN) 代替了 Baum 使用的 9 个滤光片。另外，他没有使用经验的谱能量分布，而是使用了 Bruzual 的理论模板 [7] ，这种模板对所有的星系类型都适用。除了上述提到的几点不同外，Koo 与 Baum 方法最主要的不同在于对颜色的使用上： Baum 是将颜色转化成低分辨率的光谱，而 Koo 则是将 Bruzual 的模板转化成颜色。

相对于光谱红移来说，测光方法估测红移的优点在于速度较快。用光谱观测的方法测量红移时，为了得到高信噪比的光谱，通常都需要较长的积分时间。但是对于测光来说，只需要很短的曝光时间就可以得到与光谱观测相同的信噪比。图像探测器覆盖的天区比多目标摄谱仪的大得多。这就意味着用测光的方法可以同时得到许多星系的红移。测光红移已经被视为研究星系的统计属性和演化规律的有效方法，它可以将一些观测参数 (例如：颜色、星等) 转化成星系的物理属性 (例如：红移、类型和光度) 。与光谱测红移的方法相比较，测光方法估测红移的最大缺点是精度较低，但测光红移适用于高红移、大样本数据。

测光红移已经广泛地应用到天文学各领域的许多科学研究上，并已迅速演化成研究主流观测宇宙学的重要工具。目前，测光红移在下列研究中显示出其独有的重要性：

(1) 通过观测 U 波段的 Lyman 跳变来研究红移 $z > 3$ 的原始星系 [8−10] ；

(2) 研究高红移类星体或远距离射电星系 [11] ；

(3) 研究场星系的演化或者星系的光度函数 [12−15] ；

(4) 区分星团和超星团 [16,17] ；

(5) 估测宇宙的几何结构 [3] ；

(6) 研究光度密度的演化和宇宙早期的大质量星系的数目 [18] ；

(7) 研究星系尺度的演化 [19,20] ；

(8) 确定宇宙中重子和物质密度 [21] ；

(9) 研究在 SDSS 巡天的图像数据中亮红星系的聚类 [22] ；

(10) 在微引力透镜上的应用来研究物质的分布 [23] 。

在近十年中，正在进行和已经完成的巡天项目，天文数据正以指数形式增长，天文学界面临着数据雪崩。精确的图像与光谱巡天项目，例如： SDSS 巡天 (Sloan Digital Sky Survey，简称 SDSS [24]) 、 VLT/VIRMOS 巡天 [25] 、 VST 巡天、 Keck DEEP2 巡天 [26] 等为研究宇宙的起源与演化提供了大量丰富的数据资源。为了更有效地使用这些数据集，需要开发一批有效的自动化分析工具。对于研究测光红移而言，有必要探讨各种测光红移算法，并研发相应的工具，从而帮助天文学家选取精确的、有效的测量红移的算法。对于 SDSS 巡天而言，它提供了一亿多个星系的精确测光数据，但是只对其中一百万个星系进行了光谱观测，获得了这些星系的光谱红移。对于其他的无光谱观测的星系的红移则是未知的，如果能找到行之有效的方法，利用 SDSS 大量的测光数据估测星系的红移，这将对研究星系的形成与演化和宇宙大尺度结构都具有划时代的意义。

2 测光红移的算法

随着天文数据量的指数增长，数据以 TB 量级，甚至 PB 量级计量，天文数据覆盖了各个电磁波段，天文学已步入全波段天文学时代。如此丰富的数据为各种算法的研究提供了很好的实验床，相应的数学、计算机科学、统计学、机器学、人工智能、数据库等学科的飞速发展，为新算法的出炉奠定了夯实的基础。其他相关领域的创新成果可以很方便地应用到天文学中。目前，多种方法已成功地应用到估测测光红移问题上，常用的方法分为两类：模板匹配方法和训练集方法。

模板匹配方法也就是谱能量分布 (Spectral Energy Distribution，简称 SED) 拟合方法。在 SED 拟合方法中，首先需要建立一系列模板，每个模板都是经过红化校正的，并且经过了消光改正。将经过上述操作后得到的颜色与实际观测得到的星系的颜色进行对比。通常当 χ^2 值最小时，就认为得到了该星系的红移。这种测红移的方法简单并且计算量较小，在现在的高性能计算机上很容易实现。最典型 SED 拟合方法的应用是 HyperZ 。 SED 的模板分为两类：一类来自于星族合成 (例如： Bruzual 和 Charlot [27] 的工作)；另一类是从真实星系的光谱中得到的，包含了不同星系的形态和光度 (例如： Coleman 、 Wu 和 Weedman 等人 [28] 的工作，简称 CWW) 。这两种模板都有自身的缺点：来自于星族合成的模板可能包括不正确的参数或者未包括一些已知的参数信息；那些来自于真实星系的模板几乎都是由亮的低红移星系得到，因此较难找到高红移星系的模板。

