分类号_____ 密级_____

UDC _____ 编号_____

中国科学院研究生院

博士学位论文

基于大型巡天数据的测光红移算法研究

—算法研究与工具开发

王 丹

指导教师 赵永恒 研究员 中国科学院国家天文台

张彦霞 副研究员 中国科学院国家天文台

申请学位级别 博士 学科专业名称 天文技术与方法

论文提交日期 2007 年 5 月 论文答辩日期 2007 年 7 月

培养单位 中国科学院国家天文台

学位授予单位 中国科学院研究生院

答辩委员会主席 何香涛 教授

Typeset by LaTeX 2ε at July 6, 2007

With package `CASthesis` v0.1h of CTeX.ORG

Research on Algorithms of Estimating Photometric Redshifts Based on Large Sky Survey Databases

—Algorithm Research and Tool Development

Wang Dan

Supervisor:

Prof. Zhao Yong-heng & Asoc. Prof. Zhang Yan-xia

National Astronomical Observatories, Chinese Academy of Sciences

July, 2007

Submitted in total fulfilment of the requirements for the degree of Ph.D. in Astronomical Technology and Method

摘　要

随着各种天文数据处理工具的开发和应用，虚拟天文台的功能将越来越完善。在此背景下，开发了红移测量工具和图像处理与分析工具。探讨了预测测光红移算法的原理和应用，侧重研究了四种测光红移算法，并与其他的测光红移算法进行了比较。本文的主要成果如下：

（1）综述了不同的预测测光红移的算法，包括模板匹配方法（例如：HyperZ），经验训练集方法（例如：颜色-星等-红移关系、多变量多项式回归、基于Kd树的多项式回归、贝叶斯、支持矢量机、人工神经网络）以及事例学习方法（例如：核回归、最近邻、K近邻）。并讨论了各种预测红移算法的效率及其优缺点。

（2）我们使用SDSS和2MASS巡天数据，重点研究了四种预测测光红移的算法：颜色-星等-红移关系方法、多变量多项式回归、支持矢量机和核回归。分别讨论了不同输入参数组合对支持矢量机和核回归预测红移精度的影响。从实验结果中获知，增加参数，预测精度并不一定提高，只有使用了合适的参数（例如：eClass），预测精度才会有明显的提高。而且所选择的方法不同，最优输入参数也不一定相同。对于使用了SDSS和2MASS交叉证认的星表后，预测红移精度并没有提高，这或许是因为样本集的数目减少的缘故。对于核回归来说，最关键的一步就是窗宽的选择。最优窗宽可以由三种方法确定：交叉鉴定方法、AIC和BIC。核回归在星系样本分为早型和晚型时，预测精度提高，尤其对早型星系更为明显。在这些方法中，就预测精度而言，核回归更显示了它的优越性。

（3）基于颜色-星等-红移关系方法，我们开发了测光红移测量工具，初步实现了数据的上传、结果下载等功能。而且开发了虚拟天文台图像处理工具VO_IMPAT（Image Processing and Analysis Toolkit for the Virtual Observatory of China），它可以同时访问数字巡天图像（DSS）、天文星表和其他的数据库。VO_IMPAT 实现了多波段天文数据的融合，即将不同波段的星表（例如：USNO、2MASS、NVSS、RASS）叠加在DSS图像上。VO_IMPAT还提供许多重要的图像处理功能，如放大缩小、改变颜色、等高线、画图工具、直方图、平滑化、尖锐化、旋转、标记等。

我们将进一步探索新的测光红移算法，并且完善红移测量工具和图像处理

工具，希望它们能够真正成为天文学家从事科研工作的有力助手。

关键词： 虚拟天文台-数据挖掘-星系: 距离和红移-方法: 数据分析-技术:测光

Abstract

Research on Algorithms of
Estimating Photometric Redshifts Using SDSS Data

Wang Dan (Astronomical Technology & Method)

Directed by Prof. Zhao Yong-heng & Asoc. Prof. Zhang Yan-xia

With the development and application of various toolkits for data processing in astronomy, the functions of the Virtual Observatory will be more perfect. Under this situation, we have developed tools of photometric redshift determination as well as imaging processing and analysis. The principals and implementations of algorithms for estimating photometric redshifts have been stressed. Four kinds of techniques to predict photometric redshifts have been studied and compared with other approaches. The main contributions of this dissertation are as follows:

(1)We have summarized various approaches to estimate photometric redshifts, including the template fitting technique (e.g. HyperZ), empirical training-set methods (e.g. color-magnitude-redshift relation, multiple polynomial regression, polynomial regression based on Kd-tree, bayesian method, support vector machines, artificial neural networks) and instance-learning approach (e.g. kernel regression, nearest neighbor or K-nearest neighbor). Performances of these techniques, as well as their advantages and disadvantages, have been illustrated.

(2)Based on SDSS and 2MASS databases, we have mainly investigated four methods: color-magnitude-redshift relation, multiple polynomial regression, support vector machines and kernel regression for photometric redshift estimation. We have explored the performances of support vector machines and kernel regression with different input patterns. Our experiments have shown that the more parameters considered, the accuracy doesn't always increase, and only when appropriate parameters chosen (e.g. eClass), the accuracy can improve. Moreover, for different approaches, the best input pattern is different. There is no improvement in the use of more parameters from SDSS and 2MASS catalogs probably

due to decreasing the sample size. For kernel regression, one important design decision is the choice of bandwidth. The optimal bandwidth is adopted by one of the three criterions: cross-validation, AIC and BIC. When the sample has been divided into early-type galaxies and late-type ones, the precision has increased, especially for early-type ones. Of these four methods, kernel regression has shown its superiorities in terms of accuracy.

(3)On the basis of Color-Magnitude-Redshift relation approach, we have developed a tool to estimate the photometric redshifts of celestial objects, and realized the fundamental functions, such as uploading data and downloading results. Moreover, we have implemented another toolkit, VO_IMPAT (Image Processing and Analysis Toolkit for the Virtual Observatory of China), which provides simultaneous access to images of the Digital Sky Survey (DSS), astronomical catalogs, and other databases. VO_IMPAT has realized the federation of multi-band astronomical data, i.e. overlaying catalogues (USNO, 2MASS, NVSS, RASS) from different bands upon the images of DSS. In addition, VO_IMPAT provides many essential image processing functions, which include zooming in and out, changing color-maps, contours, drawing, histogram, smoothing, sharpening, rotation, tagging and so on.

We will further explore new algorithms to estimate photometric redshifts, moreover, ameliorate the tool of photometric redshift measurement as well as VO_IMPAT, and hope that they would be true helpers for astronomical research.

Keywords: virtual observatory - data mining - galaxies: distances and redshifts - methods: data analysis - techniques: photometric

目　　录

表　　格

插　　图

第一章　河外天体的测光红移背景

当光源远离观测者时，接受到的光波频率比其固有频率低，即向红端偏移，这种现象称为“红移”。当光源接近观测者时，接受频率增高，相当于向蓝端偏移，称为“蓝移”。美国天文学家哈勃于1929年确认，遥远的星系均远离我们地球所在的银河系而去，同时，它们的红移随距离增大而成正比地增加，这一普遍规律称为哈勃定律，它成为星系退行速度及其与地球的距离之间相关的基础。哈勃发现，来自星系的光谱呈现某种系统性的红移，即星系正在远离我们；而且发现离我们越远的星系退行速度越高。哈勃定律的伟大意义，不仅在于它证实了宇宙的膨胀，而且还提供了一种估计宇宙年龄的手段。

由于宇宙的膨胀，天体相对于观察者以一定的速度退行，即天体具有红移。将天体中特定原子的光谱与地球上实验室内同种原子的光谱进行比较，可以确定光源正在以多大的速度退行。天体光谱中某一谱线相对于实验室光源的比较光谱中同一谱线向红端的位移，即红移。红移（z）的定义是：

$$z = \frac{\lambda - \lambda_0}{\lambda_0} \tag{1.1}$$

式中λ_0是实验室光源的某一谱线波长，λ 是天体的同一谱线波长。z >0，红移，波长增加；z <0，蓝移，波长减少。在红移问题中，z都大于0，因而往往简单地把z作为红移的符号。z是无量纲的标量，习惯上又总是按照多普勒效应把z换算为相应的速度。

通过公式（1.1）就可以得到天体的光谱红移。根据哈勃定律，对于小红移的天体来说，红移z与距离d存在下列关系

$$d = \frac{cz}{H_0} \tag{1.2}$$

其中c是光速，H_0是哈勃常量(75 km.s^{-1}.Mpc^{-1}), z是红移。对于大红移的天体来说，距离与红移的关系为：

$$d_{\mathrm{L}} = \frac{c}{H_0 q_0^2}\{q_0 z + (q_0 - 1)[(1 + 2q_0 z)^{\frac{1}{2}} - 1]\} \tag{1.3}$$

其中q_0是减速因子，d_{L}光度距离。绝大多数天体的距离是不能直接通过观测得到的，因此只要测出红移z，可以使用公式（1.2）或（1.3）得到天体的距离[1]。

哈勃定律揭示宇宙是在不断膨胀的，这种膨胀是一种全空间的均匀膨胀。因此，在任何一点的观测者都会看到完全一样的膨胀，从任何一个星系来看，一切星系都以它为中心向四面散开，越远的星系彼此散开的速度越大。

天体距离的测定，对于研究天体的空间位置和形成与演化，求得天体的光度函数等，均具有重要的意义。尤其对于那些无法获得光谱数据的暗源而言，利用其已有的测光数据来获得测光红移，具有更重要的研究价值。随着大型的巡天项目的发展，探测技术的提高，已经积累了海量的天文数据（如：图像数据、光谱数据和测光数据等）。这些数据为进行测光红移算法的研究提供了丰富的实验床。下面我们将要综述测光红移及各种测光红移算法的研究背景，以此为基础探讨一些测光红移算法，并比较各种算法的优劣，从而为天文学家选取合适的算法来预测测光红移提供重要的参考。

1.1 测光红移的研究现状

测光红移是指使用中波段和宽波段的测光数据或者图像得到红移，更通俗地说，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。测光红移技术已经被广泛地应用到深空大天区的巡天项目中，例如哈勃深场（Hubble Deep Field，简称HDF）、SDSS巡天。

“测光红移”并不是一个新名词，它最早出现在Puschell等人（1982）[2]的文章中。Puschell利用宽带测光数据预测暗射电星系红移。在三个方面起到了先驱作用：(1) 使用了近红外（JHK）和光学波段（RI）；(2) 采用了χ^2最小的能量光谱分布；(3) 在预测红移过程中，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）[3]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~21.5mag的星系，同时利用6个中波段的滤光片和χ^2最小的模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三个本星系的SED模板来代表所有红移值下的星系类型。

追本溯源，最早将多波段测光方法应用到红移预测工作中的是Baum。1957年Baum

[4] 提出利用测光数据研究红移，并于1962 年[5]研究了一种估计测光红移的算法，即使用光电光度计和9个滤光片，这9个滤光片覆盖了从3730Å到9875Å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）的椭圆星系的光谱能量分布，随后又得到了C1095+2044星系团（又称为Abell 0801）中的三个椭圆星系的光谱能量分布。利用波长的对数尺度，他将平均的室女星系团的光谱能量分布与平均的C10925 星系的光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。利用这种方法预测的C1095+2044星系团的红移$z = 0.19$，这与用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的这种方法用于预测红移时比较精确，但是由于其依赖于4000Å处的光谱截断特征来预测红移，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的方法来预测红移。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，他没有使用经验的光谱能量分布，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。Koo的具体做法是在双色图上绘制随光谱变化的等红移线（iso-z lines）。在一定的红移范围内，绝大多数正常的双色图（例如：U-J对J-F图、J-F 对F-N图）是耦合在一起的，为此他研究了颜色形状图。用形状来衡量SED的两端是向上摆动还是向下弯曲，也就是说，光谱是凹形的还是脊形的。颜色由光谱波长的一阶导数来衡量，形状由二阶导数来衡量。颜色使用2U-2F或者U+J-F-N，对于形状，可以使用-U+2J-F或者-U+J+F-N。利用上述方法，Koo 从UJFN星等计算了颜色和形状关系，并绘制在颜色形状图上。星系红移是由颜色形状图上的对应点与最近的等红移线来决定的。距离最近的等红移线的红移值即为该星系的红移。Koo利用这种方法，测试了100个已知红移在z=0.025到z=0.7范围内的星系红移。

相对于光谱红移来说，测光方法预测红移的优点在于速度较快。用光谱观测的方法预测红移时，来自于星系的光线被分到一些宽度只有几埃的狭缝中。每个狭缝只能得到很少的光。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，狭缝的宽度大概是1000Å，只需要很短的曝光时

间就可以得到与光谱相同的信噪比。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）。与光谱方法测红移的方法相比较，测光方法预测红移的最大缺点是精度较低。用光谱方法得到的红移的绝对误差Δz=0.001，而测光红移Δz=0.1。这种精度对于研究单个具体星系的性质是远远不够的。但是对于研究星系的形成与演化以及宇宙大尺度结构来说，这种精度足以满足大样本统计研究的需要。

测光红移已经广泛地应用到天文学的许多科学研究上，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）利用测光红移通过观测U波段的Lyman跳变来研究红移z >3的原始星系[8][9][10]；（2）利用测光红移研究高红移类星体或远距离射电星系[11]；（3）利用测光红移研究场星系的演化或者星系的光度函数[12][13][14][15]等；（4）利用测光红移区分星团和超星团[16][17]；（5）利用测光红移预测宇宙的几何结构[3]；（6）利用测光红移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）利用测光红移研究星系尺度的演化[19][20]；（8）利用测光红移确定宇宙中重子和物质密度[21]；（9）利用测光红移研究在SDSS巡天的图像数据中亮红星系的聚类[22]。

通常，星系的红移都是通过光谱观测的形式得到的。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) [23])、VLT/VIRMOS巡天[24]、VST巡天、Keck DEEP2巡天[25]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的分析工具。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，从而可以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 大量的测光数据预测星系的红移，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。

1.2　测光红移的算法研究

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、机器学习、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来。目前，多种方法已成功地应用到预测测光红移问题上，常用的方法分为三类：模板匹配方法、训练集方法和事例学习方法。

模板匹配方法也就是能量光谱分布（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。最典型的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[26]）；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，简称CWW[27]）。这两种模板都有自身的缺点：来自于星系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，采用统计理论预测红移的方法。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。光谱红移可以用来做为约束，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预测新样本红移时，极易产生偏差。然而，训练集方法也有其优点：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巡天也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[33]、Wang[29]和Budavari[30]的多项式回归方法；Firth[31]和Tagliaferri[32]的人工神经网络（Artificial Neural Networks，

简称ANNs）；Wadadekar[34]的支持矢量机方法（Supprot Vector Machines，简称SVMs）等。

除了上述提到的两种方法外，还有一种不需要训练的预测红移方法—事例学习方法，又称懒学习方法。事例学习方法和训练集方法的相同之处在于：均需要存在一个大的、有代表性的训练集。不同的地方在于：前者不需要训练过程。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

HyperZ方法是基于对光谱整体轮廓的拟合，即主要依赖于对Ly α 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方法中的模板可以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。

$$\chi^2 = \sum_{\mathrm{i}=1}^{N_{\mathrm{filters}}} [\frac{F_{\mathrm{obs,i}} - b \times F_{\mathrm{temp,i}}(z)}{\sigma_{\mathrm{i}}}]^2 \tag{1.4}$$

其中$F_{\mathrm{obs,i}}$、$F_{\mathrm{temp,i}}$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_{filters}为观测使用的滤光片数目。SED方法的最大优点在于其原理简单，且不需要光谱红移样本。由于不可能构造出适合所有星系的模板，因此模板匹配方法预测红移的精度不是很高。

（2）颜色-星等-红移关系法

另一种预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、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，因此不可以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。首先，按照SDSS巡天中的r星等将样本分成七个子区间r_1到r_7，然后每个星等区间画两张双色图，分别为$u-g$与$g-r$图和$g-r$与$r-i$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，当一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。如果落入一个格子中的样本数目少于25个时，将格子的范围扩大成两个格子的大小，再计算此时两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三个，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[62]。实际上这个过程就是自适应平滑。这样就产生了一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，比较易于天文学家的理解和接受。但是这种方法的精确度不高，预测红移的剩余残差$\sigma_{\rm rms}$=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%[36]。

（3）多项式回归

多项式回归（Polynomial Regression）是研究一个因变量与一个或多个自变量的多项式回归分析方法。如果自变量只有一个时，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。一元m次多项式回归方程为：

$$y = b_0 + b_1 x + b_2 x^2 + ... + b_{\rm m} x^{\rm m} \tag{1.5}$$

二元二次多项式回归方程为：

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2 \tag{1.6}$$

如果因变量y与自变量x的关系为非线性的，但是又找不到适当的函数曲线来拟合时，则可以尝试采用多项式回归。多项式回归的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，因为任一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。在通常的实际问题中，不论因变量与其他自变量的关系如何，我们总可以用多项式回归来进行分析。Connolly[37]使用了多项式回归的方法预测红移，拟合出了光谱红移与星等或者颜色之间的非线性映射，这个映射大多数是二次或者三次的。这种方法的预测精度不高，因为它只是光谱红移与星等之间的近似关系。为了比较准确的表示红移与颜色之间的关系，可以采用分段拟

合的方式，也就是按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。多项式回归方法的先天优势是理论简单，而且在训练的过程中不需要太长时间。但是其缺点也很明显，预测精度不高。此外当增加训练样本的个数时，需要重新训练获得回归关系。

（4）基于Kd树的多项式回归

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多项式回归。该方法在测光红移的应用见文章Csabai，预测红移的剩余残差$\sigma_{\rm rms}$=0.023[38]。

（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。用贝叶斯方法预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，预测红移的剩余残差$\sigma_{\rm rms}$=0.0476。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[39]。

（6）支持矢量机

支持向量机（Support Vector Machines，简称SVMs）是由Vapnik[40]提出的针对分类和回归问题的统计学习理论。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。与神经网络方法不同的是，SVM方法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。如果参数调解得当，即使最简单的高斯核也可以获得很理想的结果。通常选择优化的参数是比较困难的，因为每个核函数都有一个或几个参数需要调节，它们往往是耦合在一起的。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。Wadadekar[34]（2005）利用SVM方法预测测光红移，预测红移的剩余残差$\sigma_{\rm rms}$=0.027。我们也用SVM方法预测测光红移，尝试了各种参数组合，最优的

预测红移的剩余残差$\sigma_{\rm rms}$=0.027[41]。

（7）人工神经网络

人工神经网络（Artificial Neural Networks，简称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），最后一层输出预测的测光红移值，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：... ：$N_{\rm out}$，其中N_{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，再用此网络来预测测试样本的红移。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。但是神经网络有一些缺点，它的结构需要有先验知识去构造和调整，否则很难得到最优的网络。另外，在训练的过程中，神经网络会陷入局部最小。权重由网络的层数和每层节点的个数决定。随着层数和节点的增加，训练时间也将增加。目前已有多篇工作基于ANN方法来预测测光红移[31][32][42][43][44]。

（8）核回归

核回归隶属于事例学习家族，自然具有事例学习的各种优缺点，例如：所有训练样本全部存放于内存中，直到有测试样本时才学习。核回归的公式如下：

$$\hat{m}_{\rm n}(x,h_{\rm n})=\frac{\sum\limits_{\rm i=1}^{\rm n}K_{h_{\rm n}}(X_{\rm i}-x)Y_{\rm i}}{\sum\limits_{\rm i=1}^{\rm n}K_{h_{\rm n}}(X_{\rm i}-x)} \tag{1.7}$$

其中$h_{\rm n}$为窗宽，$K_{h_{\rm n}}$是核函数。

用核回归方法预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（Cross-Validation，简称CV）、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预测星系的测光红移，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。最优预测红移的剩余残差$\sigma_{\rm rms}$=0.0192[41]。

（9）最近邻或K近邻方法

同样最近邻或K近邻方法也属于事例学习的一种，对K近邻方法关键是K值的确定，通常采用交错鉴定方法来选取，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：它是非参数方法，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，非参数方法要获得较理想的结果，需要大的样本集，比参数化大得多，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，预测精度会很高。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）用中最近邻方法预测测光红移的剩余标准偏差$\sigma_{\rm rms}$=0.033[38]。

基于上面综述的各种测光红移方法的原理及其优缺点，同时考虑到红移预测精度不仅依赖于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预测红移的效果进行粗略的比较。为更清楚起见，图1.1列出了目前的九种预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。从图1.1中我们可以看出，核回归和ANNs的预测精度最高，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。

1.3　总结

本章从红移的定义谈起，引出测光红移的概念、研究现状、科学意义及其测量方法，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，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、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、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，事例学习可以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预测结果；神经网络需要较大的努力在如何选取网络层数、各层的节点数以及何时停止训练上，而且其极易限于局部极小；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，选取适当的方法来预测红移。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。

方 法	优 点	缺 点	剩余标准偏差 σ_{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。对预测没有光谱红移的星系样本红移时起到很重要的作用	预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：$\sigma_{rms}=0.0666$[35] Bruzual-Charlot 模板：$\sigma_{rms}=0.0552$[35]
颜色-星等-红移关系（CMR）	原理很简单，比较易于天文学家的理解和接受，而且运算速度较快。	预测红移精确度不高，且有不同程度的损失率。	$\sigma_{rms}=0.032$[37]
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，预测也就很不可靠。	一元回归：$\sigma_{rms}=0.057$[38] 二元回归：$\sigma_{rms}=0.047$[38] 三元回归：$\sigma_{rms}=0.042$[38]
基于 Kd 树的多项式回归	预测精度高。训练速度快。	易产生分段误差。	$\sigma_{rms}=0.023$[35]
贝叶斯方法（Bayesian method）	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，预测精度不高。	$\sigma_{rms}=0.0476$[39]
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。	$\sigma_{rms}=0.027$[33][42]
人工神经网络（ANNs）	预测红移精度相当高，而且参数越多，预测精度会相应提高。	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: $\sigma_{rms}=0.023$[44] Firth: $\sigma_{rms}=0.021$[31] Vanzella:$\sigma_{rms}=0.022$[45]
最近邻或 K 近邻（K-nearest neighbor）	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。	$\sigma_{rms}=0.033$[35]
核回归 (Kernel regression)	原理简单，预测精度较高。用于预测光谱型 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。	$\sigma_{rms}=0.0192$[42]

图 1.1: 九种预测测光红移算法的优缺点及其在天文中的应用。

第二章　所用巡天数据介绍

2.1　SDSS巡天数据

SDSS巡天（Sloan Digital Sky Survey，简称SDSS[45]）是当今世界上最雄心勃勃的天文巡天项目。SDSS巡天望远镜是在美国新墨西哥州APO（Apache Point Observatory，简称APO[46])天文台建造的一台口径2.5米的专门的天文望远镜，使用大视场拼接CCD相机和多目标光纤光谱仪两种观测模式，对1万多平方度天区进行直接成像和选源的光谱观测。这种大视场的CCD相机一次可以拍摄1.5平方度，一夜大概可以拍摄八次。其目标是确定四分之一天区内1亿个以上天体的位置和绝对亮度。若遇上有月光的夜晚或是有薄云的夜晚，将使用一对光纤光谱仪取代成像照相机，获得天体的光谱。这种观测方法与传统的望远镜的分配方法不同。SDSS巡天的主要科学目标，是通过获得百万级数目的星系和类星体的三维空间数据，开展宇宙大尺度结构的研究。SDSS一期（SDSS-I）在2005年6月完成，其涵盖了从太阳系天体、恒星、星系、星系团到宇宙大尺度结构的众多领域，取得了影响重大的科研成果。如今，SDSS巡天已经进入了新的发展时期，SDSS二期（SDSS-II）除了继续一期的星系红移巡天外，还进行银河系恒星巡天和超新星巡天观测。SDSS巡天的测光极限星等r <22.2mag, 光谱的极限星等r <17.77mag。

APO天文台和费米国家加速器实验室（Fermilab）是SDSS的两个主要执行单位。APO天文台位于美国西南新墨西哥州，有合适的天文观测条件，是SDSS巡天的天文仪器和巡天观测运行的所在地。Fermilab位于美国芝加哥西郊，其中的实测天文研究组（EAG）承担了SDSS巡天的数据处理、数据管理、巡天观测协调以及部分科学课题研究等工作。

SDSS项目是一个庞大的国际合作计划，包括了25个单位。它们是美国自然历史博物馆（the American Museum of Natural History）、波茨坦天体物理研究所（Astrophysical Institute Potsdam）、巴塞尔大学（University of Basel）、剑桥大学（Cambridge University）、凯斯西部保留地大学（Case Western Reserve University）、芝加哥大学（University of Chicago）、Drexel大学（Drexel University）、费米国家加速器实验室（Fermilab）、高等研究中心（the Institute for

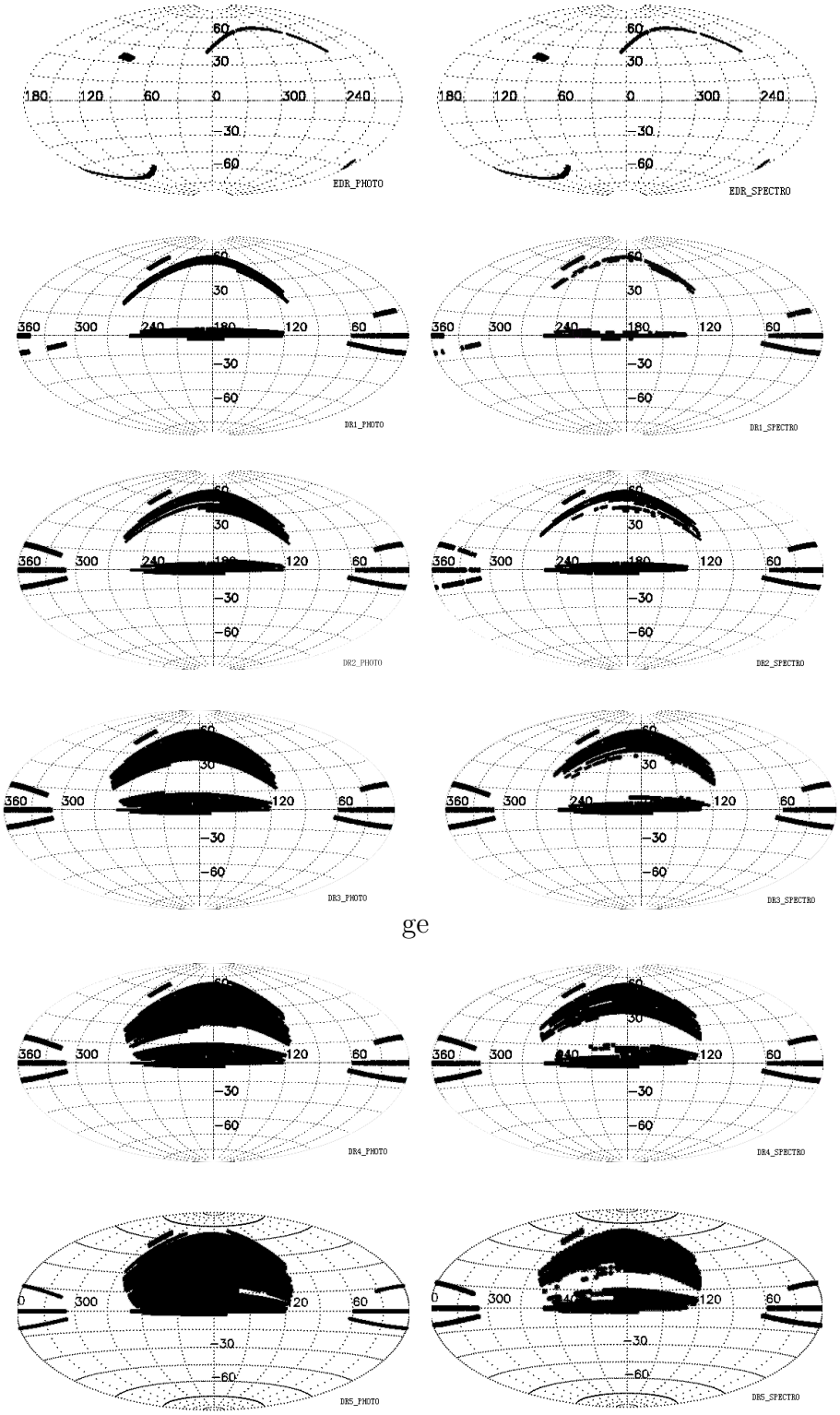
Advanced Study)、日本协作组(the Japan Participation Group)、约翰霍普金斯大学(Johns Hopkins University)、核天体联合研究所(the Joint Institute for Nuclear Astrophysics)、Kavli粒子天体物理和宇宙学研究所(the Kavli Institute for Particle Astrophysics and Cosmology)、韩国科学家团组(the Korean Scientist Group)、中国科学院LAMOST项目(the Chinese Academy of Sciences，Large sky Area Multi-Object fiber Spectroscopic Telescope)、洛杉矶Alamos国家实验室(Los Alamos National Laboratory)、马普天文研究所(MPIA)、马普天体物理研究所(MPA)、新墨西哥州立大学(New Mexico State University)、俄亥俄州州立大学(Ohio State University)、匹兹堡大学(University of Pittsburgh)、普兹茅斯大学(University of Portsmouth)、普利斯顿大学(Princeton University)、美国海军天文台(the United States Naval Observatory)、美国华盛顿大学(the University of Washington)。

2.1.1 SDSS数据产品

至2006年7月，SDSS-I期已经释放了全部的五批数据，它们是2001年6月发布的早期数据产品(EDR，[47])，2003年4月发布的第一批数据产品(DR1，[48])，2004年3月发布的第二批数据产品(DR2，[49])，2004年9月发布的第三批数据产品(DR3，[50])，2005年7月释放的第四批数据(DR4，[51])和2006年7月释放的第五批数据(DR5，[52])。不同期释放数据的信息比较见表2.1。SDSS释放的五批数据的测光和光谱覆盖范围如图2.1所示，图2.2给出了SDSS巡天的五个滤光片的响应曲线。

表 2.1: SDSS目前已经释放的数据产品

日期	名称	缩写	数据量	星系光谱数量	测光面积(deg^2)	光谱面积(deg^2)
2001-6	Early Data Release	EDR	14 $\times 10^6$	39,959	462	462
2003-4	Data Release 1	DR1	53 $\times 10^6$	134,000	2,099	1,360
2004-3	Data Release 2	DR2	88 $\times 10^6$	260,490	3,324	2,627
2004-9	Data Release 3	DR3	141 $\times 10^6$	374,767	5,282	3,732
2005-7	Data Release 4	DR4	180 $\times 10^6$	565,715	6,670	4,783
2006-7	Data Release 5	DR5	215 $\times 10^6$	674,749	8,000	5,740



ge

图 2.1: SDSS已释放数据的测光和光谱覆盖范围

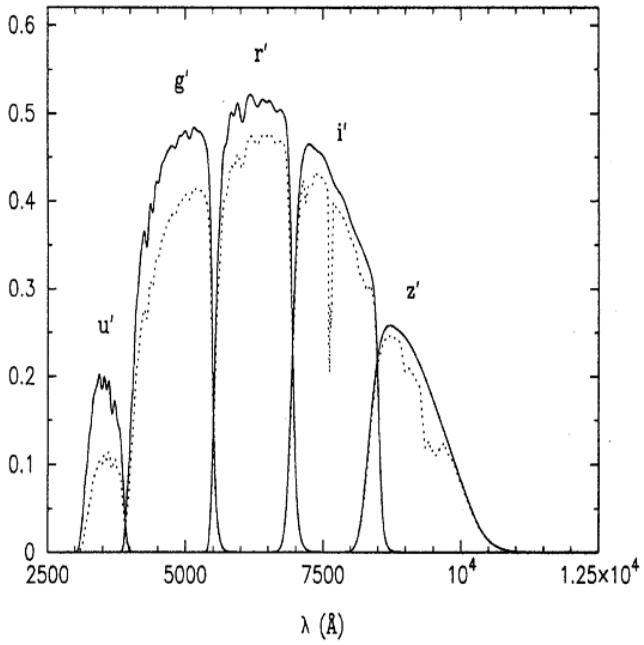


图 2.2: SDSS巡天滤光片的响应曲线。横坐标是波长，纵坐标是量子效率。图中从左至右为u', g', r', i', z'波段，虚线表示经过大气改正的结果[64]。

2.1.2 SDSS巡天样本

SDSS巡天中三类样本具有红移值：主星系样本（Main Galaxy Sample，简称MGS）、亮红星系样本（Luminous Red Galaxy，简称LRG）和类星体样本。主星系样本是通过限制Petrosian r <17.77mag，表面亮度$\mu R_{50} \leq 24.5$，天空亮度，光纤星等以及其他条件产生的。主星系样本大概每平方度中有90个星系，红移的中值为0.104。有关主星系的更多信息，可参阅Strauss等人的文章（2002，[53]）。亮红星系样本的选择在Eisenstein等人的文章（2001，[54]）中给出了详细的介绍。绝大多数的亮红星系样本红移$z < 0.38$。有些大的亮红星系样本$z \sim 0.5$。亮红星系是由一些星等$r \sim 19.5$mag带有红移值的星系团组成。类星体样本是在颜色空间中远离恒星分布的离散点或者是一些与FIRST射电巡天相匹配的点源。类星体样本的红移值可以达到6。详细介绍请参看Richard 等人的文章（2002，[55]）。

下面对SDSS巡天中涉及到的主要参数做简略的介绍：

（一）SDSS巡天中的星等

SDSS数据中提供了几种不同类型的星等，包括Petrosian星等、Model星等、Fibre星等和PSF星等。星等用反双曲正弦（inverse hyberbolic sine，简称asinh）的函数表示[56]。反双曲正弦星等相当于高信噪比下的标准星等，但对于低信噪比甚至负流量值也适用。

$$m = -\frac{2.5}{ln(10)}[asinh(\frac{f/f_0}{2b}) + ln(b)] \tag{2.1}$$

其中f是观测流量，f_0是星等为0时的流量，b是软化参数。反双曲正弦星等的更多信息可以查看Lupton[56]和Stoughton[47]的文章。

（1）Petrosian星等

因为星系并不都具有相同的表面亮度轮廓，也没有形状轮廓，所以很难测量它们的流量。SDSS巡天采用了改良的Petrosian[57]系统，用圆形孔径预测星系的流量，圆形孔径的半径是由方位平均的表面亮度轮廓定义的。Petrosian ratio（$R_{\rm p}$）的定义如下：

$$R_{\rm p}(r) \equiv \frac{\int_{0.8{\rm r}}^{1.25{\rm r}} dr' 2\pi r' I(r') / [\pi(1.25^2 - 0.8^2) r^2)]}{\int_0^{\rm r} dr' 2\pi r' I(r') / (\pi r^2)} \tag{2.2}$$

其中$I(r)$表示平均的表面亮度轮廓。Petrosian radius（$r_{\rm p}$）为在$R_{\rm p}(r_{\rm p})$等于0.2时的半径。Petrosian流量为$N_{\rm p}$等于2时Petrosian半径内的流量

$$F_{\rm P} \equiv \int_0^{N_{\rm p} r_{\rm p}} 2\pi r' dr' I(r') \tag{2.3}$$

（2）Model星等

星系的图像是由拟合光线轮廓实现的。有两种拟合方法：de Vaucouleurs拟合法

$$I(r) = I_0 exp\{-7.67[(r/r_{\rm e})^{1/4}]\} \tag{2.4}$$

和指数拟合法

$$I(r) = I_0 exp(-1.68 r/r_{\rm e}) \tag{2.5}$$

其中I_0和I(r)是半径为0和r时的强度。$r_{\rm e}$是半光度半径。Model星等是两种拟合中较好的那个星等。

（3）Fiber星等和PSF星等

Fibre星等是光纤光谱中的流量，而PSF星等是点源的星等。在星系的研究中很少用到这两个星等参数。SDSS中提供的星等都是未经过红化校正的，校正方法可参考Schlegel[58]的文章。

（二）SDSS星系样本的几个特殊参数

（1）eClass

SDSS星系样本的eClass分类是通过对光谱数据进行主分量分析（Principal Component Analysis, 简称PCA）得到的。eCoeff和eClass分别存放了五个本征系数和一个分类号。eClass是星系分类的参数，取值范围从-0.5到1。

$$\text{eClass} = atan(\frac{eCoeff2}{eCoeff1}) \tag{2.6}$$

（2）petroR50和petroR90

包括了50%和90%的Petrosian流量的半径被称为petroR50和petroR90。

（3）汇聚指数（concentration index，简写c）

汇聚指数$c = petroR90/petroR50$，c与星系的形态有关。通常对于c >2.5的星系属于早型星系，c <2.5的星系属于晚型星系。

（三）SDSS数据常用参数

为更清晰直观地了解SDSS巡天的常用参数，表2.2、2.3分别列出了SDSS星系和类星体的常用参数。

表 2.2: SDSS星系常用参数表

参数名称	简写	描述
specObjID	id	光谱编号
ra	ra	赤经（J2000）
dec	dec	赤纬（J2000）
modelMag	(model)u, g, r, i, z	Model星等
modelMagErr	(model)$uErr, ..., zErr$	Model星等误差
petroMag	(petro)u, g, r, i, z	Petrosian 星等
petroMagErr	(petro)$uErr, ..., zErr$	Petrosian星等误差
extinction	$uExt, ..., zExt$	各波段的消光
dered	(dered)u, g, r, i, z	红化较正星等(modelMag-extinction)
petroR50	$R50$	r波段的50%流量
petroR90	$R90$	r波段的90%流量
fracDev_r	$frac\_r$	r波段fracDev
petroR90/petroR50	c	汇聚指数
modelColor	(model)$u-g, ..., i-z$	model色指数
petroColor	(petro)$u-g, ..., i-z$	petro色指数
z	z	光谱红移
zErr	$zErr$	光谱红移误差
zStatus	$zStau$	光谱红移状态(>2)
zConf	$zConf$	光谱红移置信度
zWarning	zWarning	光谱红移警告
eClass	eClass	eClass光谱类型
specClass	$specClass$	光谱类型（2：星系）

表 2.3: SDSS类星体常用参数表

参数名称	简写	描述
TargetQsoTargeted	$Qsotarget$	观测目标为类星体
SpecQsoConfirmed	$SpecConfirm$	光谱确认为类星体
SpecQsoLargeZ	$largeZ$	光谱红移$z > 0.6$
SpecRa	ra	赤经（J2000）
SpecDec	dec	赤纬（J2000）
SpecZ	z	光谱红移
SpecZerr	$zErr$	光谱红移误差
SpecZConf	$zConf$	光谱红移置信度
SpecZStatus	$zStatus$	光谱红移状态（>2）
SpecZWarning	$zWarning$	光谱红移警告
SpecClass	$sClass$	光谱类型（3:QSO；4:HIZ-QSO）
bestPsfMag	(PSF)u, g, r, i, z	PSF星等
bestPsfMagErr	(PSF)$uErr, ..., zErr$	PSF星等误差
bestExtinction	(PSF)$uExt, ..., zExt$	PSF星等消光
PsfColor	(PSF)$u-g, ..., i-z$	PSF色指数

2.2 2MASS巡天数据

2MASS巡天（Two Micron All Sky Survey，简称2MASS）是由美国国家航天局（National Aeronautics and Space Administration, 简称NASA）和美国国家自然基金会（the National Science Foundation, 简称NSF）资助，由曼切斯特大学和IPAC（Infrared Processing and Analysis Center ）联合实施的近红外巡天项目。为了对整个天空进行红外巡天观测，2MASS使用了两个高度自动化的1.3米的望远镜，一台放在美国亚利桑那Hopkins山上，另一台放在智利的Tololo山上。每台望远镜均装配了3个通道的CCD相机，每个通道包括256×256像素的CCD阵列，这样就可以实现在J、H、Ks三个波段同时观测。南、北半球的望远镜分别在1998年3月和1997年6月开始观测，整个巡天计划在2001年2月

结束，共收集了近25TB的数据。表2.4显示了2MASS的数据释放信息。对全天数据进行评估后得知，2MASS最终获得的数据远好于巡天计划所提出的一级科学要求。

表 2.4: 2MASS释放的数据产品

日期	名称	覆盖面积（$\deg^2$）	百分比（%）
1998.12	2MASS数据样本	63	0.15
1999.5	第一次增量释放	2,483	6
2000.3	第二次增量释放	19,600	47
2003.3	全天释放	41,000	99.998

2.2.1　2MASS数据产品

2MASS数据分为点源星表（Point Source Catalog, 简称PSC）、展源星表（Extended Source Catalog, 简称XSC）和天空的数字图像集。点源星表中包括4.7亿个源在J、H、Ks三个波段的精确位置和测光方面的信息，其中大部分是银河系内的恒星，也包括一些不可分辨的源。由于边缘效应和亮星光的影响，PSC覆盖了99.997% 的天空，略低于2MASS巡天的全天覆盖率，其中90%的源落在$|b| < 30°$的半个天空中。在没有干扰的情况下，其完备极限星等$J \leq$15.8mag、$H \leq$15.1mag、$Ks \leq$14.3mag。

展源星表包括一百六十多万个源在J、H、Ks三波段的位置和测光信息，97%是星系。这些源大多在银道面附近5°范围内。每个展源均含有位置、星等、测光信息、与其他河外星系表的比较和表征源探测质量的标识等信息。展源的空间分辨率受前景星的影响，除了不透明的银盘区域外，探测到的展源遍及整个天空。对XSC进行统计时的测光采用基准椭圆等照度的累积流量，这样得到的星等大致包含了待测星系总流量的85%。表2.5显示了2MASS三个波段的极限星等。2MASS巡天的完备性和准确性如表2.6所示。表2.7描述了2MASS巡天的测光精度和位置精度。

2MASS数字图像集中包括J、H、Ks三波段的四百多万张覆盖全天的FITS（Flexible Image Transport System）图像。它们来自于59731个$8.5' \times 6°$的小巡天区域。每个小区域的数据在每个波段被划分为23个星空图像，其中22个

表 2.5: 2MASS三个波段的极限星等

		星等极限	
波段	波长（μm）	点源（mag）	展源（mag）
J	1.25	15.8	15.0
H	1.65	15.1	14.3
Ks	2.17	14.3	13.5

表 2.6: 2MASS巡天的完备性和准确性

	银纬覆盖范围（Galactic Latitude Range）			
银纬b	$>\|30\|^\circ$	$\|20\|-\|30\|^\circ$	$\|10\|-\|20\|^\circ$	$<\|10\|^\circ$
差量完备性（Differential Completeness）				
点源	0.99	–	–	–
展源	0.90	–	–	–
差量可靠性（Differential Reliability）				
点源	0.9995	0.9995	0.9995	0.9995
展源	0.99	0.99	0.80	–

为512×1024像素，1个为512×698像素。每张图像的大小是$8' \times 16'$，空间分辨率是$4''$[59]。

2.2.2 2MASS重大科学贡献

2MASS巡天由于其自身的优势在天文研究中发挥了重要的作用，目前已取得了一系列成果。其中最重要的有如下几个方面：

（1）描述了银河系大尺度结构。2MASS从不同于以往的角度观测银河系。由于受星际消光影响较小，2MASS揭示了发光天体的分布，也在一定程度上解释了银河系的大尺度结构。2MASS以高分辨率在近红外波段详细地、完整地描

表 2.7: 2MASS巡天的测光精度和位置精度

测光精度	
清晰的点源 (对那些SNR $\gg$ 20的源)	5%
清晰的展源 (对于等照度星等为20mag/sq.arcsec)	10%（对于H <13.8mag）
测光空间分布的均匀性	
点源	4%
展源	10%
最亮的可测量的恒星	
测光偏差（对于Ks >4mag）	<2%
可重复性	5%（对于Ks=8mag） 10%（对于4mag< Ks <8mag）
位置修正后的误差	$0.5''$

述了银河系。

（2）在Ks波段测光普查了星等亮于13.5mag的星系。2MASS最先在Ks波段对亮于13.5mag的星系进行测光普查，普查区域包括了隐带。最终获得的1,500,000个星系样本构成了具有统计意义的数据库，其中有3个波段的测光结果。

（3）提供了对天体物理学具有重要意义的稀有天体的统计研究基础。这些稀有天体或者因温度低而极端偏红，或者在可见光波段消光严重。利用2MASS数据已经找到存在非常冷的恒星的证据，它们比以往所知的所有矮星的温度都低，称之为L型矮星。同时，通过对甲烷分子吸收线的观测，也找到褐矮星存在的证据，这些星被称作T型星[59]。