训练集方法是以机器学习为基础，采用统计理论估测红移的方法。这种方法需要一个有代表性的训练集，其中包括了星系的测光数据及光谱红移值。光谱红移用来做为约束，使测光数据与光谱红移之间建立一种拟合关系。这种方法的缺点在于不能应用到纯测光数据上。而且，它不具有外推的能力，即不能超越训练集样本的限制。如果训练样本不够大且不完备，用这样的训练样本得到的回归器估测新样本红移时，极易产生偏差。然而，训练集方法有其特有的优势：它可以根据数据的信息，自动拟合，不需要额外的星系形成和演化信息。兼顾其优缺点，训练集方法特别适用于两种数据集的联合，例如 VLT/VIRMOS 巡天和 Keck DEEP2 巡天，

这两个数据集的联合可以提供超过十几万星系的光谱红移。 SDSS 巡天提供了丰富的光谱红移，可以作为理想的数据集。目前，应用最广泛的训练集方法，如 Brunner [29] 、 Wang [30] 和 Budavari [31] 的多项式回归方法； Firth [32] 和 Tagliaferri [33] 的人工神经网络 (Artificial Neural Networks，简称 ANNs)； Wadadekar [34] 的支持矢量机方法 (Support Vector Machines，简称 SVMs) 等。

在训练集方法中，还有一种不需要训练的估测红移方法 —— 事例学习方法，又称懒学习方法。事例学习方法独特之处在于不需要训练过程。所有的训练样本都存放于内存中，当新的测试样本输入时，测试样本需要遍历内存中的所有的训练样本来找到符合条件的样本点，通过计算这些样本点的加权平均值来获得测试样本的红移。事例学习方法包括最近邻、 K 近邻、局部加权回归和核回归方法等。

下面对文献中已用到的各种测光红移方法的原理和应用进行简要的介绍：

(1) HyperZ 方法 [35]

HyperZ 方法是谱能量分布方法的典型应用 [35] ，它是基于对光谱整体轮廓的拟合，即主要依赖于对 Lyα 森林、 Balmer 跳变这类显著光谱特征的探测。拟合过程是通过与从同一测光系统得到的光谱模板进行比较来实现的。 HyperZ 方法中的模板来自于实际观测或星族合成。利用 χ^2 最小化的方法，即计算星系的 SED 与同一系统下得到的星系模板之间的差别，χ^2 值取最小的模板对应的红移即被确认为该天体的测光红移。

$$\chi^2 = \sum_{i=1}^{N_{\rm filters}} \left[\frac{F_{{\rm obs},i} - b \times F_{{\rm temp},i}(z)}{\sigma_i} \right]^2$$

其中 $F_{{\rm obs},i}$ 、 $F_{{\rm temp},i}$ 和 σ_i 分别为滤光片 i 中的观测流量、模板流量及测量误差， b 为归一化常数， $N_{\rm filters}$ 为观测使用的滤光片数目。 Csabai 等人 (2003) [36] 用 CWW 模板估测红移的剩余标准偏差 $\sigma_{\rm rms} = 0.066\ 6$ ，用 Bruzual-Charlot 模板测量红移的剩余标准偏差 $\sigma_{\rm rms} = 0.055\ 2$ 。 Mobasher 等人 (2007) [37] 采用了 χ^2 最小化的方法估测了 COSMOS 巡天的测光红移值。

(2) 颜色 – 星等 – 红移关系法

另一种估测测光红移的方法称为颜色 – 星等 – 红移关系法 (Color-Magnitude-Redshift Relation，简称 CMR) 。众所周知，星系的红移不仅与它们的颜色和光谱型有关，而且也与星等有直接的关系。并且星系的测光属性与红移之间不是简单的线性关系，因此不能用简单的线性关系来描述。 CMR 方法构建了星等、颜色和红移的矩阵来数字化三者之间的关系。下面我们以王丹等人 (2006) [38] 的工作为例来详细介绍 CMR 方法。首先，按照 SDSS 巡天中的 r 星等值将样本分成 7 个子区间 r_1 到 r_7 ，然后每个星等区间画两张双色图，分别为 u-g $vs.g$-r 图和 g-r $vs.r$-i 图。这样就产生了 14 张双色图。将每张双色图等分成 400×400 个格子。将训练样本按照不同的 r 星等值在双色图中找到相应的格子，所有的训练样本都找到与之相对应的格子，当 1 个格子内落入的星系数目超过 25 个时，计算落入该格子的训练样本光谱红移的中值，并用此中值作为该格子的红移。如果落入 1 个格子中的样本数目少于 25 个时，将格子的范围扩大成 2 个格子的大小，再计算此时 2 个格子中的样本数目，如果超过 25 个，计算中值。如果落入此两个格子中的样本数目仍然没有达到 25 个，继续将格子大小扩大到 3 个，依次类推，直到格子的大小达到 5 个。此时，即使样本数目少于 25 ，也不扩大格子的大小，直