2.3　基于SDSS与2MASS交叉证认的数据样本

为了测试多波段数据对测光红移预测精度的影响，我们将SDSS巡天与2MASS巡天的测光数据进行交叉证认。交叉证认的原理为：假设在两个星表

中对应源的坐标分别为（α_1，δ_1）、（α_2，δ_2），求它们之间的角距离d。通常情况下，在d很小时角距离可取如下近似公式：

$$\delta = \frac{(\delta_1 + \delta_2)}{2} \tag{2.7}$$

$$d^2 = ((\alpha_1 - \alpha_2)cos\delta)^2 + (\delta_1 - \delta_2)^2 \tag{2.8}$$

设两个星表的误差半径分别为r_1和r_2，通常角距离应满足下列条件：

$$d \leq |r_1| + |r_2| \tag{2.9}$$

或者

$$d \leq 3\sqrt{r_1^2 + r_2^2} \tag{2.10}$$

即交叉证认的半径应满足（2.9）式或者（2.10）式[60]。在我们的工作中，r_1是SDSS测光星表的误差半径，$r_1 = 1''$；r_2是2MASS展源星表的误差半径，$r_2 = 2''$。

基于上述原理，我们交叉证认了SDSS巡天的星系测光星表与2MASS巡天的展源测光星表。考虑到对应源的准确性与可靠性，我们只考虑一对一的天体且没有缺值的测光数据。提取来自SDSS星系测光星表的参数如表2.2所示，来自2MASS展源星表的三个等照度基准的椭圆口径星等（j_m_k20fe、h_m_k20fe、k_m_k20fe）。表2.8描述了SDSS与2MASS的测光有效波长、波长范围、极限星等及半峰全宽的比较。

表 2.8: SDSS与2MASS的测光波段及各波段的特性

波段	巡天名称	$\lambda_{\rm eff}(\mathring{A})$	$\Delta\lambda(\mathring{A})$	极限星等（mag）	FWHM(arcsec)
u	SDSS	3551	600	22.0	1-2
g	SDSS	4686	1400	22.2	1-2
r	SDSS	6165	1400	22.2	1-2
i	SDSS	7481	1500	21.3	1-2
z	SDSS	8931	1200	20.5	1-2
J	2MASS	12500	1620	15.0	2-3
H	2MASS	16500	2510	14.3	2-3
Ks	2MASS	21700	2620	13.5	2-3

第三章　测光红移算法研究

回归分析以客观事物变量间的统计关系为重要研究对象，基于对客观事物进行大量实验和观察，寻找隐藏在不确定的现象中的统计规律的统计方法。既有几个自变量对一个因变量的回归问题，也有多个自变量对多个因变量的回归问题。统计回归问题就是从已知的资料出发，建立自变量x对因变量y的回归方程，以便利用这个回归方程进行新的预测。回归问题有多种分类：

（1）根据因变量y与自变量x之间是否存在线性关系，回归问题可以分为线性回归和非线性回归。非线性回归包括多项式回归、指数回归、常用的神经网络、最近邻、以及支持矢量机等。

（2）根据给定资料数据的已知条件不同，回归问题可分为参数回归和非参数回归。如果数据分布满足某种分布类型，即已知自变量和因变量之间有经验公式，此类回归问题称为参数回归。在参数回归中，主要是确定经验公式中的未知参数的值。如果数据的分布是不确定的，自变量和因变量之间关系未知，这种回归问题称为非参数回归。非参数回归首先要对数据的分布进行估计，然后才能确定回归函数[61]。

对于参数回归来说，在模型和样本数据关系确定的前提下，估计函数关系中的参数并检验所设定的关系。如果模型的函数关系通过检验被证明是成立的，那么回归结果可以外延，其推测和预测都有较高的精度，模型的参数具有明确的意义。但是对于非参数回归来说，回归函数的形式是随意的，没有任何约束，解释变量和被解释变量的分布也很少限制，因而有较大的适应性。常用的非参数回归方法包括核回归、K近邻和局部加权回归。

天文中的测光红移预测问题即属于回归问题，因而各种用于回归的方法或模型均可以用来预测红移。基于以往工作对测光红移算法的研究和探索，我们研究了四种预测测光红移的算法，分别为：颜色-星等-红移关系、多变量多项式回归、支持向量机和核回归。

3.1 颜色-星等-红移关系法

3.1.1 原理

星系的红移不仅与星系的颜色、光谱类型有关，也和星等有关。Csabai[38]和Hsieh[63]提出如果考虑一些具有相似测光属性（例如：双色图、颜色星等图）的星系，可以提高测光星系的精度。测光性质与红移间没有直接的线性关系，因此不可以用简单的函数表示它们之间的关系。为了解决这个问题，我们构建了矩阵来数字化红移、颜色和星等的关系。具体做法如下：

（1）为了表述方便，下面及以后的实验中用到的星等均为红化校正星等。将整个样本分成7个小样本集（R1∼R7），如表3.1所示。r星等小于16mag属于R1，大于16mag小于17mag属于R2，17mag和18mag之间属于R3，18mag和19mag之间属于R4，19mag和20mag之间属于R5，20mag和21mag之间属于R6，21mag和23mag之间属于R7。

表 3.1: 按r星等划分的子样本

子样本	星等范围	星系数目
R1	$r < 16.0$mag	43,459
R2	$16.0\text{mag} < r < 17.0\text{mag}$	120,373
R3	$17.0\text{mag} < r < 18.0\text{mag}$	235,456
R4	$18.0\text{mag} < r < 19.0\text{mag}$	42,921
R5	$19.0\text{mag} < r < 20.0\text{mag}$	15,462
R6	$20.0\text{mag} < r < 21.0\text{mag}$	1,952
R7	$21.0\text{mag} < r < 23.0\text{mag}$	1,429
R	整个样本	461,170

（2）将R1∼R7的样本按$u-g$与$g-r$和$g-r$与$r-i$画双色图。

（3）将每个子集等分成400×400的小方格，对于每个小方格，如果落入此小方格中的星系个数超过25个，我们就计算落入其中星系的中值作为当前方格的红移。如果落入当前小方格内的星系数目不足25，我们就将小方格扩大至2个小方格的大小，继续计算落入其中星系的个数，如果此时落入其中的星系数目超过25，就按上述方法计算红移。这种扩大最多只能扩到5个小方格的大小。这个

过程就是自适应平滑。它是将每个小方格内的红移信息进行平滑。通过上述几步，我们得到颜色星等矩阵图。对于任意星系，我们只要根据它的颜色和星等就可以在这些矩阵图中找到它们对应的红移。我们将$u-g$、$g-r$形成的矩阵称为$\mathrm{CMR_I}$，$g-r$、$r-i$形成的矩阵称为$\mathrm{CMR_{II}}$[62]。如图3.1所示：

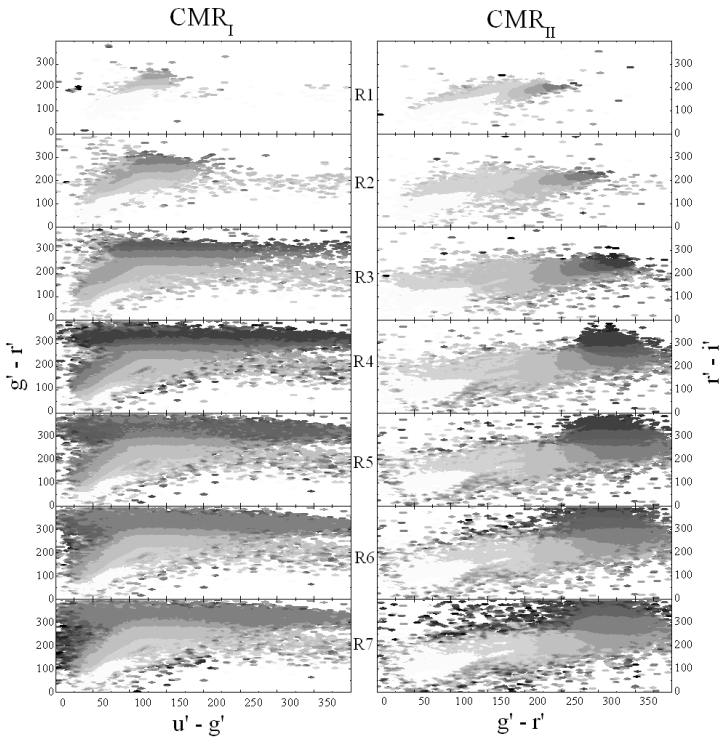


图 3.1: 以图的形式显示的CMR矩阵。红移值由灰度值表示。浅灰色表示低红移，深灰色表示高红移。图中每一排代表不同的星等值，此星等值来自于表3.1。

3.1.2　样本

从SDSS DR4星表中找到了459,584个星系的五色测光数据和光谱红移作为样本，并分别对各个波段做了消光改正。因为SDSS星系的红移由于极限星等的限制一般都小于0.5，为了测试此方法对高红移的有效性，我们在NASA/IPAC河外天体数据库（NED）中找到1,586个红移大于0.5的星系样本。由这两部分数据构成了我们的样本S，共计461,170个样本。

3.1.3 结果与讨论

我们用上述样本对CMR矩阵进行了测试。图3分别显示了$\mathrm{CMR_I}$和$\mathrm{CMR_{II}}$的测试结果。对于$\mathrm{CMR_I}$来说，当$z_\mathrm{I} < 0.3$ 时预测结果比较理想；对$\mathrm{CMR_{II}}$，

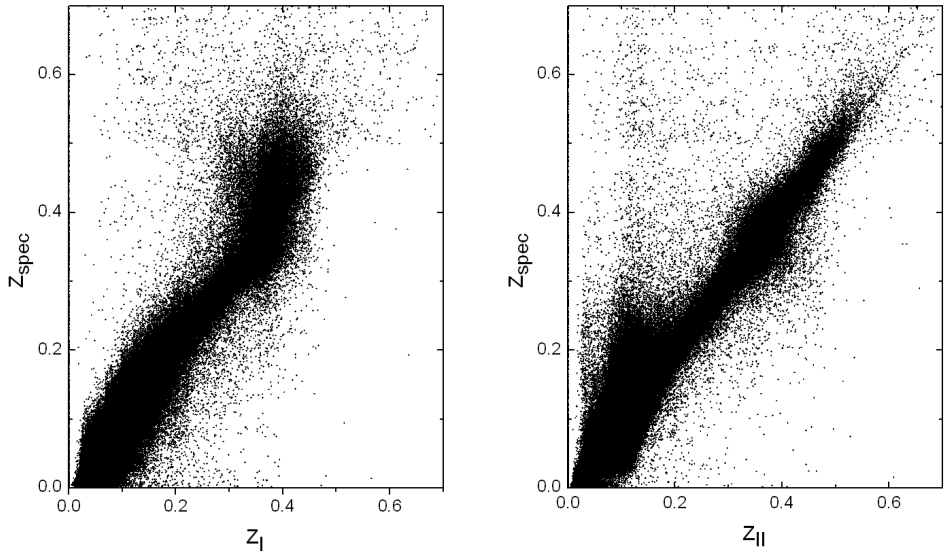


图 3.2: 用CMR_I，CMR_II方法预测的测光红移与光谱红移的对比图

当$z_\mathrm{II} > 0.2$ 时预测结果较好。这种相互补充的结果可以让我们将两个结果合并起来。首先，根据$\mathrm{CMR_{II}}$计算z_II；其次，如果$0.05 < z_\mathrm{II} < 0.2$时，根据$\mathrm{CMR_I}$计算$z_\mathrm{I}$，用$z_\mathrm{I}$的值替代$z_\mathrm{II}$。对于那些落在$z_\mathrm{I} < 0.05$并且$z_\mathrm{II} < 0.1$ 区间内的星系，我们用$\mathrm{CMR_I}$计算z_I作为星系的红移。通过这种合并的方法，我们得到的红移预测剩余标准偏差$\sigma_\mathrm{rms} = 0.0320$（见图3.3）。计算$\sigma_\mathrm{rms}$的公式为

$$\sigma_\mathrm{rms} = \sqrt{\langle (z_\mathrm{phot} - z_\mathrm{spec})^2 \rangle} \tag{3.1}$$

其中z_phot代表用CMR方法预测的测光红移的值，z_spec代表来自有SDSS星表的光谱红移的值。对于这种方法来说，如果星系的颜色落在CMR矩阵外的话，我们就无法得到红移值了。通过样本测试获知，当$r = 21$mag时，损失率大概为5%；当$r = 23$mag时，损失率大概为10%。

CMR方法从原理上来说比较简单，天文学家很容易理解，而且计算速度很快，46万数据在一分钟内就可以得到结果。但是CMR方法的预测精度不是很高。

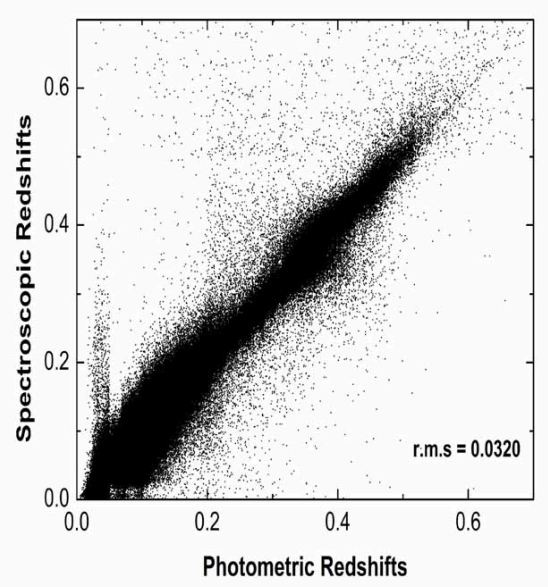


图 3.3: 用CMR方法得到的测光红移与SDSS的光谱红移的对比图.

由于CMR方法按r星等分类，所以在星等边界的星系就被强制性地划在某个小样本集内，这样未免带入一些系统误差。而且由于CMR算法本身的限制，采用自适应的平滑，使某些星系失去固有的特征，这也是精度不够高的主要原因之一。

我们利用CMR方法做了一个web服务。用户可以上传自己的数据到服务器上（如图3.4所示），经过计算后，服务器将结果返还给用户（如图3.5所示），用户既可用浏览器查看，亦可将结果下载到本地保存。

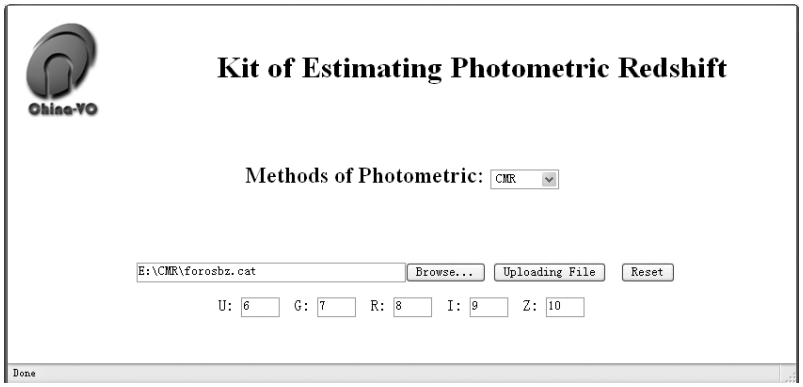


图 3.4: CMR方法的Web服务图。用户上传一个包含五色测光的文本文件。用“Browse...”按钮选择要上传的文件的位置。第二排的文本框中指定五个星等在文件中的列数，例如:“6”代表u星等在文件中的第六列...。然后点击“Uploading file”按钮。

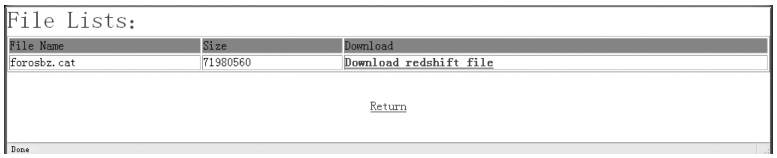


图 3.5: 用CMR方法预测红移的返回结果截图。在返回页面中显示了结果文件的名称、大小。用户点击“Download redshift file”可以用浏览器显示结果，或者右键点击下载结果文件。

3.2 多变量多项式回归

3.2.1 原理

研究一个因变量和一个或多个自变量间多项式的回归分析方法，称为多项式回归（Multiple Polynomial Regression，简称MPR）。若自变量只有一个，成为一元多项式回归；若自变量有多个，称为多元多项式回归，即多变量多项式回归。

一元m次多项式回归方程为：

$$y = b_0 + b_1 x + b_2 x^2 + ... + b_{\mathrm{m}} x^{\mathrm{m}}$$

二元二次多项式回归方程为：

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2$$

在一元回归分析中，若因变量y与自变量x的关系为非线性的，但是当找不到适当的函数曲线来拟合时，则可以采用一元多项式回归。多项式回归的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，直至适宜为止。事实上，多项式回归在回归分析中占有重要的地位，因为任意函数都可以分段由多项式来逼近，因此可以处理很多非线性回归问题。不论因变量与其自变量的关系如何，我们总可以用多项式回归来进行分析。

多变量多项式回归是一种重要的多次回归模型，是非线性的，且模型中有交叉项产生。它已经被广泛地应用到预报和相关性分析中。多变量多项式回归是多输入系统，其原理是在自变量和因变量之间产生通常的逻辑关系。在实验中，我们的训练集既包括自变量（$u-g$、$g-r$、$r-i$、$i-z$）又包括因变量（$z_{\rm spec}$）。通过训练，得到了回归关系，可以用一个数学表达式来表示。通常来说，训练样本越完备越有代表性，红移的预测精度越高。多变量多项式回归已经被应用到很多领域，例如：多变量相关性分析、时间序列和判别分析。在天文中，多变量多项式回归也得到了应用。例如Connolly（1995，[37]）使用$UJFN$四波段和红移作为训练样本，预测了370个红移最大为0.5的星系的测光红移。线性拟合的预测剩余标准偏差$\sigma_{\rm rms}=0.057$，二次拟合的预测剩余标准偏差$\sigma_{\rm rms}$=0.047。Sowards（2000，[65]）使用了2195个来自于Las Campanas巡天的带有红移的星系样本预测测光红移，这批样本都分布低红移区，最大红移为0.25，预测剩余标准偏差$\sigma_{\rm rms}$=0.035。

目前，网络上有很多MPR软件包，在考察过它们的运行效率和健壮性后，我们采用了Christian Borgelts的Java程序软件包。该程序包的优点在于：效率和预测精度较高，适合处理大数据量。在我们的实验中，训练33万数据只需1分钟。

3.2.2　样本

在研究多变量多项式回归算法时，我们使用了SDSS DR4的数据。在DR4数据中，大概有2400万星系的五色测光数据，但是只有46万星系有红移。在实验过程中，我们使用了两个样本集。样本一（S1）共有333,287个样本，由以下

三步骤得到:（1）光谱红移置信度大于等于0.95;（2）红移警告等于0;（3）五个星等必须在SDSS的极限星等之内，也就是说$u < 22.0$mag，$g < 22.2$mag，$r < 22.2$mag，$i < 21.3$mag，$z < 20.5$mag。在样本一的基础上，我们又加了额外的限制，产生了样本二（S2）。样本二的数据$r < 17.5$mag并且$0.01 < z < 0.5$，共计247,511个样本，样本的顺序是随意的、无重复的。下面的实验均是以上述两个样本为基础的。

3.2.3 结果与讨论

为了研究训练和测试样本的数目是否对测光红移的精度有影响，我们做了下面的试验。首先，我们将S1随意地分成6份。每一份进行三组试验。对每组测试和训练样本分别进行线性拟合、二次拟合和三次拟合。这样我们就得到了54组实验结果。对线性拟合来说，预测剩余标准偏差最小是0.0333，二次拟合最小的剩余标准偏差达到$\sigma_{\rm rms}$=0.0281，三次拟合最小的剩余标准偏差为$\sigma_{\rm rms}$=0.0278。表3.2显示了不同训练样本和测试样本对应的剩余标准偏差$\sigma_{\rm rms}$。

同样按照上述步骤，我们用S2做样本，得到了表3.3中的结果。

对于二次回归来说，在表3.2中最小的剩余标准偏差$\sigma_{\rm rms}$达到0.0281，而在表3.3中最小的剩余标准偏差为0.0256，下降了大约9%。由于S2样本去掉了高红移的星系样本，所以弥散减小。

在S2样本中，当训练样本为100,000，测试样本为147,511时，线性回归得到的剩余标准偏差$\sigma_{\rm rms}$达到最小值0.0291。线性回归公式为：

$z_{\rm phot}$ = -0.065630 $(u'-g')$ + 0.251205 $(g'-r')$ + 0.004186 $(r'-i')$ - 0.096274 $(i'-z')$ + 0.024336。

在训练样本为200,000时，三组测试样本的剩余标准偏差都达到最小值0.0256。其中测试样本为47,511时，二次回归的公式为：

$z_{\rm phot}$ = 0.025692 $(u'-g')^2$ + 0.125040 $(u'-g')$ $(g'-r')$ - 0.024228 $(g'-r')^2$ -0.093342 $(u'-g')$ $(r'-i')$ - 0.020939 $(g'-r')$ $(r'-i')$ + 0.034685 $(r'-i')^2$ -0.030826 $(u'-g')$ $(i'-z')$ - 0.023166 $(g'-r'$ $(i'-z')$ - 0.029854 $(r'-i')$ $(i'-z')$ -0.013204 $(i'-z')^2$ - 0.227018 $(u'-g')$ + 0.119712 $(g'-r')$ + 0.220097 $(r'-i')$ -0.023471 $(i'-z')$ + 0.148448

图3.6中左图是用S1样本得到的二次回归剩余标准偏差最小时的测光红移和光谱红移散点图，右图是用S2为样本得到的二次回归剩余标准偏差最小时的

表 3.2: S1样本中不同训练样本与测试样本对应的剩余标准偏差$\sigma_{\rm rms}$

训练样本数	测试样本数目	线性回归$\sigma_{\rm rms}$	二次回归$\sigma_{\rm rms}$	三次回归$\sigma_{\rm rms}$
50,000	50,000[a]	0.0343	0.0285	0.0273
	283,287[b]	0.0334	0.029	0.0342
	333,287[c]	0.0336	0.0289	0.0333
100,000	100,000[a]	0.0339	0.0284	0.0277
	233,287[b]	0.0333	0.0291	0.029
	333,287[c]	0.0335	0.0289	0.0287
150,000	150,000[a]	0.0338	0.0286	0.0278
	183,287[b]	0.0333	0.0289	0.0286
	333,287[c]	0.0335	0.0288	0.0283
200,000	200,000[a]	0.0334	0.0287	0.0278
	133,287[b]	0.0336	0.0285	0.029
	333,287[c]	0.0335	0.0287	0.0283
250,000	250,000[a]	0.0335	0.0287	0.0279
	83,287[b]	0.0334	0.0285	0.0278
	333,287[c]	0.0335	0.0286	0.0278
300,000	300,000[a]	0.0335	0.0287	0.0278
	33,287[b]	0.0333	0.0281	0.0278
	333,287[c]	0.0335	0.0286	0.0278

[a] 训练和测试用同样的样本

[b] S1中的一部分作为训练样本，另外一部分作为测试样本

[c] S1中的一部分作为训练样本，整个S1作为测试样本

测光红移和光谱红移散点图。 从图3.7中我们可以看出，随着训练样本数的增大，线性和二次回归的剩余标准偏差变化比较缓慢，而三次回归的结果变化很大。对于线性和二次回归来说，只要训练样本有代表性，即使训练样本数目很小（50,000）剩余标准偏差仍然较小。所以样本数目对线性和二次回归的影响不是很大。对三次回归来说训练集的样本越大，偏差越小。训练样本对三次回归

表 3.3: S2样本中，不同测试和训练样本对应的剩余标准偏差$\sigma_{\rm rms}$

训练样本数	测试样本数目	线性回归$\sigma_{\rm rms}$	二次回归$\sigma_{\rm rms}$	三次回归$\sigma_{\rm rms}$
50,000	50,000[a]	0.0294	0.025	0.0242
	197,511[b]	0.0293	0.027	0.0384
	247,511[c]	0.0293	0.0267	0.0359
100,000	100,000[a]	0.0296	0.0257	0.0248
	147,511[b]	0.0291	0.0263	0.0348
	247,511[c]	0.0293	0.0261	0.0311
150,000	150,000[a]	0.029	0.0257	0.0247
	97,511[b]	0.0296	0.0263	0.0369
	247,511[c]	0.0293	0.0261	0.03
200,000	200,000[a]	0.0293	0.0256	0.0248
	47,511[b]	0.0292	0.0256	0.0263
	247,511[c]	0.0293	0.0256	0.0251

[a] 训练和测试用同样的样本

[b] S2中的一部分作为训练样本，另外一部分作为测试样本

[c] S2中的一部分作为训练样本，整个S2作为测试样本

的影响很大。综合考虑了剩余标准偏差的大小，回归算法的健壮性和容易理解性，我们认为二次回归是比较理想的预测测光红移的方法。虽然在某些情况下，三次回归的剩余标准偏差小于二次回归，但是由于三次回归的波动性过大，所以我们不予以考虑。此外，为了验证小训练样本对线性和二次回归算法的影响，我们随机从S2 中选择了10,000个星系作为训练样本，结果得到线性剩余标准偏差为0.0294，二次回归的剩余标准偏差为0.0263。这再一次证明前面的说法：只要训练样本有代表性，即使样本数很少，MPR方法仍然可以得到令人满意的结果。

与其它预测测光红移方法比较来说，MPR是训练集方法中原理最简单的一种。它可以在有限的步骤内绝对收敛，产生具有最小参数的拟合结果，因此MPR避免了过度拟合的问题。虽然用ANN方法预测测光红移的剩余标准偏差

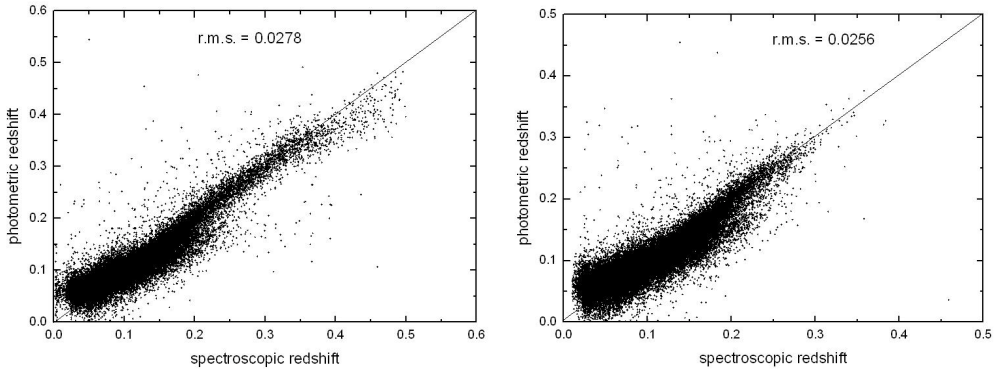


图 3.6: 左图: S1样本的测光和光谱红移对比散点图。其中训练样本为300,000，测试样本为33,287。右图: S2样本的测光和光谱红移对比散点图。其中训练样本为200,000，测试样本为47,511。

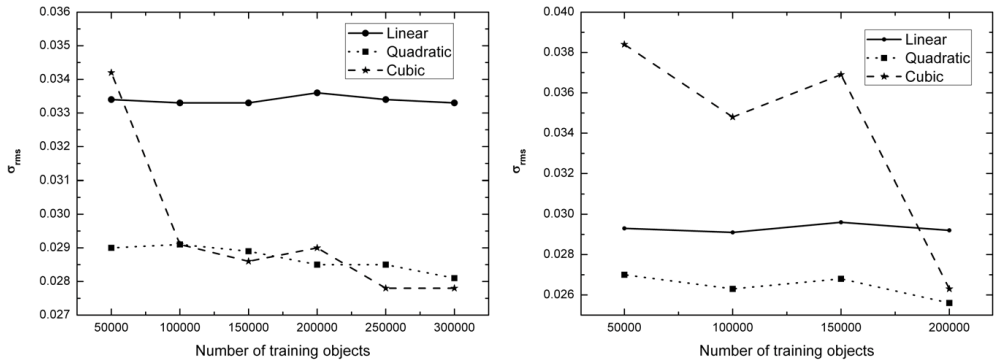


图 3.7: 左图: S1中，训练样本数与剩余标准偏差$\sigma_{\rm rms}$的关系。右图: S2中，训练样本数与剩余标准偏差$\sigma_{\rm rms}$的关系。

小于MPR，但需花费很多时间训练网络。而MPR不需要选择训练模型，它生成的数学表达式简化了拟合过程。另外，MPR算法的计算量较小，所以适合大数据量的研究。与其他的训练集方法一样，MPR不具有预测红移外推的能力，也就是说对于暗的或者大红移的星系，由于缺少训练样本，所以预测精度很低。对于那些训练样本中没有出现的样本，但在测试样本中出现，也就无法得到精确的红移值。在此次实验中，MPR方法只适用于r星等亮于22.2的星系。MPR方法存在的另一个问题是不适合高维的输入。如果输入参数过多时，可以采用降维方法如主分量分析（Principal Component Analysis，简称PCA）方法减少输入参量，然后再用MPR预测红移。

3.3 支持矢量机

支持矢量机（Support Vector Machines，简称SVMs）是一种基于统计学习理论的一般性构造学习方法，其理论是由Vapnik[40]于1995年提出，主要思想是：在高维空间内利用线性函数的对偶核，并通过内积空间的向量运算来处理线性不可分的数据。其优点在于优化对偶理论使高维特征空间中的模型参数易于计算，且运算的复杂度与问题的维数关系不大。

3.3.1 原理

支持矢量机利用结构风险最小化的原理，采用最小的VC维数（Vapnik-Chervonenkis Dimension）创建分类器。若VC维数很低，误差概率会很小，这意味着有较好的推广性。用线性分割的超平面构造分类器。而对一些问题在原始空间中是线形不可分的情况，其将原始空间非线性地转化到更高维的特征空间中去。在这个特征空间中，其很容易找到一个最优的线性分割平面，即相对于训练样本，分类器具有很大的分界面。

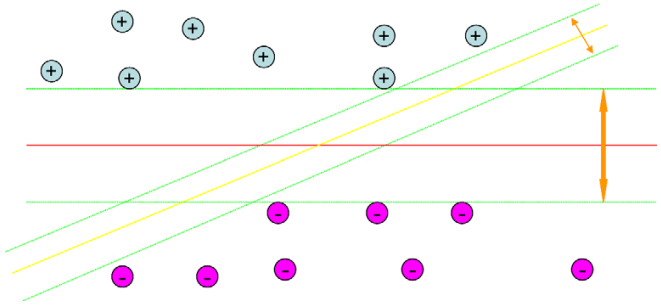


图 3.8: SVM分类问题："+" 代表一类；"-" 代表另一类

考虑线性可分的两类问题如图3.8，"+" 代表一类；"-" 代表另一类。假设存在训练样本（x_1, y_1），...，（x_l, y_l），$x \in R^n$, $y \in \{+1, -1\}$，l为样本数，n为输入维数，在线性可分的情况下就会有一个超平面$(\omega \cdot x) + b = 0$使得这两类样本完全分开，该超平面满足条件：

$$(\omega \cdot x_i) + b > 0, \qquad y_i = +1$$

$$(\omega \cdot x_i) + b < 0, \qquad y_i = -1$$

这等价于

$$\begin{aligned} (\omega \cdot x_i) + b \geq 1, \qquad & y_{\mathrm{i}} = +1 \\ (\omega \cdot x_i) + b \leq -1, \qquad & y_{\mathrm{i}} = -1 \end{aligned} \tag{3.2}$$

也可表示为

$$y_i[(\omega \cdot x_i) + b] \geq 1, \qquad i = 1, ..., l \tag{3.3}$$

如果训练数据可以无误差地被分开，而且每一类数据离超平面最近的向量与超平面之间的距离最大，则称这个超平面为最优超平面。如图3.8所示，水平的分界超平面为最优超平面。

为得到最优超平面，需要找到满足上述条件的超平面，最大化超平面与任意类训练样本的最小距离或最大化分类边界距离。处于最大的分类边界上的点为支持向量。离超平面距离最近的两个点到超平面的距离之和为：

$$\rho(\omega, b) = min_{\{x_{\mathrm{i}}|y_{\mathrm{i}}=1\}} \frac{\omega \cdot x_{\mathrm{i}} + b}{|\omega|} - max_{\{x_{\mathrm{i}}|y_{\mathrm{i}}=-1\}} \frac{\omega \cdot x_i + b}{|\omega|} \tag{3.4}$$

要想使（3.4）式的值最大，由（3.1）式可得：

$$\rho(\omega, b) = min_{\{x_{\mathrm{i}}|y_{\mathrm{i}}=1\}} \frac{1}{|\omega|} - max_{\{x_{\mathrm{i}}|y_{\mathrm{i}}=-1\}} \frac{-1}{|\omega|} \tag{3.5}$$

$$\Leftrightarrow \rho(\omega, b) = \frac{2}{\sqrt{\omega \cdot \omega}} \tag{3.6}$$

因此，求解最优超平面，即为相对于矢量ω和标量b，求解下式的最小值

$$\phi(\omega) = \frac{1}{2}\omega \cdot \omega = \frac{1}{2}\|\omega\|^2 \tag{3.7}$$

优化函数$\phi(\omega)$为二次型，约束条件是线性的，因此这是个典型的二次规划问题，可由拉格朗日乘子法求解，引入拉格朗日乘子$\alpha_{\mathrm{i}} \geq 0, i = 1, 2, ..., l$:

$$L(\omega, b, \alpha) = \frac{1}{2}\omega \cdot \omega - \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}\{[(x_{\mathrm{i}} \cdot \omega) + b]y_{\mathrm{i}} - 1\} \tag{3.8}$$

相对于矢量ω和标量b，L取最小值，此时的矢量ω和标量b分别代表ω_0和标量b_0; 相对于拉格朗日乘子α_{i}，L取最大值，α_{i}记为α_{i}^0。L的极值点称为鞍点。对L求导

$$\frac{\partial L(\omega, b, \alpha)}{\partial b} = 0 \tag{3.9}$$

$$\Leftrightarrow \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}^{0} y_{\mathrm{i}} = 0 \tag{3.10}$$

$$\frac{\partial L(\omega, b, \alpha)}{\partial \omega} = 0 \tag{3.11}$$

$$\Leftrightarrow \omega_0 - \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}^{0} x_{\mathrm{i}} y_{\mathrm{i}} = 0 \tag{3.12}$$

从而得到最优超平面的几个特征：

（1）从式（3.10）可以得到参数α_{i}^{0}的约束方程：

$$\sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}^{0} y_{\mathrm{i}} = 0 \qquad \alpha_{\mathrm{i}}^{0} \geq 0 \qquad i = 1, 2, ..., l \tag{3.13}$$

（2）由公式（3.12）可得矢量ω_0是训练样本的矢量的线性叠加：

$$\omega_0 = \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}^{0} x_{\mathrm{i}} y_{\mathrm{i}} \qquad \alpha_{\mathrm{i}}^{0} \geq 0 \qquad i = 1, 2, ..., l \tag{3.14}$$

（3）在矢量ω_0的展开式中，只有那些支持矢量的参数ω_0^0值不为零：

$$\omega_0 = \sum_{support vectors} \alpha_{\mathrm{i}}^{0} x_{\mathrm{i}} y_{\mathrm{i}} \qquad \alpha_{\mathrm{i}}^{0} \geq 0 \tag{3.15}$$

由Kühn-Tucker定理可知：最优超平面的充分必要条件是分割超平面要满足下面的条件：

$$\alpha_{\mathrm{i}}^{0}\{[(x_{\mathrm{i}} \cdot \omega_0) + b_0] y_{\mathrm{i}} - 1\} = 0, \qquad i = 1, 2, ..., l \tag{3.16}$$

将这些结果代入L中：

$$W(\alpha) = \frac{1}{2}\omega \cdot \omega - \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}\{[(x_{\mathrm{i}} \cdot \omega) + b] y_{\mathrm{i}} - 1\} \tag{3.17}$$

$$= \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}} - \frac{1}{2} \sum_{\mathrm{i,j}=1}^{l} \alpha_{\mathrm{i}} \alpha_{\mathrm{j}} y_{\mathrm{i}} y_{\mathrm{j}} (x_{\mathrm{i}} \cdot x_{\mathrm{j}}) \tag{3.18}$$

上式取最大值需在非负象限中，即

$$\alpha_{\mathrm{i}}^{0} \geq 0, \qquad i = 1, 2, ..., l \tag{3.19}$$

且在下面的条件下

$$\sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}} y_{\mathrm{i}} = 0 \tag{3.20}$$

得到这个问题的解，就可以建立指示函数：

$$f(x) = sign(\sum_{\mathrm{supportvectors}} y_i \alpha_{\mathrm{i}}^{0} (x_{\mathrm{i}} \cdot x) - b_0) \tag{3.21}$$

这里x_{i}为支持向量，α_{i}^{0}为拉格朗日乘子，b_0为临界值：

$$b_0 = \frac{1}{2}[(\omega_0 \cdot x^*(1)) + (\omega_0 \cdot x^*(-1))] \tag{3.22}$$

其中$x^*(1)$是任何属于第一类的支持矢量，$x^*(-1)$则为任何属于第二类的支持矢量。

这种解法仅对线性可分的数据适用，而对线性不可分的数据需略加修改，即

$$0 \leq \alpha_{\mathrm{i}}^{0} \leq C$$

其中C是一个预先假设的常数。

从上面的推导，可得到最优超平面的优点：

（1）在拉格朗日乘子α_{i}不为零的情况下，最优超平面主要是由支持向量来定义的；

（2）最优超平面的建立不直接依赖于所处理问题的维数；

（3）最优超平面的描述也不直接依赖于所处理问题的维数。

在线性不可分的情况下，支持向量机的主要思想是将输入矢量非线性地映射到高维特征空间中，在高维空间中寻找最优超平面。

非线性映射就是将上面的标积$x_{\mathrm{i}} \cdot x$变为核函数$K(x_{\mathrm{i}} \cdot x)$，这样3.18变成

$$W(\alpha) = \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}} - \frac{1}{2} \sum_{\mathrm{i,j}=1}^{l} \alpha_{\mathrm{i}} \alpha_{\mathrm{j}} y_{\mathrm{i}} y_{\mathrm{j}} K(x_{\mathrm{i}} \cdot x_{\mathrm{j}}) \tag{3.23}$$

满足上面的条件下

$$\sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}} y_{\mathrm{i}} = 0$$

$$0 \leq \alpha_{\mathrm{i}}^{0} \leq C$$

其中C是一个预先假设的常数。求解式（3.23），就可以建立指示函数：

$$f(x) = sign(\sum_{\mathrm{supportvectors}} y_i\alpha_{\mathrm{i}}^0 K(x_{\mathrm{i}} \cdot x) - b_0) \tag{3.24}$$

其中

$$\omega_0 \cdot x = \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}} y_{\mathrm{i}} K(x_{\mathrm{i}} \cdot x)$$

$$b_0 = \frac{1}{2}\sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}} y_{\mathrm{i}} [K(x_{\mathrm{i}}, x^*(1) + K(x_{\mathrm{i}}, x^*(-1))]$$

偏差b_0由两个支持矢量来计算，但为了可靠性，可以用边界上的所有支持矢量计算求得。如果核函数中含有偏差项，偏差可以融入到核函数中。这样分类器会简化：

$$f(x) = sign(\sum_{\mathrm{supportvectors}} y_i\alpha_{\mathrm{i}}^0 K(x_{\mathrm{i}} \cdot x)) \tag{3.25}$$

从而简化了最优化问题[60]。

通过引入一个可选的损失函数，支持矢量机也能用于回归问题。这个损失函数必须进行适当回归，使之包含一个距离的度量。图3.9描述了几种可选的损失函数。 其中（a）对应于最小平方误差标准；（b）是一种拉普拉斯损失函数，对于异常点不如（a）敏感；当数据的分布未知时，Huber提出的（c）具有最佳性能。然而上述三个损失函数并不能使得支持向量稀疏。为了解决这个问题，Vapnik提出了（d），这是对Huber损失函数的一个近似，能使得支持矢量变得很稀疏。

（一）线性回归

考虑下面的拟合问题，

$$(y_1, x_1), ..., (y_l, x_l), x \in R^{\mathrm{n}}, y \in R \tag{3.26}$$

其中线性函数

$$f(x) = (\omega \cdot x) + b \tag{3.27}$$

通过使得下式

$$\phi(\omega, \xi^+, \xi^-) = \frac{1}{2}\|\omega\|^2 + C(\sum_{\mathrm{i}}^{l} \xi_{\mathrm{i}}^- + \sum_{\mathrm{i}}^{l} \xi_{\mathrm{i}}^+) \tag{3.28}$$

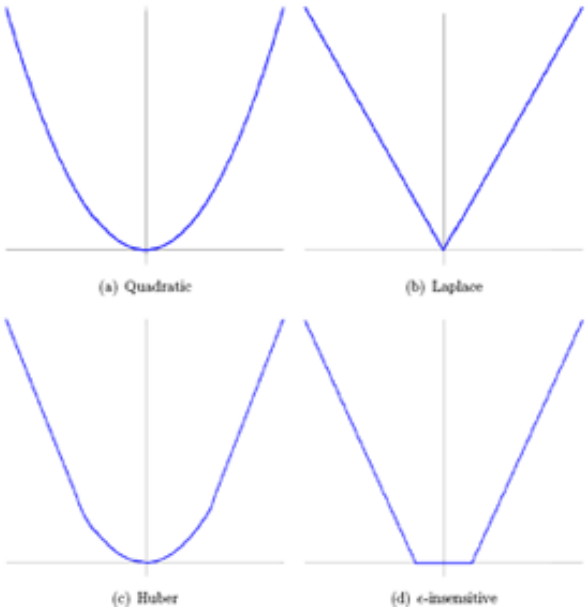


图 3.9: SVMs几种损失函数

最小，获得最佳的回归函数。其中C是一个事先确定的数，ξ^+，ξ^-是表征系统输出上下限的松弛变量。

（1）ε-迟钝损失函数

使用图3.9(d)描述的ε-迟钝损失函数：

$$L_\varepsilon(y) = \begin{cases} 0, & for \quad |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & otherwise \end{cases} \tag{3.29}$$

其解由下面的最大值问题给出：

$$\max_{\alpha,\alpha^*} W(\alpha, \alpha^*) = \max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i.x_j)$$

$$+\sum_{i=1}^{l}\alpha_i(y_i - \varepsilon) - \alpha_i^*(y_i + \varepsilon) \tag{3.30}$$

或者

$$\bar{\alpha}, \bar{\alpha}^* = arg\min_{\alpha,\alpha^*} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i.x_j)$$

$$-\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}}-\alpha_{\mathrm{i}}^{*})y_{\mathrm{i}}+\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}}+\alpha_{\mathrm{i}}^{*})\varepsilon \tag{3.31}$$

约束如下，

$$0\leq\alpha_{\mathrm{i}},\alpha_{\mathrm{i}}^{*}\leq C,i=1,...,l$$

$$\sum_{i=1}^{l}(\alpha_{\mathrm{i}}-\alpha_{\mathrm{i}}^{*})=0 \tag{3.32}$$

解（3.32）约束的（3.30）式得到拉格朗日乘数α，α'，回归函数由式（3.27）表示，其中

$$\bar{\omega}=\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}}-\alpha_{\mathrm{i}}^{*})x_{\mathrm{i}}$$

$$\bar{b}=-\frac{1}{2}[\bar{\omega}.(x_{\mathrm{r}}+x_{\mathrm{s}})] \tag{3.33}$$

解满足的Karush-Kuhn-Tucker（KKT）条件如下：

$$\bar{\alpha}_{\mathrm{i}},\bar{\alpha}_{i}^{*}=0,i=1,...,l \tag{3.34}$$

这样，支持矢量就是那些大于0的拉格朗日乘数对应的样本点。ε =0时，我们使用L1损失函数，优化问题就可以简化：

$$\min_{\beta}\frac{1}{2}\sum_{\mathrm{i}=1}^{l}\sum_{\mathrm{j}=1}^{l}\beta_{\mathrm{i}}\beta_{\mathrm{j}}(x_{\mathrm{i}},x_{\mathrm{j}})-\sum_{\mathrm{i}=1}^{l}\beta_{\mathrm{i}}y_{\mathrm{i}} \tag{3.35}$$