接计算样本红移的中值 [39] 。实际上这个过程就是自适应平滑。这样就产生了 1 个颜色和红移的矩阵。双色图中不同的灰度值代表不同的红移值。测试样本只要在产生的矩阵中找到对应点，就可以根据该点的灰度值得到红移值。这种方法原理很简单，天文学家比较容易理解和接受。但是这种方法的精确度不高，估测红移的剩余残差 $\sigma_{\rm rms}=0.032$ ，而且还有不同程度的损失率，当星等 $r=21$ mag 时，损失率为 5% ，而当星等 $r=23$ mag 时，损失率增至 10% 。

(3) 多项式回归法

多项式回归 (Polynomial Regression) 法是研究 1 个因变量与 1 个或多个自变量的多项式回归分析方法。如果自变量只有 1 个时，称为一元多项式回归。如果自变量有多个时，称为多元多项式回归。其最大优点就是可以通过增加 x 的高次项对实测点进行逼近，直至满意为止。多项式回归可以处理很多非线性问题，因为任何函数都可以分段用多项式来逼近，因此它在回归分析中占有重要的地位。 Connolly [40] 使用了多项式回归的方法估测红移，拟合出了光谱红移与星等或者颜色之间的非线性映射。一元回归的剩余残差 $\sigma_{\rm rms}=0.057$ ，二元回归的剩余残差 $\sigma_{\rm rms}=0.047$ ，三元回归的剩余残差 $\sigma_{\rm rms}=0.042$ 。显然，随着回归次数的增加，剩余偏差明显减小，但是由于三元线形回归的波动性比较大，因此通常研究中使用二元线形回归的较多。从上述的结果中可以看出，这种方法的估测精度不高，因为它只是光谱红移与星等之间的近似关系。为了比较准确地表示红移与颜色之间的关系，可以采用分段拟合的方式，也就是利用红移将样本分成几个区间，不同红移区间内使用不同的多项式关系进行拟合。

(4) 基于 Kd 树的多项式回归法

Kd 树是应用了 K 近邻原理将查询区域分区的快速查找方法 [41] 。基于 Kd 树的多项式回归的基本原理为：先用 Kd 树的方法将颜色空间分成若干小空间，每个小空间中样本的数目是一样的，然后在每个小空间内进行二次多项式回归。该方法估测红移的剩余残差 $\sigma_{\rm rms}=0.023$ [35] 。

(5) 贝叶斯方法

用贝叶斯方法测红移实际是集成了模板匹配方法和贝叶斯方法。由于使用了先验概率和边缘化技术，可以包含一些其他测红移方法容易忽略的相关信息，例如红移分布的预期形状、星系类型。用贝叶斯方法估测红移很明显地缩小了测光红移的弥散。在 $z<6$ 时，没有异常值和系统偏差，估测红移的剩余残差 $\sigma_{\rm rms}=0.047\,6$ 。如果先验信息缺乏，可以使用获得测光红移的数据的先验分布来替代。在高红移下，得到如此小的误差是任何训练集方法都无法实现的。在数据缺乏时，这种方法可能带来有价值的结果。而且如果先验概率与星系的测光属性没关系，基于贝叶斯理论的测红移方法可以有效地提高测光红移的精度。尽管贝叶斯的方法优点很显著，但是这种方法有可能带来一些虚假现象 [42] 。

(6) 支持向量机法

支持向量机 (Support Vector Machines，简称 SVMs) 是由 Vapnik [43] 提出的针对分类和回归问题的统计学理论。 SVM 方法具有许多引人注目的特点和有前途的试验性能，越来越受到重视。 SVM 方法是基于结构风险最小化原理的方法，明显优于传统的基于经验风险最小化的神经网络方法。常用的核函数是高斯核函数，该核函数只有一个可调参数，便于操作，因而通常将其作为默认的核函数。 SVMs 不存在 ANNs 的诸多缺点，例如：过度拟合、局部极小等。 Wadadekar(2005) [34] 利用 SVM 方法估测测光红移，估测红移的剩余残差 $\sigma_{\rm rms}=0.027$ 。我们也用 SVMs 方法估测测光红移，尝试了各种参数组合，最优的

估测红移的剩余残差 $\sigma_{\rm rms}=0.027$ [44] 。

(7) 人工神经网络法

人工神经网络 (ANNs)，又称神经网络，分为监督式的学习方法和非监督式学习方法。在监督式的学习方法中，网络是在学习过程中建立的，而且是根据权重不断调整的，结果在最后阶段产生。在非监督式学习方法中，训练过程中不提供自行调整网络，其一般适用于各种数据压缩，例如降维或聚类。人工神经网络具有前向式和后向式两种网络拓扑结构。在前向式网络中，不允许有回路，因此很快地产生结果。在后向式的网络中，允许回路产生，因此较容易产生理想的结果，但是需较长时间方可完成。其中在测光红移应用中最广泛的是多层神经网络 (Multi-layer Perceptron，简称 MLP)。 MLP 是由层和节点组成。第一层是输入层 (星等或颜色，以及光谱红移)，最后一层输出估测的测光红移值，介于输入和输出层之间的层称为隐藏层。隐藏层的个数以及节点数是变化的，同一层的节点必须与相邻层的节点相连。神经网络的结构可以写成 $N_{\rm in}$:N_1:N_2: $\cdots$: $N_{\rm out}$，其中 $N_{\rm in}$ 是输入层的节点数， N_1 是第一个隐藏层的节点数。例如：结构为 9:6:1 表示输入层有 9 个输入参数，隐藏层只有一层包含 6 个节点，输出层输出 1 个参数。每个节点都有一个权重。在将 ANNs 应用到测光红移的估计之前，需要评估样本和测试样本来选取网络结构和训练样本的参数，直到达到了最优网络，再用此网络来估测测试样本的红移。用 ANN 方法得到的测光红移精确度远远高于模板匹配方法得到的精度。目前已有多篇工作是基于 ANN 方法来估测测光红移 [32,33,44−48] 。

(8) 最近邻或 K 近邻方法

最近邻或 K 近邻方法均属于事例学习，对 K 近邻方法关键是 K 值的确定，通常采用交错鉴定方法 (Cross-Validation，简称 CV) 来选取， CV 误差最小值时对应的 K 值为最优 K 值。最近邻方法在测红移时，每一个测试样本需要在颜色空间内找到训练样本中离其最近的训练样本的红移作为该测试样本的测光红移； K 近邻则是取离测试样本最近的邻域内 K 个训练样本红移的平均值作为测试样本的测光红移。近邻法的好处是：它不用引入模型的形式，因此特别适合没有先验知识的情况 (没有先验知识就无法假定模型的形式)，但是实际上，要获得较理想的结果，需要比参数化大得多的样本集，相当于用大数据来弥补先验知识的不足。近邻法一个严重弱点是需要存储全部训练样本于内存中，这需要耗费大的内存，以及繁重的距离计算。在理想情况下，如果训练量样本足够大，而且包含了所有的星系类型，估测精度会很高。目前，这样有代表性的训练集是很难找到的。从技术角度上来分析，大的训练集预示着训练时间的增加。这就需要使用有效的多维查找技术，例如：用 Kd 树代替线性查询方式。目前，对其改进的方法大致分为两种：一种是对样本集进行组织与整理，分群分层，尽可能将计算压缩到在接近测试样本邻域的小范围内，避免盲目地与训练样本集中每个样本进行距离计算；另一种则是在原有样本集中挑选出对分类计算有效的样本，使样本总数合理地减少，这样既可以达到减少计算量，又可减少存储量的双重效果。 Csabai (2003) [36] 用最近邻方法估测测光红移的剩余标准偏差 $\sigma_{\rm rms}=0.033$ 。

(9) 核回归方法

核回归也属于事例学习家族中的一员，它不需要训练过程，将训练样本存放在内存中，当得到测试样本时，每个测试样本都需要遍历训练样本，找到在一定窗宽范围内的训练样本，

对其进行加权平均。核回归的公式如下：

$$\hat{m}_n(x,h_n)=\frac{\sum\limits_{i=1}^{n}K_{h_n}(X_i-x)Y_i}{\sum\limits_{i=1}^{n}K_{h_n}(X_i-x)}\ ,$$

其中， h_n 为窗宽， K_{h_n} 是核函数。用核回归方法估测测光红移的关键在于窗宽的确定。确定核回归窗宽的方法有多种，如交错鉴定方法、赤池信息准则 (Akaike information criterion，简称 AIC) 、施瓦茨准则 (Schwarz Information criterion，简称 SIC) 等。其中最简便的方法是交错鉴定方法，当 CV 误差达到最小值时对应的窗宽为最优窗宽。我们首次尝试用核回归的方法来估测星系的测光红移，发现其精度高于一般的训练集方法，远远优于模板匹配的方法。最优估测红移的剩余残差 $\sigma_{\rm rms}=0.019\ 2$ [44] 。类似于 Collister 和 Lahav [46] 利用 ANNs，Wadadekar [34] 利用 SVMs 估测星系的光谱类型 eClass(ANNs： $\sigma_{\rm rms}=0.052$; SVMs： $\sigma_{\rm rms}=0.057$)，王丹等人 (2007) [44] 用核回归方法对 eClass 进行了估测，并取得了令人满意的结果 ($\sigma_{\rm rms}=0.033\ 7$) 。