限制为，

$$-C\leq\beta_{\mathrm{i}}\leq C,i=1,...,l$$

$$\sum_{i=1}^{l}\beta_{i}=0 \tag{3.36}$$

回归函数仍为式（3.27）所示，其中：

$$\bar{\omega}=\sum_{\mathrm{i}=1}^{l}\bar{\beta}_{\mathrm{i}}x_{\mathrm{i}}$$

$$\bar{b}=-\frac{1}{2}[\bar{\omega}.(x_{\mathrm{r}}+x_{\mathrm{s}})] \tag{3.37}$$

（2）二次损失函数

使用图3.9(a)所示的二次损失函数：

$$L_{quad}(f(x)-y)=(f(x)-y)^2 \tag{3.38}$$

这时候解由下式给出：

$$\max_{\alpha,\alpha^*} W(\alpha,\alpha^*)=\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{\mathrm{i}=1}^{l}\sum_{\mathrm{j}=1}^{l}(\alpha_{\mathrm{i}}-\alpha_{\mathrm{i}}^*)(\alpha_{\mathrm{j}}-\alpha_{\mathrm{j}}^*)(x_{\mathrm{i}}.x_{\mathrm{j}})$$

$$+\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}}-\alpha_{\mathrm{i}}^*)y_{\mathrm{i}}-\frac{1}{2C}\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}}^2+(\alpha_{\mathrm{i}}^*)^2) \tag{3.39}$$

对应的优化问题可以使用KKT条件而得到简化，注意这意味着$\beta i'=|\beta i|$，这样优化问题变成：

$$\min_{\beta}\frac{1}{2}\sum_{\mathrm{i}=1}^{l}\sum_{\mathrm{j}=1}^{l}\beta_{\mathrm{i}}\beta_{\mathrm{j}}(x_{\mathrm{i}}.x_{\mathrm{j}})-\sum_{\mathrm{i}=1}^{l}\beta_{\mathrm{i}}y_{\mathrm{i}}+\frac{1}{2C}\sum_{\mathrm{i}=1}^{l}\beta_{\mathrm{i}}^2 \tag{3.40}$$

约束为：

$$\sum_{\mathrm{i}=1}^{l}\beta_{\mathrm{i}}=0 \tag{3.41}$$

回归函数由式（3.27）和式（3.37）给出。

（3）Huber损失函数

使用图3.9(c)所示的Huber损失函数：

$$L_{huber}(f(x)-y)=\begin{cases}\frac{1}{2}(f(x)-y)^2, & for \quad |f(x)-y|<\mu\\ \mu|f(x)-y|-\frac{\mu^2}{2}, & otherwise\end{cases} \tag{3.42}$$

解由下式给出：

$$\max_{\alpha,\alpha^*} W(\alpha,\alpha^*)=\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{\mathrm{i}=1}^{l}\sum_{\mathrm{j}=1}^{l}(\alpha_{\mathrm{i}}-\alpha_{\mathrm{i}}^*)(\alpha_{\mathrm{j}}-\alpha_{\mathrm{j}}^*)(x_{\mathrm{i}}.x_{\mathrm{j}})$$

$$+\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}}-\alpha_{\mathrm{i}}^*)y_{\mathrm{i}}-\frac{1}{2C}\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}}^2+(\alpha_{\mathrm{i}}^*)^2)\mu \tag{3.43}$$

最后优化问题如下：

$$\min_{\beta}\frac{1}{2}\sum_{\mathrm{i}=1}^{l}\sum_{\mathrm{j}=1}^{l}\beta_{\mathrm{i}}\beta_{\mathrm{j}}(x_{\mathrm{i}}.x_{\mathrm{j}})-\sum_{\mathrm{i}=1}^{l}\beta_{\mathrm{i}}y_{\mathrm{i}}+\frac{1}{2C}\sum_{\mathrm{i}=1}^{l}\beta_{\mathrm{i}}^2\mu \tag{3.44}$$

约束为：

$$-C \leq \beta_{\mathrm{i}} \leq C, i = 1, ..., l$$
$$\sum_{\mathrm{i}=1}^{l} \beta_{\mathrm{i}} = 0 \tag{3.45}$$

回归函数由式3.27和式3.37给出。

（二）非线性回归

与分类问题类似，非线性模式一般需要适当的模型数据。与支持矢量分类器的方法相似，用一个非线性映射，把数据映射到一个高维特征空间，在这个空间里使用线性的方法解决。核函数同样被引入，用来解决过高的维数带来的麻烦。使用图3.9(c)描述的$\varepsilon-$迟钝损失函数，非线性回归问题的解如下给出：

$$\max_{\alpha,\alpha^*} W(\alpha, \alpha^*) = \max_{\alpha,\alpha^*} \sum_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}^*(y_{\mathrm{i}} - \varepsilon) - \alpha_{\mathrm{i}}(y_{\mathrm{i}} + \varepsilon)$$
$$-\frac{1}{2}\sum_{\mathrm{i}=1}^{l}\sum_{\mathrm{j}=1}^{l}(\alpha_{\mathrm{i}}^* - \alpha_{\mathrm{i}})(\alpha_{\mathrm{j}}^* - \alpha_{\mathrm{j}})K(x_{\mathrm{i}}.x_{\mathrm{j}}) \tag{3.46}$$

约束如下：

$$-C \leq \alpha_{\mathrm{i}}, \alpha_{\mathrm{i}}^* \leq C, i = 1, ..., l$$
$$\sum_{\mathrm{i}=1}^{l}(\alpha_{\mathrm{i}} - \alpha_{\mathrm{i}}^*) = 0 \tag{3.47}$$

求解（3.47）约束的（3.46）式得到拉格朗日乘数α，α'，回归函数由下式给出：

$$f(x) = \sum_{\mathrm{SVMs}} (\bar{\alpha}_{\mathrm{i}} - \bar{\alpha}_{\mathrm{i}}^*)K(x_{\mathrm{i}}, x) + \bar{b} \tag{3.48}$$

其中：

$$\bar{\omega}.x = \sum_{\mathrm{SVMs}} (\bar{\alpha}_{\mathrm{i}} - \bar{\alpha_{\mathrm{i}}}^*)K(x_{\mathrm{i}}, x)$$
$$\bar{b} = -\frac{1}{2}\sum_{\mathrm{SVMs}} (\bar{\alpha}_{\mathrm{i}} - \bar{\alpha_{\mathrm{i}}}^*)[K(x_{\mathrm{i}}, x_{\mathrm{r}}) + K(x_{\mathrm{i}}, x_{\mathrm{s}})] \tag{3.49}$$

同样，如果核包含偏差b，等式约束就可以去掉。这样回归函数为：

$$f(x) = \sum_{\mathrm{SVMs}} (\bar{\alpha}_{\mathrm{i}} - \bar{\alpha_{\mathrm{i}}}^*)K(x_{\mathrm{i}}, x) \tag{3.50}$$

其他的损失函数其优化过程都是类似的，只要用核函数替换高维空间的点积就行了。ε-迟钝损失函数是很吸引人的，它和二次损失函数以及Huber损失函数不一样，后两者的所有样本点都将是支持矢量，而它的支持矢量将是稀疏的。二次损失函数产生的解和脊形回归（ridge regression)的解是等价的。当正则化参数$\lambda = \frac{1}{2C}$时，零级正则化（zeroth order regularisation）的解也和它等价。

3.3.2　SVM核函数

核函数的引入是为了建立一个到高维特征空间的映射，其基本思想是使各种操作在输入空间中进行，而非在高维特征空间中进行，因此，内积并不需要在特征空间评估。但计算仍严格地依赖于训练样本的种类数。对于高维问题，要想获得好的数据分布通常要求足够大的训练样本。

下面的理论是建立重构带核的希尔伯特空间。特征空间的内积具有与输入空间平等的核函数，

$$K(x, x^{'}) =< \phi(x), \phi(x^{'}) >, \tag{3.51}$$

如果某种条件成立，假设K是一个正的对称的确定函数，且满足Mercer条件：

$$K(x, x^{'}) = \sum_{\mathrm{m}}^{\infty} \alpha_{\mathrm{m}} \phi_{\mathrm{m}}(x) \phi_{\mathrm{m}}(x^{'}), \qquad a_{\mathrm{m}} \geq 0, \tag{3.52}$$

$$\int\int K(x, x^{'}) g(x) g(x^{'}) dx dx^{'} > 0, \qquad g \in L_2 \tag{3.53}$$

那么该核函数代表特征空间的合理内积。除非特别声明，满足Mercer条件的有效的核函数对任何x和$x^{'}$都成立。

核函数有多种形式，这也是支持矢量机的研究热点之一。

（1）多项式（Polynomial）

多项式映射对非线性模型是一种比较流行的方法，

$$K(x, x^{'}) =< x, x^{'} >^{d}, \tag{3.54}$$

$$K(x, x^{'}) = (< x, x^{'} > +1)^{d}, \tag{3.55}$$

为避免赫赛函数(hessian)为零，通常采用第二种。

（2）高斯径向基函数（Gaussian Radial Basis Function）

径向基函数备受关注，具有如下的高斯形式

$$K(x,x^{'})=exp(-\frac{\|x-x^{'}\|^2}{2\sigma^2}) \tag{3.56}$$

利用径向基函数的古典技巧使用了某种确定的子族中心。典型的是一种聚类方法首先用于选择子族中心。支持矢量机的突出特征是这种选择不明显，用每一个支持矢量构造一个集中于某点的局部高斯函数。

（3）指数径向基函数（Exponential Radial BasisFunction）

径向基函数具有如下的形式：

$$K(x,x^{'})=exp(-\frac{\|x-x^{'}\|}{2\sigma^2}) \tag{3.57}$$

当非连续存在时，可以得到分段的线性解。

（4）多层感知机（Multi-Layer Perceptron）

具有单一个隐层的多层感知机有一个有效的核函数表达式：

$$K(x,x^{'})=tanh(\rho<x,x^{'}>+\tau) \tag{3.58}$$

其中ρ为尺度因子，τ为偏差因子，支持矢量对应第一层，拉格朗日对应权重。

（5）傅立叶级数（Fourier Series）

可以认为傅立叶级数在2N+1维特征空间中展开。核函数定义在$[-\frac{\pi}{2},\frac{\pi}{2}]$区间上，

$$K(x,x^{'})=\frac{sin(N+\frac{1}{2})(x-x^{'})}{sin(\frac{1}{2}(x-x^{'}))} \tag{3.59}$$

有傅立叶转化可知，该核函数的正则能力很差，因此其不是最好的选择。

（6）样条函数（Splines）

由于样条函数的灵活性，它们常被采用。一个有限的κ级样条函数在τ_s处有N项：

$$K(x,x^{'})=\sum_{\mathrm{r}=0}^{\kappa}x^{r}x^{'r}+\sum_{\mathrm{s}=1}^{N}(x-\tau_{\mathrm{s}})_{+}^{\kappa}(x^{'}-\tau_{\mathrm{s}})_{+}^{\kappa} \tag{3.60}$$

定义在区间[0，1）上的无限的样条函数采取如下形式：

$$K(x,x^{'})=\sum_{\mathrm{r}=0}^{\kappa}x^{r}x^{'r}+\int_{0}^{1}(x-\tau_{\mathrm{s}})_{+}^{\kappa}(x^{'}-\tau_{\mathrm{s}})_{+}^{\kappa}d\tau \tag{3.61}$$

当$\kappa = 1$时，核函数变为

$$K(x,x^{'}) = 1+ < x,x^{'} > + \frac{1}{2} < x,x^{'} > min(x,x^{'}) - \frac{1}{6} min(x,x^{'})^3 \qquad (3.62)$$

其解为分段的三次解。

（7）B样条函数（B splines）

B样条函数是另外一种普遍的样条函数。核函数定义在[-1，1]，具体如下：

$$K(x,x^{'}) = B_{2N+1}(x - x^{'}) \qquad (3.63)$$

（8）叠加的核函数（Addictive Kernels）

较为复杂的核函数可以通过核函数的叠加获得

$$K(x,x^{'}) = \sum_{\mathrm{i}} K_{\mathrm{i}}(x,x^{'}) \qquad (3.64)$$

（9）张量积（Tensor Product）

多维核函数可由核函数的张量积得到

$$K(x,x^{'}) = \prod K_{\mathrm{i}}(x_i,x_i^{'}) \qquad (3.65)$$

这对建立多维样条核函数尤为重要，其可直接由单变量的核函数内积得到。

面对诸多映射，究竟选择哪一个比较合适？研究者通常以步步为营法（bootstrapping）和交叉确认法（cross-validation）选择核函数。

3.3.3　SVMs在测光红移中的应用

（1）实验一：

选择了SDSS DR5中所有已知光谱红移的星系与2MASS展源星表进行交叉证认，得到了150,000 个星系。在这些星系样本中，我们又进行了如下的限制：（1）SDSS光谱红移置信度大于等于0.95；（2）红移警告为0；（3）SDSS r星等小于等于17.5mag。经过上述三个条件的约束，我们得到了含有62,083个星系的样本集。

用训练集方法预测红移的好处在于可以添加额外的参数，这些参数可能使预测精度提高。为了研究这些参数（例如：petroR50 、petroR90、fracDeV_r）在红移预测过程中是否起到提高预测精度的作用，我们设计了不同的输入参数组合，结果如表3.4所示。

表 3.4: 不同输入参数用SVMs方法预测红移的剩余标准偏差$\sigma_{\rm rms}$

输入参数	$\sigma_{\rm rms}$
u, g, r, i, z	0.0291
u, g, r, i, z, J, H, Ks	0.0278
$u-g, g-r, r-i, i-z$	0.0273
$u-g, g-r, r-i, i-z, r$	0.0284
$u-g, g-r, r-i, i-z, z-J, J-H, H-Ks$	0.0273
$u-g, g-r, r-i, i-z, z-J, J-H, H-Ks, r$	0.0275
$u-g, g-r, r-i, i-z$,fracDev_r	0.0306
$u-g, g-r, r-i, i-z$,petroR50,petroR90	0.0330

注：—-petroR50是r波段50%的光度半径；petroR90 是r波段90%的光度半径；fracDeV_r是r波段的fracDeV。

从表3.4的实验结果我们可以看出，在输入参数为色指数的时候，SVMs 可以得到最佳的预测结果，也就是说，当输入4个色指数（$u-g$、$g-r$、$r-i$、$i-z$）或者7个色指数（$u-g$、$g-r$、$r-i$、$i-z$、$z-j$、$j-h$、$h-k$）时，预测精度都达到了最优值$\sigma_{\rm rms}$=0.0273。而7个色指数加上SDSS r星等的预测精度要高于8个星等，更优于4个色指数加上r星等作为输入的组合。当4个色指数加上fracDev_r或者petroR50、petroR90时预测精度都明显下降。SVMs并不象ANN算法，输入的参数越多预测精度越高。对于SVMs来说，只有与色指数直接相关的参数才能提高红移的预测精度。

图3.10显示了用SVM算法预测的测光红移与SDSS光谱红移的对比散点图，输入参数为4个色指数（$u-g$、$g-r$、$r-i$、$i-z$）。 从图中明显可见偏差。虽然SVMs不需要象ANNs那样需要训练网络，也避免了陷入局部极小和过度拟合的困境。但是SVMs需要先验的经验调整参数，SVMs提供了不同的核函数。如果参数调整有效，即使最简单的高斯核函数，也可以取得理想的结果。SVMs与ANNs一样，当训练样本集发生变化的时候，需要重新训练获得回归器才能进行预测。

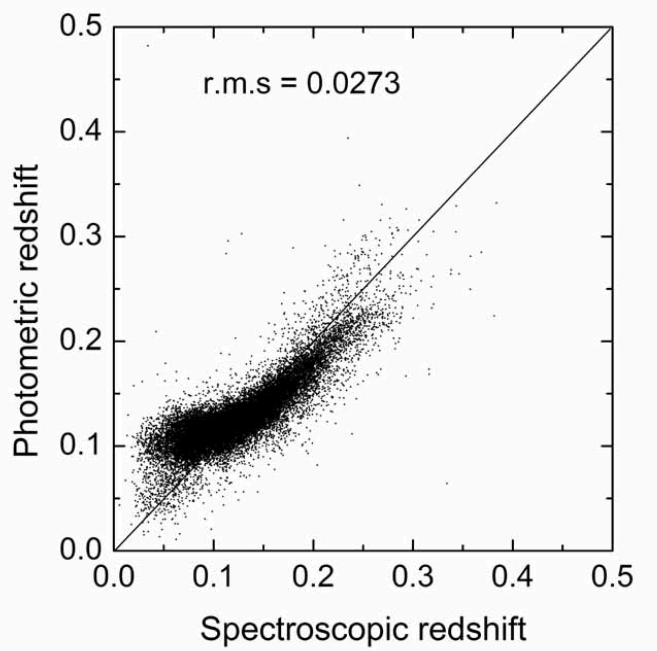


图 3.10: 光谱红移和用SVMs预测的测光红移的对比散点图，数据来源于SDSS DR5 与2MASS交叉证认的样本集。

（2）实验二

我们选择了SDSS DR5中所有具有光谱红移的类星体样本，且此样本需要满足三个条件：(a) SpecClass = 3（证明此光谱是类星体）；(b)SpecZWarining = 0；(c) SpecZStatus > 2。基于这些条件获得67,492类星体样本。我们用SVMs方法预测了67,492样本的测光红移，其预测精度如表3.5所示。 在表中我们采用不同的衡量标准。如何在6种衡量标准中找到最优的模型参数，是问题的关键所在。我们考虑选择绝对误差（$|\Delta z| < 0.3$）所占比例最大的模型参数组合作为最优回归器。当选用高斯核SVMs，模型的可调参数有两个g和c，采用固定g，调c，而后随着在$|\Delta z| < 0.3$区域的预测百分比增大到一定程度时，再固定c，调g。从表3.5可以看出g在取小点的值时，增大c值有助于提高处于$|\Delta z| < 0.3$区域的预测百分比；当g调到20，而后调c，可以发现c增到2后，若再增加，百分比则会降低；若固定c，继续增加g，百分比也会降低；当$g = 20$，$c = 2$时，SVMs预测类星体测光红移的精度最优，分别为当$|\Delta z| < 0.1$时占48.94%，当$|\Delta z| < 0.2$时占70.71%，当$|\Delta z| < 0.3$时占78.12%，当$|\frac{\Delta z}{1+z}| < 0.25$占86.59%，方差为0.119。当然理想的标准应该是处于三个区域（$|\Delta z| < 0.1$、$|\Delta z| < 0.2$、$|\Delta z| < 0.3$）和满

表 3.5: 高斯核SVMs在不同的输入参数下预测类星体测光红移的预测精度

输入参数	$\|\Delta z\| < 0.1$	$\|\Delta z\| < 0.2$	$\|\Delta z\| < 0.3$	$\|\frac{\Delta z}{1+z}\| < 0.25$	方差
g=0.1 c=0.01	27.89%	46.75%	56.83%	69.4%	0.162
g=0.1 c=0.1	32.36%	51.89%	62.18%	72.67%	0.153
g=0.1 c=1	32.81%	53.66%	63.6%	74.6%	0.132
g=0.1 c=10	32.84%	53.31%	64.08%	75.6%	0.124
g=5 c=0.01	37.71%	57.99%	67.1%	78.22%	0.160
g=5 c=0.1	41.02%	63.83%	72.95%	83.59%	0.113
g=5 c=1	44.46%	67.24%	75.59%	85.34%	0.100
g=5 c=10	46.19%	68.82%	76.73%	86.15%	0.101
g=5 c=100	42.61%	67.48%	76.8%	86.08%	0.109
g=10 c=0.1	44.09%	65.84%	74.06%	84.23%	0.123
g=10 c=1	47.34%	69.84%	77.29%	86.29%	0.107
g=15 c=1	48.7%	70.38%	77.91%	86.39%	0.114
g=20 c=0.1	46.34%	67.4%	74.6%	83.76%	0.154
g=20 c=1	49.17%	70.59%	77.86%	86.32%	0.120
g=20 c=2	**48.94%**	**70.71%**	**78.12%**	**86.59%**	**0.119**
g=20 c=5	48.79%	70.28%	77.89%	86.53%	0.122
g=20 c=10	48.28%	70.05%	77.61%	86.2%	0.125
g=25 c=1	49.1%	70.47%	77.73%	86.24%	0.126
g=50 c=1	46.93%	68.32%	75.85%	83.99%	0.155

足$|\frac{\Delta z}{1+z}| < 0.25$的预测源的百分比越高，同时方差最小时，这时的模型最优，而实际应用中很难同时满足这些条件。天文学家可以根据自己课题的需要来选取适当的标准。

3.4 核回归

3.4.1 原理

设Y为因变量，X是自变量。X为d维解释变量向量，是影响Y的若干重要因素，既可为确定性变量，也可为随机性变量。给定样本观测值（X_1, Y_1），（X_2,

Y_2），...，（X_n, Y_n），假定$\{Y_\mathrm{i}\}$独立同分布，便可建立多元非参数回归模型：

$$Y_i = m(X_\mathrm{i}) + u_\mathrm{i}, \qquad i = 1, ..., n \tag{3.66}$$

其中$m(.)$是未知的函数，u_i为随机误差项，它反映了除解释变量外其他影响被解释变量的可观察或不可观察的因素对被解释变量的影响以及模型的设定误差等。当解释变量为确定性变量时，假定随机误差项的数学期望为零，即Eu_i=0。此时，被解释变量的数学期望E$Y_\mathrm{i} = m(X_\mathrm{i})$。当解释变量为随机变量时，假定解释变量与随机变量独立，假定随机误差项的条件数学期望为零，即E$(u_\mathrm{i}|X_\mathrm{i}) = 0$。此时，被解释变量的条件数学期望E$(Y_\mathrm{i}|X_\mathrm{i}) = m(X_\mathrm{i})$。当Var$(u_\mathrm{i}|X_\mathrm{i}) = \sigma_\mathrm{u}^2$时，称随机误差项为同方差，否则，称随机误差项为异方差。当影响被解释变量的其他因素随着解释变量的水平的变化，对被解释变量的影响程度也随之变化时，就产生了异方差现象。对于模型（3.66）的估计方法有许多种，例如：核估计、局部线形估计、近邻估计、正交序列估计和样条估计等。在本论文中，我们用的是核估计，即核回归（Kernel Regression，简称KR）[66]。

多元核回归模型（3.66）的不变窗宽核回归为

$$\hat{m}_\mathrm{n}(x, h_\mathrm{n}) = \frac{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}} K_{h_\mathrm{n}}(X_\mathrm{i} - x) Y_\mathrm{i}}{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}} K_{h_\mathrm{n}}(X_\mathrm{i} - x)} \tag{3.67}$$