王丹等人 (2007) [44] 用核回归估测测光红移时，使用了光谱类型参数 eClass 作为输入参数。 eClass 是 SDSS 星表中的一个用来标志光谱类型的参数，是通过对光谱数据进行主分量分析得到的一种类型参数，是一个 0.5 ∼1 范围内的实数，值越小表示为早型星系，值大时对应晚型星系。增加该参数可以有效地改进测光红移的估测精度 ($\sigma_{\rm rms}=0.018\ 9$) 。考虑到光谱类型有助于提高测光红移的精度，将星系先分类而后估测红移。按照汇聚指数 (c) 将样本分为早型星系与晚型星系独立估测红移。对早型星系来说，精度提高得很可观。可见对将星系先分类而后估测红移是不错的方法。

3 各种方法优缺点比较

基于上面综述的各种估测测光红移方法的原理及其优缺点，同时考虑到估测的红移精度不仅依赖于估测红移的方法而且还依赖于所使用的样本及所选择的参数，因此我们只能对各个方法估测红移的效果进行粗略的比较。为更清楚起见，表 1 列出了目前的九种估测测光红移方法的优缺点，以及用于测量红移时的剩余标准偏差 $\sigma_{\rm rms}$ 值。从表 1 中我们可以看出，核回归和 ANNs 的估测精度最高，均优于 SVMs 、基于 Kd 树的多项式回归、 CMR 和多项式回归，并且远远好于模板匹配方法。

模板匹配方法是基于对光谱整体轮廓的拟合，即主要依赖于对 Lyα 森林、Balmer 跳变这类显著光谱特征的探测。每个星系的测光数据被构造成谱能量分布 (SED)，通过与从同一测光系统得到的光谱模板进行匹配，计算模板和实际光谱之间的 χ^2 值，使 χ^2 最小化来确定红移。用模板匹配的方法不仅可以得到红移值，还可以同时得到所要研究星系的类型和光度。由于模板匹配技术充分利用了星系的 SED，因此它在估测那些很少或没有光谱红移的星系样本的红移时起到很重要的作用。并且其最突出的优点是原理简单，不需要光谱样本。模板匹配方法的估测精度强烈地依赖于准确的具有代表性的 SED 模板。通常的模板来自于星族模型和真实的星系模板。对前者而言，通常包含了不真实的参数或者未包含全部已知样本信息；对后者

而言，模板一般来自亮的低红移星系样本，而忽略了高红移星系样本。对模板匹配而言，最大的缺点就是模板构造的不完备性，导致其估测的红移精度要劣于其他方法得到的结果。

表 1　9 种估测测光红移算法的优缺点及其在天文学中的应用

方 法	优　点	缺　点	剩余标准偏差 $\sigma_{\rm rms}$
模板匹配 (SED)	原理简单，不需要光谱样本，可以同时得到星系的类型和光度。对估测没有光谱红移的星系样本红移时起到很重要的作用。	估测精度强烈依赖于准确的具有代表性的 SED 模板或实测模板。通常构造完善的模板是比较困难的。	CWW 模板：$\sigma_{\rm rms}=0.066\ 6$ [36] Bruzual-Charlot 模板：$\sigma_{\rm rms}=0.055\ 2$ [36]
颜色 − 星等 − 红移关系 (CMR)	原理很简单，天文学家容易理解和接受，而且运算速度较快。	估测红移精确度不高，且有不同程度的损失率。	$\sigma_{\rm rms}=0.032$ [38]
多项式回归	理论简单，而且在训练的过程中不需要太长时间。非线性关系均可以采用多项式回归。	拟合的函数关系会随着不同观测系统和样本集的变化而变化，在高红移区，光谱红移样本很不完备，估测也就很不可靠。	一元回归：$\sigma_{\rm rms}=0.057$[40] 二元回归：$\sigma_{\rm rms}=0.047$[40] 三元回归：$\sigma_{\rm rms}=0.042$[40]
基于 Kd 树的多项式回归	估测精度较高。	查询算法速度较慢，对大数据量训练样本需要花费很长时间。	$\sigma_{\rm rms}=0.023$ [35]
贝叶斯方法	模板匹配方法和贝叶斯方法的结合。可以作为处理那些没有红移数据的补充方法。	先验概率的引入可能带入虚假现象，估测精度不高。	$\sigma_{\rm rms}=0.047\ 6$ [41]
支持向量机 (SVMs)	不需要调节网络结构，只需选择合适的核函数和临界的参数值。参数调节得当，高斯核可以得到较为理想的结果。	有的核函数的可调参数不止一个，而且参数关系是耦合在一起，调节起来比较困难，需要借助先验经验。	$\sigma_{\rm rms}=0.027$ [34,44]
人工神经网络 (ANNs)	估测红移精度相当高，而且参数越多，估测精度会相应提高。	训练网络的选取较为复杂，内部结构十分复杂，可解释性差，易造成过度拟合和陷入局部极小，训练时间较长。	Collister: $\sigma_{\rm rms}=0.023$ [46] Firth: $\sigma_{\rm rms}=0.021$ [32] Vanzella: $\sigma_{\rm rms}=0.022$ [47]
最近邻或 K 近邻	不需要创建模型，直接依赖于数据，适合没有先验知识的情况。	存储全部训练样本于内存中，耗费大的内存，以及繁重的距离计算。	$\sigma_{\rm rms}=0.033$ [36]
核回归	原理简单，估测精度较高。用于估测光谱型 eClass 时也取得了令人满意的结果。	必须选择最优窗宽。存储全部训练样本于内存中，耗费大的内存和计算时间。窗宽较小时易造成损失点的增加。	$\sigma_{\rm rms}=0.019\ 2$ [44]

颜色 – 星等 – 红移关系方法是通过将样本按 r 星等值分区，对每个星等区间建立自己的双色图。通过双色图上的灰度值来估测红移。这种方法很容易实现，在原理上更接近于天文学家的思路。但是由于此方法是按星等区间划分，对那些在星等分界线附近的星系红移估测不够准确。而且在估测红移时，超出某个星等时，红移估测损失率会加大，这种损失是不可弥补的。

多项式回归方法在估测红移时是将红移拟合成星等或者颜色的函数关系，利用这种函数关系来估测那些未知红移的星系样本，并且不需要知道星系光谱演化信息。其先天优势是原理上比较简单，天文学家容易理解。而且在训练过程中不需要太长时间，估测红移速度很快，对于大样本而言略有优势。但是其缺点也很明显，即估测精度不高。多项式拟合的函数关系会随着不同观测系统和样本集的变化而变化，在高红移区的光谱红移样本很不完备，这样对于高红移星系样本的红移估测也就很不可靠。

Benitz 提出了一种 HyperZ 和 Bayesian marginalization 合并的方法。该方法有助于揭示天体的相关信息，例如：红移分布的预期形状和星系类型。这点正是其他方法所忽略的。这种方法大大提高了 HyperZ 方法估测红移的精度。但是在应用过程中某些具体的课题可能渗入虚假因素的影响。而且估测红移精度不高。因此，这种合并方法只能作为处理那些没有红移数据的补充方法。与 ANNs 相比较而言， SVMs 简化了训练过程， SVMs 不需要训练网络，它只需要选择合适的核函数和参数。如果参数调解得当，最简单的高斯核函数也可以取得较好的估测结果。但是 SVMs 的参数调整需要先验知识，并且参数间的耦合关系使参数调整的过程更加复杂，训练时间也较长。

ANNs 犹如 “黑箱”，只能看到它的输入和输出，实际上它的内部结构十分复杂，可解释性差，并且它没有固定的网络结构。没有经验的用户要在实践中花费很多的时间和精力来摸索如何调整网络和参数。当 ANN 网络越复杂，输入参数越多时， ANNs 试图对数据进行较精确的拟合，则很容易造成过度拟合。而且在参数空间进行局域搜索时，常常陷入局域极小值。另外，当添加层数和节点数时需要重新训练网络，相应的训练时间也会增加。从估测红移精度而言， ANN 方法无疑是很好的选择，但其训练速度相对较慢。

核回归具备事例学习的一切优缺点。其特点是当训练样本数目很大时，需要花费较长的测试时间，并且占用很大的内存；另一个特点是当窗宽较小时，有些测试样本在对应的窗宽内找不到训练样本，这样测试样本就无法得到红移值，这些点称其为损失点。当窗宽增大时，损失点的比率会骤减。所以对于核回归方法的改进是需要提供高效率的查找方法和采用自适应窗宽，兼顾它的独特优势 —— 估测精度相当高，采用改进的核回归方法将有效地提高估测测光红移的效率。