其中h_n为窗宽，$K_{h_\mathrm{n}}(u) = h_\mathrm{n}^{-d} K(h_\mathrm{n}^{-1} u)$，$K(.)$是$d$维对称密度函数，$K(u) \geq 0$，$\int K(u)du = 1$。最常见的核函数有：均匀核$K(u) = 0.5$、高斯核$K(u) = \frac{1}{\sqrt{(2\pi)}} exp(-\frac{1}{2}u^2)$和抛物线核$K(u) = 0.75(1 - u^2)$。我们采用了高斯核，因此式（3.67）可以转化为：

$$\hat{m}_\mathrm{n}(x, h_\mathrm{n}) = \frac{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}} exp[-\frac{1}{2}(\frac{X_\mathrm{i} - x}{h_\mathrm{n}})^2] Y_\mathrm{i}}{\sum\limits_{\mathrm{i}=1}^{n} exp[-\frac{1}{2}(\frac{X_\mathrm{i} - x}{h_\mathrm{n}})^2]} \tag{3.68}$$

$K(.)$是概率密度函数，核回归就是Y_i的加权算术平均值。当X_i落在离x越近时，权数就越大；落在离x越远时，权数就越小；当X_i落在$[x - 3h_\mathrm{n}, x + 3h_\mathrm{n}]$之外时，权数基本上为零。

3.4.2 窗宽的选择

窗宽是控制核回归精度的重要参数。从公式（3.68）可以看出，太小的窗宽得到除了数据点外其他点的函数值都为零的函数。所以，太小的窗宽会使得随机误差项产生的噪音没有被排出，是没有意义的估计。而太大的窗宽得到过份光滑的曲线，接近于直线，此时的估计也是没有任何意义的。在核估计的实际应用中，如果回归函数的估计接近于一条直线，则窗宽肯定过大，参加局部加权的观测点过多，此时，可减少窗宽。如果回归函数的估计很不光滑，则窗宽肯定过小，此时，随机误差项产生的噪音没有被排除，应该加大窗宽，使得在局部参加加权平均的观测点增多，从而更多的消除随机误差项产生的噪声。最佳的窗宽应当是既不过小也不过大。

（一）理论窗宽的最佳选择

回归函数估计的渐进方差随着窗宽减小而增大，渐进偏随着窗宽的减小而减小。所以，核回归是在估计的渐进偏和渐进方差中寻求平衡，使得均方误差达到最小。使得均方误差达到最小的最佳理论窗宽具有以下形式：

$$h_n = cn^{-\frac{1}{4+d}} \tag{3.69}$$

其中c与n无关，只与回归函数、解释变量的密度函数和核函数有关。d是维数。所以，在实际应用中最佳窗宽的选择是不断地调整c，使得核回归达到满意的估计结果。

（二）窗宽的交错鉴定选择方法

交错鉴定方法（Cross-Validation，简称CV）是选择窗宽的一个常用方法，分为留一法（leave one out）和$k-$折交叉确认（k-fold validation）两种。

留一法的原理为：每个局部观察点$x = X_{\mathrm{i}}$，首先，在样本中提出该观察点（$X_{\mathrm{i}}, Y_{\mathrm{i}}$），其次，将剩下的$n-1$个观察点作为训练样本进行核回归，最后，通过比较平方拟合误差

$$CV(h_{\mathrm{n}}) = n^{-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(Y - \hat{Y})^2 w(x_{\mathrm{i}}) \tag{3.70}$$

选择使平方拟合误差达到最小的窗宽h_{n}。其中$w(x_i) \geq 0$为某权数，Y代表光谱红移，$\hat{Y}$代表用核回归方法得到的测光红移，n代表样本总数。该方法的关键是在样本中突出观察点（$X_{\mathrm{i}}, Y_{\mathrm{i}}$）。如果不这样的话，由于核权函数$W_{\mathrm{ni}}(x)$在观察

点$x = X_{\mathrm{i}}$ 处达到最大值，就会使得$x = X_{\mathrm{i}}$的重要当对解释变量数据的分布掌握一定信息，尤其对非均匀分布时，此时若采用不变窗宽估计，则当不变窗宽太小时，在解释变量密度小的x处，因参加平均的观测点少，故核估计的偏差较大。而当不变窗宽过大时，虽然提高了解释变量密度小处核估计的精度，但在解释变量密度大的x处，因参加平均的观测点过多，造成和估计的偏差增大。在现实中，变量的分布是均匀分布的情况较少，更多的是非均匀分布的情况。对于解释变量程度过分夸大其他观察点数据的重要程度降低。所以采用交错鉴定方法就避免了因没剔除观察点（$X_{\mathrm{i}}, Y_{\mathrm{i}}$）而将有用的数据排除在外的情况。

$k-$折交叉确认法原理为：将训练样本集随机地分成k个互不相交的子集，每个折的大小大致相等。利用$k-1$个作为训练样本，对给定的一组参数建立回归模型，利用剩下的一个子集做测试样本。根据以上过程重复k次，因此每个子集都有机会进行测试，根据k次迭代后得到的CV值，参考公式（3.71）。对同一组参数建立的回归模型来说，CV值最小时对应的窗宽就是该组参数建立回归模型中的最佳窗宽。

$$CV(h_n) = \frac{1}{K}\sum_{K=1}^{K}\frac{1}{m}\sum_{i=1}^{m}(Y-\hat{Y})^2 w(x_i) \tag{3.71}$$

其中，m代表每折中的样本数，通常情况下m是相等的；$w(x_{\mathrm{i}}) \geq 0$为某权数；Y代表光谱红移；$\hat{Y}$代表用核回归方法得到的测光红移。本文中我们采用了10-折交叉确认法，$w(x_{\mathrm{i}})$均取1。

3.4.3　赤池信息准则和贝叶斯信息准则

赤池信息准则（Akaike information criterion，简称AIC）是由日本统计学家赤池弘次在1974年根据极大似然估计原理提出的一种较为普遍的的模型选择准则，它既可用于回归变量选择中，又可用于时间序列分析的自回归模型的定阶上。AIC值越小，统计模型越优。显然，AIC准则使得模型的评价和选择工作变得简单实用。AIC定义如下：

$$AIC = -2\ln L_{\max} + 2k \tag{3.72}$$

其中$L_{\max}$是最大似然函数，k是模型中自由参数的个数。假设模型误差是正态分布的，AIC也可以写如下形式：

$$AIC = 2\sigma^2 k + RSS \tag{3.73}$$

其中，RSS是剩余方差平方和。

施瓦茨准则（Schwarz Information criterion，简称SIC）是由Schwarz在1978年提出的以贝叶斯原理为基础，因此又叫做贝叶斯信息准则（Bayesian Information Criterion，简称BIC）。与赤池信息准则一样，也可用于模型的选择，值也是越小越好。BIC克服了AIC自我回归高估的情况。通常，AIC适合小样本，而BIC适合大样本情况。BIC适用于参数个数相等的模型选择。BIC定义如下：

$$BIC = -2\ln L_{\max} + k\ln N \tag{3.74}$$

其中$L_{\max}$是最大似然函数，k是模型中自由参数的个数，N为样本数。

通过比较AIC和BIC的性质可知，BIC惩罚复杂度性能优于AIC。对于无限数目样本，BIC要优越。当模型维数有限时，BIC的结果要好于AIC。BIC最大化模型后验概率，因此，如果真正模型是候选模型之一，BIC最可能选出真正模型，如果真实模型不在候选模型中，BIC趋向选取简单模型。相反，AIC则是最小化模型和真实分布之间的不同。如果真实模型不是候选模型，AIC将选取具有最小均方差的模型。通常情况下，样本数目是有限的，真实模型也难以构建，因此AIC更具有优越性，被广泛地使用。

3.4.4 KR在测光红移中的应用

（1）实验一：

我们选择了SDSS DR5中所有有光谱红移的星系与2MASS展源星表进行交叉证认，得到了150,000 个星系。从这些星系样本中，我们又做了如下的限制：（a）SDSS光谱红移置信度（zConfidence）大于等于0.95；（b）红移警告（zWarning）为0；（c）SDSS r星等小于等于17.5mag；经过上述三个条件的约束，我们得到了一个含有62,083个星系的样本集。

为了测试多波段测光数据对核回归预测红移精度的作用。我们进行了多输入参数组合的实验，结果如表3.6中所示。表3.6显示出不同输入参数的剩余标准偏差$\sigma_{\rm rms}$以及对应的最优窗宽，此处的最优窗宽是通过10-折交错确认方法得到的。

从表3.6可以看出，当SDSS四个色指数作为参数时，剩余标准偏差$\sigma_{\rm rms}$达到了0.0193。当添加了额外的参数fracDev_r时，剩余标准偏差$\sigma_{\rm rms}$有了细微的变化，变成了0.0192，这说明fracDev_r在预测红移的时候起到了积极的作用。在上

表 3.6: 不同输入参数的$\sigma_{\rm rms}$值（样本为SDSS和2MASS交叉证认得到的星系）

输入参数	剩余标准偏差$\sigma_{\rm rms}$	最佳窗宽
u,g,r,i,z	0.0208	$h=0.025$
u,g,r,i,z,j,h,k	0.0254	$h=0.015$
$u-g,g-r,r-i,i-z$	0.0193	$h=0.020$
$u-g,g-r,r-i,i-z,r$	0.0196	$h=0.025$
$u-g,g-r,r-i,i-z,z-j,j-h,h-k$	0.0210	$h=0.045$
$u-g,g-r,r-i,i-z,z-j,j-h,h-k,r$	0.0235	$h=0.055$
$u-g,g-r,r-i,i-z$, fracDev_r	0.0192	$h=0.020$
$u-g,g-r,r-i,i-z$,petroR50,petroR90	0.0218	$h=0.040$

注：— petroR50是r波段50%的光度半径；petroR90 是r波段90%的光度半径；fracDeV_r是r波段的fracDeV。

述两种输入参数下，最优窗宽（h）均是0.02。当加r星等时，预测精度则下降，剩余标准偏差$\sigma_{\rm rms}$变成了0.0196，这说明在这批样本中，增加参数r星等对红移的预测意义不大。当输入参数为SDSS的五个星等或者七个色指数时，预测精度明显变差，剩余标准偏差$\sigma_{\rm rms}$分别为0.0208、0.0210。2MASS的三个星等在非参数回归中作用甚微。在本实验中，最差的输入参数是八个星等，剩余标准偏差$\sigma_{\rm rms}$=0.0254。此时预测已经出现了很大的偏差。此外，petroR50、petroR90参数也没有起到提高精度的作用。从以上实验结果可以看出，对于核回归来说，并非参数的个数越多越好，只有增加那些与星系类型有关的参数才能提高预测精度，否则反而会使预测精度恶化。图3.11显示了用核回归的方法预测的测光红移与SDSS光谱红移的对比散点图，输入参数为SDSS四个色指数的结果。核回归不同于其他方法的地方在于：当窗宽（h）很小时，测试样本在此窗宽内找不到训练样本时导致预测失败，我们称之为损失点。损失点随着窗宽的增大而骤减。在该实验中，不同输入参数最优窗宽下的损失率均小于1%，这么大的损失率对大样本的统一研究是可以接受的。

（2）实验二：

挑选在SDSS DR5所有具有光谱红移的星系中光谱红移置信度大于等于0.95，并且红移警告等于0的样本，满足该条件的星系样本数为399,929。

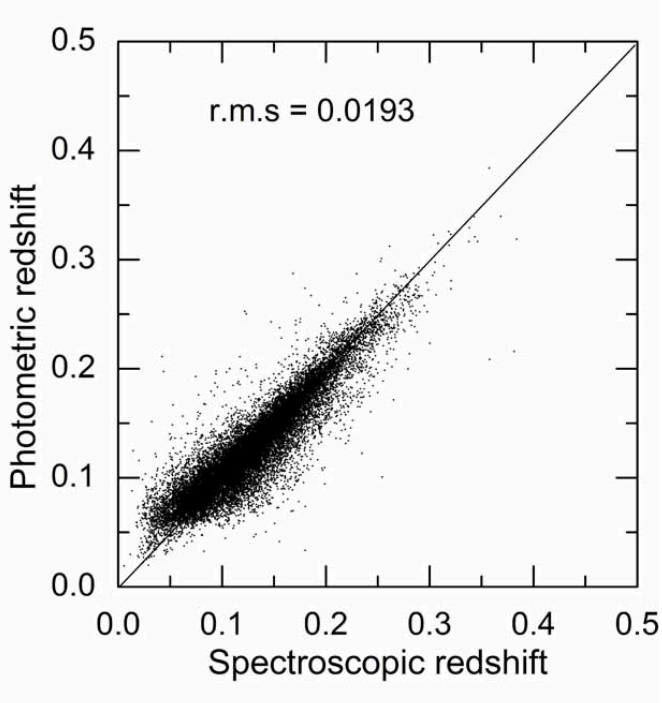


图 3.11: SDSS光谱红移与用核回归方法得到的测光红移的对比散点图。样本集是62,083个SDSS和2MASS交叉证认得到的星系样本。

在这个实验中我们探究了星系的光谱类型eClass、汇聚指数（Concentration index，简写c）对预测测光红移的作用，以及AIC、BIC标准来评判输入参数模型的优劣。

在以往的实验中，我们按经验方法将样本集按2：1的比例分成了训练样本和测试样本。为了测试这种经验方法的有效性，我们做了如下实验。实验结果如图3.12所示 从图中我们可以看到，当训练样本数目为200,000~250,000，而测试样本数目为150,000~200,000时，剩余标准偏差$\sigma_{\rm rms}$同时达到最小值0.0206。即在训练样本与测试样本比例为1：1或者2：1时，预测都可以达到最优。这充分证明使用260,000个星系作为训练样本，139,929个星系样本作为测试样本的可靠性。

在窗宽的选择中，我们使用了10-折交错鉴定、赤池信息准则和贝叶斯信息准则。以SDSS四色指数（$u-g$、$g-r$、$r-i$、$i-z$）为输入参数。实验结果如表3.7所示。 从表中我们可以发现：当$h=0.02$时，CV、AIC和BIC同时达到了最小值，因此对于四个色指数为输入参数时，最优窗宽为0.02，此时剩余标准偏差$\sigma_{\rm rms}$=0.0220也最小。有一点需要特别指出的是CV、AIC、BIC和剩余标

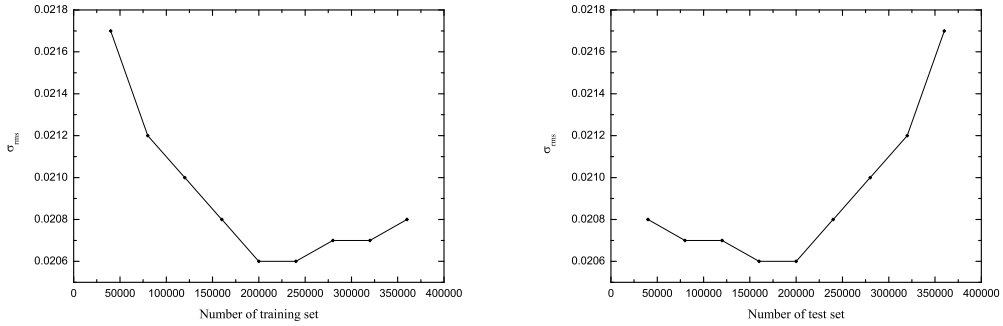


图 3.12: 左图为训练样本数目与剩余标准偏差$\sigma_{\rm rms}$的关系图；右图为测试样本数目与剩余标准偏差$\sigma_{\rm rms}$的关系图.

准偏差$\sigma_{\rm rms}$的变化趋势不是始终一致的。也就是说CV最小的时候，剩余标准偏差$\sigma_{\rm rms}$未必最小，因为选择最优窗宽的原则是既要考虑预测的精度，同时又要兼顾损失点不能太多。在我们的事例中是一种巧合。通常为避免多种规则的干扰，可以只用交叉鉴定方法来选择最优窗宽。

我们用了不同的参数作为输入组合，实验结果如表3.8所示 在该实验中，样本没有对r星等做限制，因此样本集中包含了一些大红移的样本。从表3.8中可见，输入参数为color+r+eClass时，预测精度很高，剩余标准偏差$\sigma_{\rm rms}$达到0.0189，此时的最优窗宽为0.025。$ugriz$+eClass作为输入参数时，剩余标准偏差$\sigma_{\rm rms}$为0.0198，最佳窗宽为0.025，预测精度略低于color+r+eClass的情况。而当输入为色指数、星等（例如color、color+r、color+$r+c$以及$ugriz$时），预测的弥散度明显要大于eClass作为输入参数的组合。参数fracDeV_r没有明显改善预测精度，color+fracDeV_r预测结果与color的相同。参数petroR50、petroR90对红移的预测精度非但没有提高，反而使预测的弥散度更大。这更证明了我们在实验一中得到的结论，只有那些与星系类型相关的参数（例如：eClass）才能有效地提高红移的预测精度，对核回归来说并不是参数越多越好，这点与人工神经网络完全不同。

图3.13显示了输入参数为4个色指数和r星等（$u-g$、$g-r$、$r-i$、$i-z$、r）时，用核回归方法预测的测光红移与SDSS光谱红移的对比散点图，其中训练样本为260,000，测试样本为139,929，剩余标准偏差$\sigma_{\rm rms}$=0.0206。

图3.14显示了输入参数为color+r+eClass时，核回归预测的测光红移与SDSS光

表 3.7: 对应于交错鉴定(CV)、赤池信息准则(AIC)和贝叶斯信息准则(BIC)的窗宽和剩余标准偏差$\sigma_{\rm rms}$

h	CV	AIC	BIC	剩余标准偏差$\sigma_{\rm rms}$
0.010	22.669	69.900	79.750	0.0225
0.015	22.292	67.644	77.494	0.0221
0.020	**22.231**	**67.555**	**77.405**	**0.0220**
0.025	22.475	68.458	78.308	0.0222
0.030	22.894	69.747	79.597	0.0224
0.035	23.318	71.262	81.112	0.0226
0.040	23.889	73.069	82.919	0.0229
0.045	24.442	75.111	84.961	0.0232
0.050	25.053	77.193	87.043	0.0235
0.055	25.700	79.317	89.167	0.0239
0.060	26.400	81.673	91.523	0.0242
0.065	27.172	84.220	94.070	0.0246
0.070	27.957	85.889	95.739	0.0250
0.075	28.773	89.743	99.593	0.0254
0.080	29.636	92.696	102.546	0.0258
0.085	30.547	95.756	105.606	0.0262
0.090	31.502	98.848	108.698	0.0266

谱红移的对比散点图。从图中我们可以看出，弥散度远远小于输入参数为4个色指数+r的组合。此时的剩余标准偏差$\sigma_{\rm rms}$=0.0189。eClass是SDSS星表中一个用来标志光谱类型的参数，它是一个连续的值为-0.5∼1 的实数，值越小表示早型星，值大时对应晚型星。

基于测光数据，Collister和Lahav利用ANNs，Wadadekar利用SVM预测了星系的光谱类型（eClass）。仿效他们的做法，核回归也可以用来预测星系的光谱类型。我们用color+r和已知的eClass作为输入参数来预测光谱类型eClass。与预测红移不同的地方在于红移预测的输出为红移（z）而预测星系类型的输出为光谱类型（eClass）。我们得到的星系类型的剩余标准

表 3.8: 不同输入参数的剩余标准偏差$\sigma_{\rm rms}$及对应的最优窗宽

输入参数*	剩余标准偏差$\sigma_{\rm rms}$	窗宽h
$ugriz$	0.0215	0.025
$ugriz$+petroR50+petroR90	0.0247	0.070
$ugriz$+fracDeV_r	0.0223	0.035
$ugriz$+eClass	0.0198	0.025
color	0.0220	0.020
color+r	0.0206	0.030
color+$r+c$	0.0206	0.035
color+r+petroR50+petroR90	0.0226	0.050
color+fracDeV_r	0.0220	0.025
color+$ugriz$	0.0210	0.040
color+r+eClass	0.0189	0.025

注：—-petroR50是r波段50%的光度半径；petroR90 是r波段90%的光度半径；fracDeV_r是r波段的fracDeV；color是色指数，如：$u-g$、$g-r$、$r-i$、$i-z$；c =petroR90/petroR50.