4 总结与展望

从测光红移的概念、研究现状、科学意义及其测量方法出发，重点论述了测光红移算法的分类，各种测光红移算法的原理及其应用。这些方法基本涵盖了该领域的方方面面：基于模板匹配的 HyperZ 法、基于物理参量与红移关系的颜色 – 星等 – 红移关系法、基于统计学原理的多项式回归法、基于 Kd 树的多项式回归和贝叶斯方法、基于核理论的支持矢量机、基于机

器学习的人工神经网络法、基于事例学习的核回归和最近邻或 K 近邻方法。每种方法各有其优缺点。从天文学家易于理解和操作及速度的角度考虑，模板匹配、颜色 — 星等 — 红移关系法和多项式回归是不错的选择；从精度而言，神经网络、核回归较好；既考虑精度又考虑可控制性，支持矢量机方法要好些；考虑样本的不完备和非均匀性，事例学习尽可能避免样本的这些缺陷。模板匹配的关键之处在于模板的创建，模板是否优越直接影响估测结果；神经网络需要在如何选取网络层数、各层的节点数以及何时停止训练上作较大的努力，而且其极易限于局部最小；支持矢量机在于选择合适的核函数及其相应参数的调整；核回归重要的工作在于最优窗宽的确定； K 近邻则是在于 K 值的选择。因此在实际的应用中，需综合考虑上述的各种因素，选取适当的方法来估测红移。正是由于各种方法仍存在不足之处，才使得在这方面的探索生生不息，也正是这些不足成为推动此领域发展的强大动力。

在未来的工作中，考虑采用不同的数据样本和改进算法。对于数据样本，将应用其他波段的多色测光数据，如来自于紫外波段的 GALEX 数据、来自于红外波段的 Spitzer 数据。核回归有很大的灵活性，它不仅可以替换核函数、距离公式，或者使用自适应窗宽的核回归，其中自适应窗宽的核回归可以避免损失点的问题，而且在估测精度上也要高于固定窗宽。为了加快核回归的效率，可以使用 Deng 和 Moore [49] 提出的多分辨率事例学习方法，这样可以缩短事例学习所用的时间。这种方法有两个突出的优点：灵活地操作局部和全局数据；当训练样本改变时不需要重新训练。对 SVMs 可以尝试用不同的核函数，或者利用支持矢量机的改进算法最小二乘矢量机估测红移。另外，在数据预处理阶段，对高维数据进行特征提取、特征选择来有效地降维。改进的核回归方法或多种方法的混合方法将是估测测光红移方法的完美选择。

这些方法不仅适合于测光数据，也可以用于光谱数据。例如，河外天体光谱红移的测量、恒星物理参数 (有效温度、金属丰度、表面重力加速度) 的测定。尤其对将来大型光谱巡天 LAMOST 望远镜产生的海量光谱数据的自动化处理具有重要的参考和应用价值。

参考文献：

[1] Puschell J J, Owen F N, Laing R A. ApJ, 1982, 257: 57
[2] Loh E D, Spillar E J. ApJ, 1986, 303: 54
[3] D'Abrusco R, Staiano A, Longo G, et al. ApJ, 2007, 663: 752
[4] Baum W A. AJ, 1957, 62: 6
[5] Baum W A. Proceedings from IAU Symposium no. 15, Macmillan Press, 1962: 390
[6] Koo D C. AJ, 1985, 90: 418
[7] Bruzual A G. AJ, 1983, 273: 105
[8] Cowie L L, Lilly S J, Gardner J, et al. ApJ, 1988, 332: 29
[9] Guhathakurta P, Tyson J A, Majewski S R. ApJ, 1990, 357: 9
[10] Steidel C C, Hamilton D. ApJ, 1993, 105: 2017
[11] Puschell J J, Owen F N, Laing R A. ApJ, 1982, 257: 57
[12] Guiderdoni B. Proceedings of the Third IAP Work-shop. France: IAP, 1987: 271
[13] Koo D C. ApJ, 1986, 311: 651
[14] Lilly S J, Cowie L L, Gardner J P. ApJ, 1991, 369: 79
[15] Subbarao M U, Connolly A J, Szalay A S, et al. ApJ, 1996, 112: 929

[16] Connolly A J, Szalay A S, Koo D, et al. ApJ, 1996, 473: 67
[17] Koo D C. ApJ, 1981, 251: 75
[18] Fontana A, Dodorico S, Poli F, et al. AJ, 2000, 120: 2206
[19] Poli F, Giallongo E, Menci N, et al. AJ, 1999, 527: 662
[20] Giallongo E, Menci N, Poli F, et al. ApJ, 2000, 530: 73
[21] Blake C, Collister A, Bridle S, et al. MNRAS, 2007, 374: 1527
[22] Padmanabhan N, White M, Eisenstein D J. MNRAS, 2007, 376: 1702
[23] Edmondson E, Miller L, Wolf C. MNRAS, 2006, 371: 1693
[24] York D G, Adelman J, Anderson J E, et al. AJ, 2000, 120: 1579
[25] Le Fèvre, Vettolani G, Maccagni D, et al. The Messenger, 2003, 111: 18
[26] Davis M, Sandra M F, Jeffrey N, et al. Proceeding of SPIE. 2003, 4834: 161
[27] Bruzual A G, Charlot S. ApJ, 1993, 405: 538
[28] Coleman G D, Wu C C, Weedman D W. ApJS, 1980, 43: 393
[29] Brunner R J, Connolly A J, Szalazy A S. ApJ, 1997, 482: 21
[30] Wang Y, Bahcall N, Turner E L. AJ, 1998, 116: 2081
[31] Budavári T, Szalay A S, Charlot S, et al. ApJ, 2005, 619: 31
[32] Firth A E, Lahav O, Somerville R S. MNRAS, 2003, 339: 1195
[33] Tagliaferri R. Lecture Notes in Computer Science. 2003, 2859: 226
[34] Wadadekar Y. PASP, 2005, 117: 79
[35] Bolzonella M, Miralles J M, Pell ó R. A&A, 2000, 363: 476
[36] Csabai I, Budavári T, Connolly A J, et al. AJ, 2003, 125: 580
[37] Mobasher B, Capak P, Scoville N, et al. ApJS, 2007, 172:117
[38 Wang D, Zhang Y, Cui C, et al. SPIE, 2006, 6274: 13
[39] Yang Y, Huo Z, Zhou X, et al. ApJ, 2004, 614: 692
[40] Connolly A J, Csabai I, Szalay A S, et al. AJ, 1995, 110: 2655
[41] Bentley J L. Communications of the ACM, 1979, 19:509
[42] Benítez N. ApJ, 2000, 536: 571
[43] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer Verlag 1995
[44] Wang D, Zhang Y Z, Liu C, et al. ChJAA, 2008, 8: 119
[45] Ball N M, Loveday J, Fukugita M, et al. MNRAS, 2004, 348: 1038
[46] Collister A A, Lahav O. PASP, 2004, 116: 345
[47] Vanzella E, Cristiani S, Fontana A, et al. A&A, 2004, 423: 761
[48] Li L, Zhang Y, Zhao Y, et al. ChJAA, 2007, 7: 448
[49] Deng K, Moore A. Proceedings of the Twelfth International Joint Conference on Artificial Intelligence San Francisco: Morgan Kaufmann, 1995: 1233

Overview of Photometric Redshift Estimators

WANG Dan, ZHANG Yan-xia, ZHAO Yong-heng

(*National Astronomical Observatories, Chinese Academy of Sciences, Beijing 10012*)

Abstract: With the establishment and development of large digital sky survey projects, astronomic data are measured by TB, even PB, including various photometric and spectroscopic data. Photometric redshifts have shown their superiority compared to spectroscopic ones. So far pho-

tometric redshifts have been regarded as an efficient and effective measure for studying the statistical properties of the large-scale structure of the universe and the formation and evolution of galaxies. We illustrate the conception, background and approaches of photometric redshifts, as well as its application in astronomy, then mainly summarize nine approaches to determine photometric redshifts, namely HyperZ, Color-Magnitude-Redshift relation (CMR), polynomial regression, polynomial regression based on KdTree, Bayesian method, Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), K-nearest neighbor and kernel regression. Photometric redshift techniques have been divided into two broad categories: template matching method and empirical training-set method. The former includes HyperZ, and the latter contains CMR, polynomial regression, polynomial regression based on KdTree, Bayesian method, SVMs, ANNs. Another interpolative training-set methods are instance-based learning techniques, which are composed of nearest neighbor, K-nearest neighbor and kernel regression. There are advantages and disadvantages to each approach. Template matching technique relies on fitting model galaxy spectral energy distributions (SEDs) to the photometric data, where the models span a range of expected galaxy redshifts and spectral types. The Achilles heel of the technique is the shortage of large and complete template sets. The training set method depends on representative and complete training sets, moreover it is difficult to extrapolate to regions that are not well sampled by the training set. Unlike the traditional training methods, the best merit of instance-based learning approach is the ability to make predictions with different parameters without needing a retraining phase, moreover it doesn't seriously depend on the size of sample. Nevertheless, instance-based learning approach has the obvious disadvantage that is a significant computational cost on large data sets. In summary, only regarding the accuracy of estimating photometric redshifts, ANNs and kernel regression are best choices. In terms of understanding and easy implementation, HyperZ, CMR and polynomial regression are better. In the end, the prospect of algorithms for measuring photometric redshifts is described.

Key words: astrophysics; photometric redshifts; distances; statistical methods