偏差为$\sigma_{\rm rms}$ =0.0337。Wadadekar利用SVM的方法对10,000个样本预测，得到的eClass剩余标准偏差为$\sigma_{\rm rms}$ =0.057；Collister和Lahav用了64,175个星系样本预测eClass，剩余标准偏差$\sigma_{\rm rms}$ =0.052。图3.15显示了用非参数方法预测的测光eClass与光谱eClass的对比散点图。

为了优化不变窗宽的结果，我们将样本按红移升序排列，在红移0∼0.5(最大红移)间等分成33份，红移差为0.015，将落入不同红移区间内的样本分别进行核回归，找到在此红移区间内的最优窗宽，这样我们绘制出了红移（z）与最优窗宽（h）间的关系图，如图3.16所示。根据图3.16红移与最优窗宽的关系，我们采用了两种方法拟合这种关系：样条拟合方法（如图3.17所示）、多项式回归方法（如图3.18所示）。

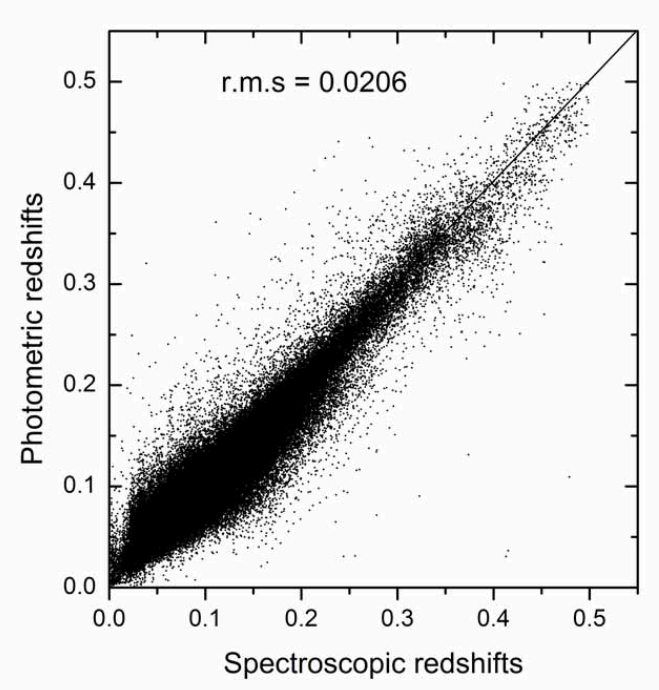


图 3.13: SDSS光谱红移与用核回归方法得到的测光红移的对比散点图。训练样本是260,000，测试样本是139,929。输入参数为四个色指数和r星等，即$u-g$、$g-r$、$r-i$、$i-z$、r。

图3.17中所示的黑色小正方形代表33个红移区间的最优窗宽。虚线代表将33个点连起来的连接线。实线代表用样条拟合方法拟合的红移与最优窗宽的关系。 图3.18中所示的黑色小正方形代表33个红移区间的最优窗宽。虚线代表将33个点连起来的连接线。实线代表用多项式回归的方法拟合的红移与最优窗宽的关系。

利用3.17和3.18中拟合的两种关系，我们做了如下实验：首先，我们将样本分成260,000做训练和139,929做测试。利用核回归的方法，输入参数为四个色指数，窗宽h为0.02，对上述样本进行回归估计，得到输出为固定窗宽红移（$z_{\rm fixed}$）。根据图3.17和3.18拟合的红移与窗宽的关系，每个$z_{\rm fixed}$都对应一个新的窗宽（$h_{\rm new}$），利用这个新窗宽，输入参数仍为四个色指数，重新计算变窗宽的红移（$z_{\rm new}$）。此时的$z_{\rm new}$ 为基于不变窗宽得到的变窗宽核回归预测的红移。利用z_{new}与SDSS的光谱红移计算新的剩余标准偏差$\sigma_{\rm rms}$。计算结果如表3.9所示。从表3.9中我们可以看到，基于不变窗宽基础上的自适应窗宽并没有达到我们预

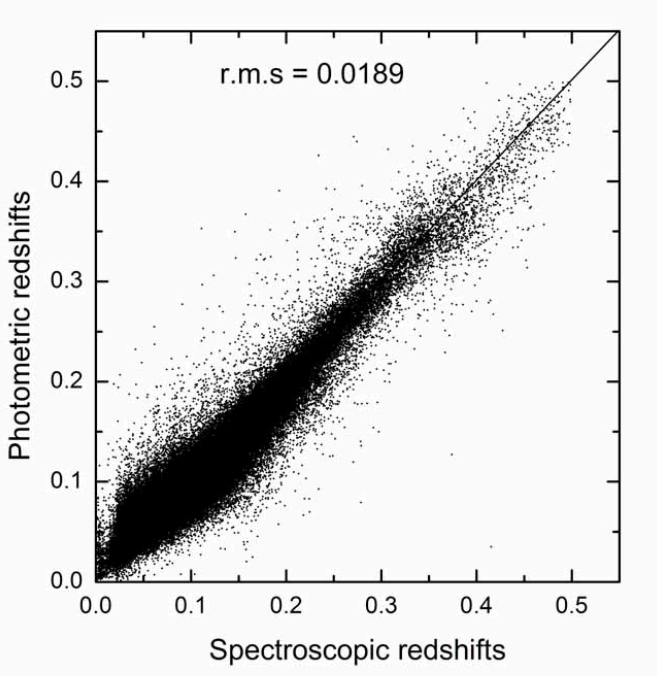


图 3.14: SDSS光谱红移与用核回归方法得到的测光红移的对比散点图。训练样本是260,000，测试样本是139,929。输入参数为四个色指数和r星等，即$u-g$、$g-r$、$r-i$、$i-z$、r 和eClass。

期的优化不变窗宽预测精度的目的。它们的弥散比不变窗宽的弥散还要大。也就是说这种重复叠代的方法是发散的，并不收敛，所以用这种方法是不能优化不变窗宽下的预测精度。

既然eClass有助于提高预测测光红移的精度，我们考虑对星系进行分类是否也有同样的效果。我们将样本分成早型星系和晚型星系两部分。根据Strateva[68]研究显示，当汇聚指数c >2.5时此星系为早型星系，c <2.5时为晚型星系。这样我们得到了251,794个早型星，148,135个晚型星。我们用这两个新的样本集做了表3.10中的实验。从表3.10中我们可以看出，早型星系的预测结果远远好于晚型星系。当输入参数为color时，早型星系的剩余标准偏差$\sigma_{\rm rms}$ =0.0197，而晚型星系的剩余标准偏差$\sigma_{\rm rms}$ =0.0247，早型星系与晚型星系合并后的剩余标准偏差$\sigma_{\rm rms}$ =0.0215，合并公式如下所示。合并的剩余标准偏差$\sigma_{\rm rms}$为0.0215要好于原样本的0.0220。同样对于输入参数为color+r和color+eClass时，早型星系

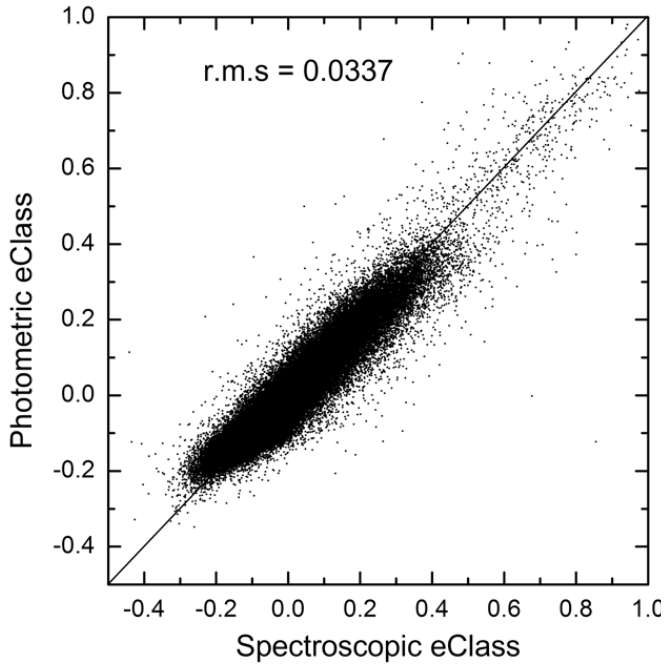


图 3.15: SDSS光谱eClass与核回归方法预测的测光eClass的对比散点图。训练样本是260,000，测试样本是139,929。

的预测结果都十分令人满意。

$$\sigma_{\rm rms}^{W}=\sqrt{\frac{1}{N_{\rm E}+N_{\rm L}}(\sum_{\rm i=1}^{N_{\rm E}}(z_{\rm Es}-z_{\rm Ep})^2+\sum_{\rm i=1}^{N_{\rm L}}(z_{\rm Ls}-z_{\rm Lp})^{2)}} \tag{3.75}$$

其中$N_{\rm E}$、$N_{\rm L}$分别是早型星系和晚型星系的个数；$z_{\rm Es}$、$z_{\rm Ep}$为早型星系的光谱红移和测光红移；$z_{\rm Ls}$、$z_{\rm Lp}$为晚型星系的光谱红移和测光红移；

表 3.9: 变窗宽下的剩余标准偏差$\sigma_{\rm rms}$

窗宽h	剩余标准偏差$\sigma_{\rm rms}$	拟合方法
固定窗宽(0.02)	0.0220	没用拟合方法
变窗宽	0.0224	多项式拟合
变窗宽	0.0222	样条拟合

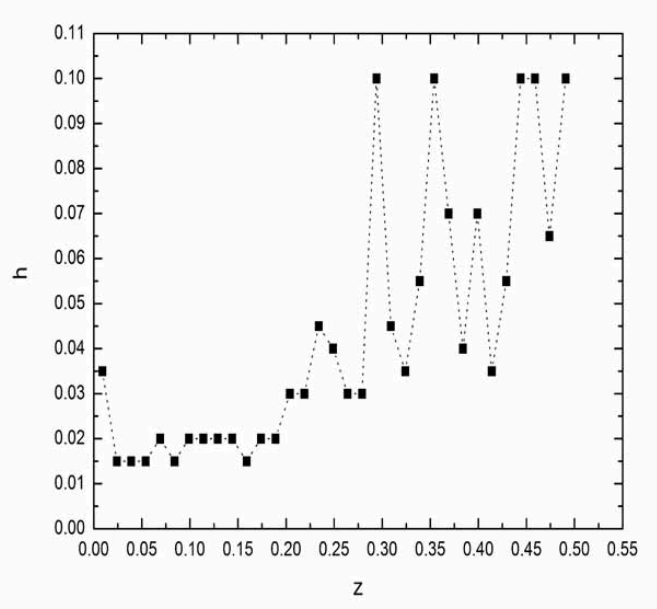


图 3.16: 不同红移区间的最优窗宽和红移关系图.

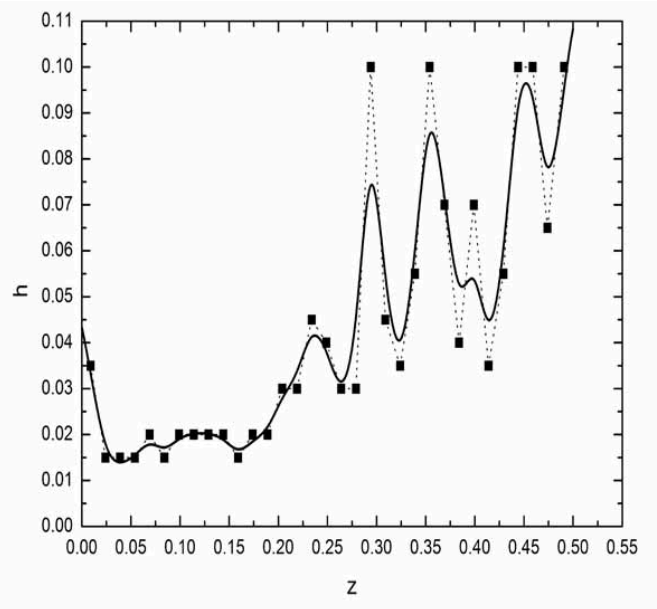


图 3.17: 用样条拟合方法拟合的红移与最优窗宽的关系。

（3）实验三

我们使用SDSS DR5中所有有光谱红移的类星体样本，且需要满足以下

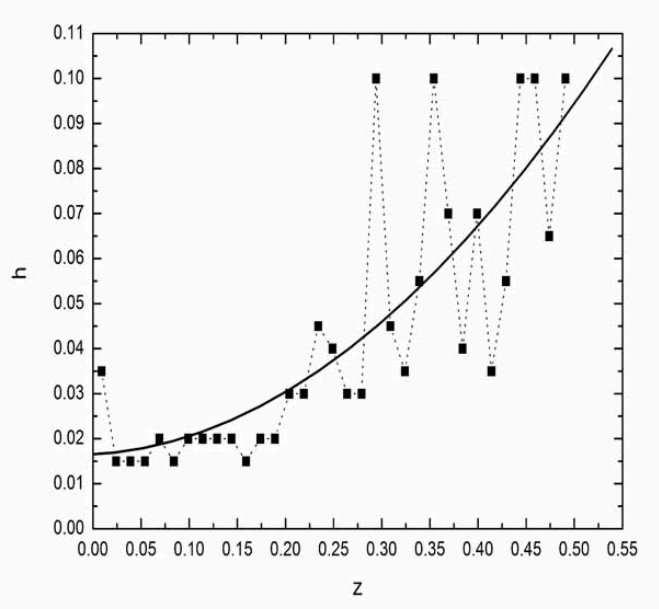


图 3.18: 用多项式回归方法拟合的红移与最优窗宽的关系。

表 3.10: 早型、晚型样本的剩余标准偏差$\sigma_{\rm rms}$与合并后的剩余标准偏差$\sigma_{\rm rms}$对比

输入参数	$\sigma^E_{\rm rms}$	$\sigma^L_{\rm rms}$	$\sigma^W_{\rm rms}$	$\sigma_{\rm rms}$
color	0.0197	0.0247	0.0215	0.0220
color+r	0.0186	0.0230	0.0204	0.0206
color+eClass	0.0164	0.0222	0.0187	0.0189

注—$\sigma^E_{\rm rms}$是早型星系的$\sigma_{\rm rms}$; $\sigma^L_{\rm rms}$是晚型星系的$\sigma_{\rm rms}$; $\sigma^W_{\rm rms}$是早型星系和晚型星系样本合并后产生的σ_{rms}; $\sigma_{\rm rms}$是来自于表3.8。

三个条件:（a）SpecClass = 3（证明此光谱是类星体）；(b)SpecZWarining = 0;（c）SpecZStatus > 2。满足以上条件的类星体样本为67,492。

我们利用此样本，预测了类星体的红移，输入参数为color+r，输出结果为类星体的测光红移，表3.11列举了六种评判预测精度优劣的标准。

从表3.11中可以看到，当窗宽为0.015时，预测精度很高，$|\Delta z| < 0.3$的占84.72%。但是此时的损失点所占比率很高，达到了12.04%。固定窗宽的核回归算法自身有缺陷，即总是存在损失点，因此在应用的时候不仅要考虑精度而且要考虑损失点所占比率不能太高。兼顾这两个因素，我们使用了CV最

表 3.11: 用核回归方法预测类星体测光红移的预测精度

窗宽h	$\|\Delta z\|<0.1$	$\|\Delta z\|<0.2$	$\|\Delta z\|<0.3$	$\|\frac{\Delta z}{1+z}\|<0.25$	σ^2_{rms}	方差	CV($\times10^{-2}$)	损失率
0.015	62.53%	78.82%	84.72%	88.61%	0.237	0.147	1.66	12.04%
0.020	61.40%	77.96%	83.79%	88.38%	0.233	0.138	1.66	6.31%
0.025	59.82%	76.91%	82.63%	88.12%	0.225	0.126	1.58	3.89%
0.030	58.19%	75.56%	81.74%	87.97%	0.218	0.115	1.52	2.59%
0.035	56.23%	74.22%	80.61%	87.91%	0.213	0.107	1.48	1.77%
0.040	54.24%	72.65%	79.52%	87.69%	0.210	0.101	1.46	1.29%
0.045	**52.63%**	**71.46%**	**78.59%**	**87.43%**	**0.207**	**0.096**	**1.45**	**1%**
0.050	50.94%	70.39%	77.62%	87.04%	0.206	0.093	1.47	0.75%
0.055	49.20%	68.95%	76.50%	86.56%	0.207	0.091	1.47	0.61%
0.060	47.61%	67.70%	75.58%	85.95%	0.208	0.090	1.49	0.45%
0.065	46.21%	66.28%	74.65%	85.45%	0.210	0.089	1.50	0.36%
0.070	44.74%	65.05%	73.66%	84.97%	0.212	0.088	1.57	0.27%

小作为评判标准。因此，当h =0.045时，CV得到最小值1.45×10^{-2}，认为此时得到的回归器最好。即$|\Delta z|<0.1$的预测百分比占52.63%，$|\Delta z|<0.2$ 的占71.46%，$|\Delta z|<0.3$的占78.59%，$|\frac{\Delta z}{1+z}|<0.25$的占87.43%，$\sigma^2_{\rm rms}$=0.207；方差为0.096，损失率为1%。

对于核回归来说，最主要的工作就是窗宽的选择。最优窗宽是随着输入参数的不同而不同的。在以上三个实验中我们可以看出，只有与星系类型相关的参数才可以提高红移的预测精度，例如：光谱类型eClass和汇聚指数c；其它一些无关的参数（例如petroR50、petroR90）不但不能提高预测精度反而会产生更大的弥散。用核回归方法预测eClass也取得了令人满意的结果（$\sigma_{\rm rms}$=0.0337）。并且在实验中，我们也发现对星系样本分成早型和晚型时，对提高红移的预测精度也起到了一定的作用。尤其值得注意的是，对于早型星系来说，红移的预测精度已经达到了0.016。这是其它测红移方法无法比拟的。

核回归也有一些不足，首先由于核回归没有所谓的训练过程，它是将所有的训练样本都储存在内存中，在测试过程中，每个测试样本都要遍历训练样本，这就对计算机的内存要求很高，而且当样本数很大时，需要很长的测试时间。此外，核回归有“盲点”，也就是说有些测试点在固定窗宽内找不到训练样本，导致无法预测，这样的点就被损失了。当然对损失点可以采用增大窗宽的方法解决。但是随着窗宽的增大，预测精度会降低，所以如何找到损失点与预测精度

之间的平衡点是个重要的问题。

对于固定窗宽的核回归，当样本是非均匀分布时，此时若采用定窗宽估计，则当窗宽太小时，在样本密度小的时候，因参加平均的观测点小，故核估计的偏差较大。而当窗宽过大时，虽然提高了低密度样本区的估计的精度，但是对于高密度的区域，因参加平均的观测点过多，造成估计的偏差增大。在现实中，样本的分布多为非均匀的情况，因此用自适应窗宽代替不变窗宽，这样既解决了损失点的问题同时也提高了预测精度。自适应窗宽的原理如公式（3.75）：

$$\hat{m}_{\mathrm{n}}(x,h_{\mathrm{n}},\alpha)=\frac{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}K_{h_{\mathrm{n}}/\alpha(X_{\mathrm{i}})}(X_{\mathrm{i}}-x)Y_{\mathrm{i}}}{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}K_{h_{\mathrm{n}}/\alpha(X_{\mathrm{i}})}(X_{\mathrm{i}}-x)} \tag{3.76}$$

其中h_{n}为不变窗宽，$\alpha(X_{\mathrm{i}})$为与解释变量密度函数有关的变窗宽函数。同时可以试验用不同的距离函数代替现在的欧式距离。

3.5 结论与展望

本章主要讨论了四种预测红移的方法即：颜色-星等-红移关系（CMR）、多变量多项式回归（MPR）、支持矢量机（SVMs）和核回归（KR）。所用样本分别是基于SDSS DR4、DR5和2MASS的巡天数据的。应用了SDSS DR4的星系数据，DR5的星系和类星体数据以及2MASS与SDSS DR5交叉证认的数据，进行了多组实验，并且研究了SVMs和KR方法对不同输入模型对预测测光红移精度的影响。在核回归和SVM方法中，我们尝试用PCA对高维数据进行降维。然后用降维后的数据进行回归估计。结果表明，经过PCA降维后的数据并没有提高红移的预测精度，而且随着波段的增加，预测精度也不一定提高。例如：在SDSS数据基础上考虑2MASS数据，预测精度并未提高。

目前，应用于测光红移的方法远不限我们采用的以上四种方法，还有模板匹配、神经网络等。每种算法都有自己的优缺点。为了比较不同方法的性能，我们将实验结果与他人的结果进行了比较，如表3.12所示。该表列出了各种测光算法预测红移的剩余标准偏差σ_{rms}值。

表 3.12: 不同测光红移方法所用数据样本及其对应的剩余标准偏差$\sigma_{\rm rms}$

方法	剩余标准偏差$\sigma_{\rm rms}$	数据样本	输入参数
CWW[1]	0.0666	SDSS-EDR	$ugriz$
Bruzual-Charlot[1]	0.0552	SDSS-EDR	$ugriz$
Bayesian[2]	0.0476	HDF-N	$UBVIJHK$(PCA)
CMR[3]	0.0320	SDSS DR4	$ugriz$
Interpolated[1]	0.0451	SDSS-EDR	$ugriz$
Polynomial[1]	0.0318	SDSS-EDR	$ugriz$
MPR[3]	0.0256	SDSS DR5	color
Kd-tree[1]	0.0254	SDSS-EDR	$ugriz$
ClassX[4]	0.0340	SDSS-DR2	$ugriz$
SVMs[5]	0.027	SDSS-DR2	$ugriz$
	0.0230	SDSS-DR2	$ugriz$+petroR50+petroR90
SVMs[3]	0.0273	SDSS DR5,2MASS	color
ANNs[6]	0.0229	SDSS-DR1	$ugriz$
Polynomial[7]	0.025	SDSS-DR1,GALEX	$ugriz+nuv$
KR[3]	0.0215	SDSS-DR5	$ugriz$
	0.0206	SDSS-DR5	color+r
	0.0189	SDSS-DR5	color+eClass
	0.0193	SDSS-DR5,2MASS	color

NOTE.— SDSS-EDR表示SDSS的早期数据(Stoughton等人，2002)，SDSS-DR1表示SDSS第一批数据(Abazajian等人，2003)，SDSS-DR2表示SDSS第二批数据(Abazajian等人，2004)，SDSS-DR5表示SDSS第五批数据(Adelman-McCarthy等人，2007). petroR50是r波段50% 光度半径，petroR90是r波段90%光度半径，fracDeV_r是r波段的fracDeV，color是色指数，例如：$u-g$、$g-r$、$r-i$、$i-z$.

(1)来自文献Csabai 2003[38]; (2)来自文献Benitz 2000[39]; (3)来自于我们的工作; (4)来自文献Suchkov, Hanisch & Margonet 2005[67];

(5)来自文献Wadadekar 2005[34]; (6)来自文献Collister & Lahav 2004[43]; (7)来自文献Budavári et al. 2005[30]。

因为红移预测精度不仅依赖于测红移的方法而且还依赖于使用样本及所用参数，因此我们只能对各个方法预测红移的效果进行粗略的比较。从表3.12中我们可以看出，核回归和ANNs的预测精度最高，均优于SVMs、Kd树、CMR、ClassX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$Ly\alpha$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光谱能量分布（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，因此它在预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。模板匹配方法的预测精度强烈的依赖于准确的具有代表性的SED 模板。通常的模板是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。换而言之，对模板匹配而言，最大的缺点就是模板构造的不完备性，导致其在预测红移时精度要劣于其它方法。

Benitz[39]提出了一种HyperZ和Bayesian marginalization合并的方法。该方法有助于包含天体的相关信息，例如：红移分布的预期形状和星系类型。这点正是其它方法所忽略掉的。这种方法大大提高了HyperZ方法预测红移的精度。但是在应用过程中对某些具体的课题可能参入虚假因素的影响。而且预测红移精度不高。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。

颜色-星等-红移关系（CMR）方法是通过将样本按r星等分区，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预测红移。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预测不够准确。而且对于星等r大于23mag时无法预测。在星等r <21mag时，红移预测损失率为5%，这种损失是不可弥补的。

多项式回归方法原理上比较简单，它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预测那些未知红移的星系样本。多项式回归方法不需要知道星系光谱演化信息，易于天文学家的理解。不过其预测红移速度很快，对于大样本而言略有优势。但是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，而且在高红移区，光谱红移样本很不完备，这样对于高红移星系样本的红移预测也就很不可靠。

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，常常限入局部极小值。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[34]。从预测红移精度而言，ANN方法无疑是很好的选择，但其速训练度相对较慢。

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预测结果。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。不过，SVMs不存在ANNs的诸多缺点，例如：过度拟合，局部极小等。

核回归属于事例学习家族中的一员，具备事例学习的一切优缺点。它们典型的优点是不需要训练过程，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。求出对应测试样本的红移。由核回归的特点可知，当训练样本数目很大的时候，需要花费较长的测试时间，并且占用很大的内存。核回归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—预测精度相当高，改进的核回归方法将是预测测光红移方法的完美选择。

与其他人的工作相比较，我们工作的创新之处在于首次应用了核回归来预测星系和类星体的测光红移。核回归有很大的灵活性，它不仅可以替换核函数而且也可以使用不同的距离公式，在我们的实验中使用了高斯核和欧式距离。对于核回归方法最重要的一环就是窗宽的确定，可以使用不同的方法来确定最优窗宽，例如交错鉴定方法、AIC和BIC。从实验结果可以看出，不同的输入参数对应不同的最优窗宽，而且不是参数个数越多预测精度就越高，这点完全有别于ANNs，合适的几个参数就可以达到理想的效果。例如：输入参数为四个色指数时，剩余标准偏差$\sigma_{\rm rms}$=0.0220; 输入参数为四个色指数加r星等时，剩余标准偏差$\sigma_{\rm rms}$=0.0206; 对于SDSS DR5与2MASS交叉证认的样本，只考虑SDSS的四个色指数时，剩余标准偏差$\sigma_{\rm rms}$=0.0193。核回归对于高红移样本密度比较稀疏的区域更显示出它的优越性，克服掉训练集方法不可以外推的

缺陷。类似于Collister和Lahav利用ANNs，Wadadekar利用SVMs预测星系的光谱类型eClass（ANNs: $\sigma_{\rm rms}$ =0.052; SVMs：$\sigma_{\rm rms}$ =0.057），我们用核回归方法对eClass进行了预测，并取得了令人满意的结果（$\sigma_{\rm rms}$=0.0337）。

其次，我们用核回归预测测光红移时，首次使用了别人从未考虑过的光谱类型参数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，是一个连续值，-0.5∼1 的实数，值越小表示早型星系，值大时对应晚型星系。实验结果表明，增加该参数可以有效地改进测光红移的预测精度（$\sigma_{\rm rms}$=0.0189），如表3.12。只要那些与星系类型相关的参数（例如：eClass）才可能有效地提高预测红移的精度。

再者，考虑到光谱类型有助于提高测光红移的精度，因此，我们尝试将星系先分类而后预测红移。按照汇聚指数（c）将样本分为早型星系与晚型星系独立预测红移。发现合并后的精度比原来整体预测的精度要高。当输入参数为四个色指数时，合并剩余标准偏差为σ_{rms}=0.0215，原来整体精度为σ_{rms}=0.0220；当输入参数是四个色指数加r星等时，合并剩余标准偏差为$\sigma_{\rm rms}$=0.0204，原来整体剩余标准偏差为$\sigma_{\rm rms}$=0.0206；当输入参数为四个色指数加r和eClass时，合并剩余标准偏差为$\sigma_{\rm rms}$=0.0187，原来整体剩余标准偏差为$\sigma_{\rm rms}$=0.0189。尤其对早型星系来说，精度提高得很可观。当输入参数为四个色指数时，剩余标准偏差$\sigma_{\rm rms}$=0.0197；当输入参数是四个色指数加r星等时，剩余标准偏差$\sigma_{\rm rms}$=0.0186；当输入参数为四个色指数加r和eClass时，剩余标准偏差$\sigma_{\rm rms}$=0.0164。对晚型星而言，精度似乎不是很理想。

在未来的工作中，我们将考虑采用不同的数据样本和改进算法。对于数据样本，将应用其它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。对于核回归而言，尝试用不同的距离函数和核函数，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，而且在预测精度上也要高于固定窗宽。为了加快核回归的效率，我们将使用Deng&Moore[69] 提出的多分辨率事例学习方法，这样可以大大缩短了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—最小二乘矢量机预测红移。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。

第四章　虚拟天文台图像处理与分析工具的设计和实现

4.1　多波段天文学

一直以来，多波段天文观测是天文学家的梦想。上个世纪后五十年中，随着探测器和空间技术的发展，天文观测从可见光、射电波段扩展到红外、紫外、X射线和γ射线在内的电磁波各个波段，形成了多波段天文学，甚至全波段天文学。天文研究进入了一个全新的阶段。

4.1.1　光学天文学

光学天文学是利用天体在光学波段的辐射来研究天体现象的学科，是天文学中发展最早的学科。现在的光学天文学主要是利用大口径光学望远镜及其焦面附属仪器来研究天体的形态、结构、运动特性、物理状态、演化阶段和化学成分的一门学科。尽管近十年来天文学已经进入了全波段天文学时期，但是光学天文学仍然是天文学的核心。

下面以几个光学波段的巡天和光学星表为例，来介绍光学天文学的发展。

SDSS（Sloan Digital Sky Survey，简称SDSS [45]）是目前最有雄心的训天计划。它使用一架口径2.5m的大视场望远镜进行测光和多光纤光谱观测，测光极限星等22.2mag（R），光谱极限星等17.77mag。SDSS计划对全天的1/4天区进行训天，得到近亿颗天体的准确坐标和绝对光度等观测数据，重点是河外天体，尤其是活动星系核，计划观测数目在百万颗以上。目前SDSS已经发现了一大批高红移的类星体和星系，其中包括迄今为止发现的最大红移的星系和类星系。

USNO（the United States Naval Observatory Astrometric，简称USNO [72]）星表是覆盖全天的巡天表，在低于极限星等B≈20mag时含有多于5亿个未确定源，它们的位置可为天体测量做参考。这些源是通过PMM（The Precision Measuring Machine）探测到的。整个巡天的全部数据量超过了10TB。该表由一系列二进制文件组成，并且这些文件是以源的位置组织的。既然源的密度随位置的不同而不同，因此每个文件中源的个数也各不相同。要准确地提取源的参数，需提供与数据匹配的软件工具。该星表包括源的位置，即赤经、赤纬，还有每个恒星的蓝星等和红星等.

大天区面积多目标光纤光谱天文望远镜（Large Sky Area Multi-Object Fiber Spectroscopic Telescope, 简称LAMOST [73]）于1997年正式立项，小系统于今年6月底出光。LAMOST是一台横卧于南北方向的中星仪式反射施密特望远镜，可观测天区的赤纬从-10度到+90度。相应于5度视场、直径为1.75米的焦面上放置4000根光纤。采用并行可控的光纤定位技术，可在最短的时间将光纤按星表位置精确定位，并提供了光纤位置微调的可能。这将在光纤定位技术上突破目前世界上同时定位640根光纤的技术。通过这样的构思和设计，解决了大视场的施密特望远镜投射改正板很难做大，大口径反射望远镜视场较小的问题，使LAMOST成为大口径兼大视场光学望远镜的世界之最。由于它的4米口径，在1.5小时曝光时间内以1纳米的光谱分辨率可以观测到20.5mag的暗弱天体的光谱；由于它相应于5°视场的1.75米焦面上可以放置数千根光纤，连接到多台光谱仪上，同时获得4000个天体的光谱，成为世纪上光谱获取率最高的望远镜。

4.1.2 射电天文学

射电天文学是通过观测天体的无限电波来研究天文现象的一门学科。理论上以近代物理为基础来分析研究天体的物理特性、化学组成和结构演变。由于大气的阻挡，从天体来的无限电波只有波长约1毫米到30米左右的才能到达地面，迄今为止，绝大部分的射电天文研究都是在这个波段内进行的。其中类星体、脉冲星和2.7K微波背景的发现是射电天文对近代天体物理的三大贡献。

NVSS（the National Radio Astronomical Observatory, Very Large Array, Sky Survey ，简称NVSS[74]）是一个射电巡天项目，覆盖赤纬-40°以北的天区。巡天星表包括超过180万个孤立源的全部强度和现行偏振图像测量量，其分辨率为45角秒和完备极限流量为25mJy。NVSS的主要数据是一组2326个连续映射立方体，每个立方体覆盖了4°×4°天区，其具有三个平面包含Stocke I、Q和U图像，还有关于这些图像的孤立源星表。每张大的图像是由大于100张更小的原始快照图像得到的。

FAST（A Five Hundred Meter Aperture Spherical Telescope，简称FAST [75])是500米口径球面射电望远镜，具有三项自主创新：利用贵州独一无二的天然喀斯特洼坑作为台址；洼坑内铺设约2000块单元组成500米球冠状主动反射面；采用轻型索拖动机构和并联机器人，实现望远镜接收机的高精度定位。FAST突破了望远镜的百米工程极限，开创了建造巨型射电望远镜新模

式。FAST将是国际上最大的望远镜，拥有30个足球场大的接收面积，与号称“地面最大的机器”德国波恩100米望远镜相比，其灵敏度提高约10倍；与排在阿波罗登月之前、被评为人类20世纪十大工程之首的美国Arecibo300米望远镜相比，其综合性能约高10倍。

4.1.3　红外天文学

红外天文学是利用电磁波的红外波段研究天体的一门学科。整个红外波段，包括波长0.7~25μm的近红外区和25~1000μm的远红外区。随着半导体物理学的发展和军事侦察的需要，研制出了灵敏度很高而热噪音很低的单元和阵列红外监测器件，红外天文学在近年获得了突飞猛进的发展。

2MASS（the Two Micron All-Sky Survey，简称2MASS [76]）是一个近红外（J、H和Ks）的全天巡天项目。该项目始于1997年，数据于2002年度释放完毕。2MASS利用了两台高度自动的1.3米的望远镜，每台配备一个具有三个通道的照相机，能够同时在三个波段J（1.25μm）、H（1.65μm）和Ks（2.17μm）观测天空。当巡天结束时，2MASS星表包括近3亿颗恒星、50万星系和星云在三个波段的天体测量和测光属性，以及多于12TB的图像数据。

IRAS（the NASA Infrared Processing and Analysis Center Infrared Science Archive，简称IRAS[77]）提供了获得各种数据的服务，主要针对红外数据。IRSA也提供了交叉证认工具，用户可以根据要求从各种源中选取候选目标。IRSA主要提供以浏览为基础的查询服务，包括星表和图像。IRSA含有多于15TB的数据，大部分与2MASS巡天相关。

4.1.4　X射电天文学

X射电天文学是通过X射线波段（波长0.01~100埃的电磁辐射）研究天体的一门学科。因为天体的X射线会受到地球大气的严重阻碍，所以主要通过卫星进行探测。因此，虽然X射线的探测始于二十世纪四十年代，但是成为一门学科，则是在人造地球卫星上天以后。

RASS（the ROSAT X-Ray All-Sky Survey，简称RASS[78]）是在1990~1991年，通过使用ROSAT（伦琴号卫星）位置灵敏正比计数器（ROSAT Position Sensitive Proportional Counter，简称PSPC）和ROSAT X射线望远镜（the ROSAT X-ray Telescope，简称XRT）得到的X射线星表。ROSAT卫星于1990年6月1日发

射，所携带的观测设备比以前的X射线观测仪器具备更高的定点观测的灵敏度。ROSAT发现了60,000多个X射线源，包括25,000多个活动星系核、20,000多个恒星和5,000多个星系团，并且得到了分子云、中子星、彗星等天体的X射线辐射的观测资料。

另一个很有名气的X射线数据中心是Chandra数据库。Chandra射电天文台是1999年7月由哥伦比亚号航天飞机发射升空的。Chandra 比以前的X射线望远镜有更高分辨率、更高灵敏度以及更大的接受面积，适合于观测暗弱源。不仅有助于研究黑洞、超新星、暗物质等天体，还有增进人们对宇宙起源、演化过程的认识[79]。

4.1.5 紫外天文学

利用天体在100到4000埃的紫外波长的辐射来研究天文现象的学科。由于大气对紫外波段的吸收十分严重，因此需要在高空或大气外进行观测。

1990年发射的ROSAT卫星利用WFC（UK Wide Field Camera）进行的第一次极紫外全天巡天（the Rosat WFC All-Sky Survey）。1992年发射的极远紫外探测卫星（the Extreme Ultraviolet Explorer, 简称EUVE）和1999年发射升空的远紫外光谱探测卫星（Far Ultraviolet Spectroscopic Explorer，简称FUSE）进行的巡天观测。

4.2 虚拟天文台

随着望远镜、探测器、计算机和互联网技术的发展以及大量天文数据和资料的网络共享。天文学界认为有能力且有必要建立全球性的望远镜－虚拟天文台（Virtual Observatory，简称VO），将全球的天文数据统一到一个实体中，为任何地方和领域的人们所利用。如果说利用γ射线巡天、X射电巡天、紫外巡天、光学巡天、红外巡天和射电巡天所得到的观测数据，用适合的方法对数据进行统一规范的整理、归档，便可以构成一个全波段的数字虚拟天空；而根据用户需求获得某天区的各类数据，就仿佛是在使用一架虚拟的天文望远镜；如果再根据科学研究的要求开发出功能强大的计算工具、统计分析工具和数据挖掘工具，这就相当于拥有了虚拟的各种探测设备。这样，有虚拟的数字天空、虚拟的望远镜和虚拟的探测设备所组成的机构便是一个独一无二的虚拟天文台。虚拟天文台是互联网时代天文学发展的必然产物[80]。

虚拟天文台最早是由美国国家科学院在天文与天体物理发展的新十年展望中提出的。美国国家虚拟天文台（the National Virtual Observatory, 简称NVO[81]）的概念一经提出，立即引起天文界的广泛重视。各国也纷纷相继出台了自己的虚拟天文台计划。中国虚拟天文台（Virtual Observatory of China, 简称China-VO [82]）于2002年建立。以美国、英国、欧洲为首的虚拟天文台组织在2002年6月在德国的“国际虚拟天文台大会”上决定成立国际虚拟天文台联盟（International Virtual Observatory Alliance，简称IVOA [83]）。同年，中国加入了国际虚拟天文台联盟。至今国际虚拟天文台已发展到了16个成员。

中国虚拟天文台的建设是很有必要的:

1.只有加入国际虚拟天文台联盟，才能以平等的身份全方位共享IVOA的技术与资源；

2.建设China-VO是实现LAMOST、BATC巡天等我国自产数据与IVO数据融合的最佳途径；

3.建设China-VO可以培养一批与IVO相适应的天文学家和技术人才，为未来中国天文学的发展提供智力支持；

4.可以利用IVO丰富的资源加强教育与科学普及工作，提高我国公众的科学素质；

5.为国内网格技术研究提供最好的试验场，推动国内IT技术的发展。

大样本光谱巡天是LAMOST项目的主要目标，把LAMOST建设成为面向VO的LAMOST（VO-Oriented LAMOST），不但是LAMOST本身的需要，也将是中国对世界天文的重大贡献。它的观测数据将成为China-VO数据库的重要组成部分，为国际虚拟天文台和天文学的发展做出应有的贡献。我们要建设有“中国特色”的中国虚拟天文台。科学上，China-VO将采用与LAMOST项目紧密结合的方式，充分发挥LAMOST光学光谱数据中心的作用，使光谱数据及其相关处理技术成为China-VO的核心与特色。技术上，China-VO必须最大程度地集中国内各天文研究机构、IT研究机构和数学等相关领域的人才资源，共同努力实现目标。此外，China-VO必须采用开放的运作方式，与国际上各虚拟天文台计划开展充分的合作，在IVOA中发挥积极作用，充分发挥沟通国内和国际各种资源服务与共享的桥梁作用。IVO的提出与发展离不开计算机与网络技术的直接推动，但同时也为IT技术提出了挑战，比如分布式多数据库联合查询、分布式异构数据的互操作性、大规模多维数据统计分析与可视化、授权认证管理等

等。基于海量的LAMOST自产数据的自动处理的流程将逐步成为China-VO数据服务的重要范例。为实现这一范例，China-VO除了IVO共同面临的一些挑战外，还必须解决下面的问题：1.海量多光纤光谱观测数据的自动处理；2.光纤光谱谱线的自动提取；3.光谱自动分类；4.光谱红移的自动测量；5.光纤光谱数据与其它类型天文数据的融合；6.光谱数据的可视化。

国际虚拟天文台联盟自建立以来，就开始组织和协调各国虚拟天文台为实现虚拟天文台的宏伟蓝图而积极努力，开发各种软件和工具为天文学家提供良好的科研平台。到目前为止，一些VO项目已经率先开发出一系列实用的工具，例如：交叉证认工具OpenSkyQuery，数据分析、挖掘和可视化工具Mirage，数据融合工具与整合工具VizieR，光谱处理工具SpecView，智能天文"鼠标"Skymouse [85]，而且还有一些图像处理工具，对FITS图像进行操作的工具有fv [86]，FTOOLS [87]和FITSIL/CFITSIO [88]，天文图像可视化的工具有Aladin [89]，DS9 [90]，SAOImage [91]，VOPlot [84]。

4.3 虚拟天文台图像处理与分析工具

中国虚拟天文台作为国际虚拟天文台的一员，也在开发自己的工具或服务来参与国际合作。上小节提到的图像处理工具不能简单地集成到中国虚拟天文台的软件包中，为此，我们借鉴法国斯特拉斯堡数据中心开发的数据整合工具Aladin的思想和我们项目的需要开发了中国虚拟天文台图像处理分析工具（VO_IMPAT [70]），它提供了对数字巡天图像数据（Digital Sky Survey, 简称DSS [93]）、天文星表以及其他数据库的交互访问，可互动地可视化天空任何一部分图像，并可与天文星表或用户上传的文件叠加。

VO_IMPAT星表数据来自于北京天文数据中心（Beijing Astronomical Data Center，简称BADC [92]）。BADC是WDC（World Data Center，简称WDC）系统里唯一的天文数据中心。BADC从80年代开始做数据共享的工作。通过广泛的数据征集，国际数据交换和镜像，迅速成为天文学界著名的数据交流中心，它使用的是关系型数据库。目前比较流行的天文星表、数据、图像都可以在BADC上找到。DSS是一个地面上包括E、V、J、R和N波段的全天图像巡天，是由Palomar和UK施密特望远镜实现的。巡天产生的照相底片在STScI转化成数字信息并生成哈勃导星数据库（the Hubble Guide Stars Catalog，简称GSC）。每一张巡天底片覆盖了6.5° × 6.5°的天空，是通过一个修改后的图像数字仪（PDS）

测微密度计进行数字化的。数字图像每个像素的大小为25微米（1.7角秒/像素）或者15微米（1.0角秒/像素），每面有14000 × 14000 或者23040 × 23040个像素。这些图像被存放在12英寸的光学介质上，很难快速存取。为了方便地访问这些数据，使用了H-transform技术对图像进行压缩。数据的压缩率是1/10。这些压缩后的数据被写在CD上放入自动播放机进行快速存取。这样用户就可以快速地访问任何一部分天空的数字图像了。DSS可以提供FITS或者JPG的图像格式。图像可以通过Web得到，也可以通过StarView、Visual Target Tuner（VTT）或者Aladin这样的应用程序进行在线访问得到。

VO_IMPAT是个交互性的图像处理工具，其基本操作流程（流程图如图4.1所示）：当使用VO_IMPAT时，用户输入查找的区域和半径后，发送请求，通过Web Service传送到BADC服务器上，服务器分析查找请求后查询数据库，将查找到的数据返回给用户。VO_IMPAT允许用户可视化任何一部分天空，并将光学、红外、射电、X射线波段的星表数据叠加在DSS底图上。VO_IMPAT的设计目的是实现多波段天文数据的融合。可以将不同波段的星表叠加到DSS底图上，如光学波段的USNO星表、近红外波段的2MASS 星表、射电波段的NVSS星表和X射线波段的RASS星表。同时VO_IMPAT还可以对图像进行放大、缩小、伪彩色、等高线、直方图、尖锐化、平滑化、旋转等处理，不同的星表数据可以采用不同的颜色和图标显示。

4.3.1　用户界面

VO_IMPAT是基于Java语言开发的应用程序，显示以查找坐标为中心，查找半径为半径的矩形区域的天空图像，默认的查找半径是20″。它允许用户交互性地查找图像、星表和其他的数据，而且还提供了图像处理的基本功能。

VO_IMPAT的用户界面包括以下几部分，如图4.2所示。

（1）显示窗口：用户界面中间部分是显示窗口，用来显示图像和星表的投影；

（2）位置窗口：位于显示窗口的上方的矩形区域，用户需要在此处输入查

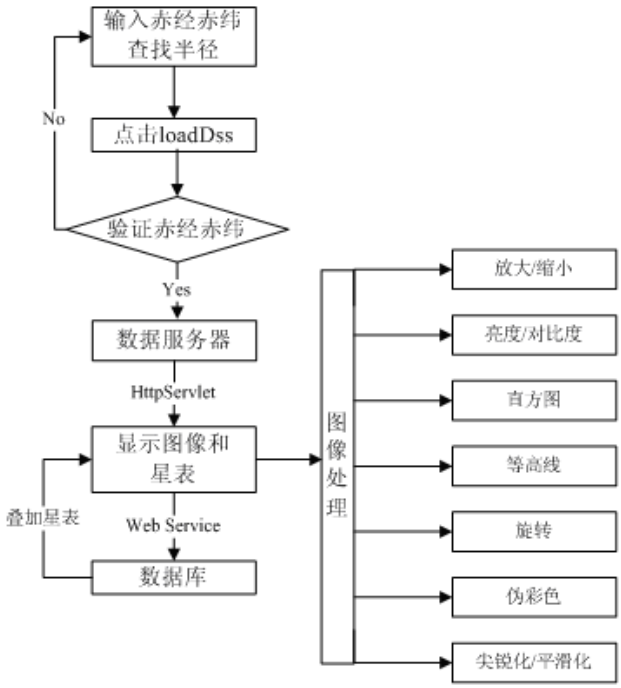


图 4.1: VO_IMPAT流程图。

找位置的赤经、赤纬和半径；

（3）导航窗口：位于用户界面的右上角方形区域，用于快速预览整个图像；

（4）放大缩小窗口：位于导航窗口的下方，用来显示放大或缩小的图像；

（5）鼠标窗口：位于放大窗口的下方，用来显示当前鼠标的位置；

（6）波段窗口：位于位置窗口的下方，有四种波段选择：光学（USNO）、红外（2MASS）、射电（NVSS）和X射线（RASS）；

（7）测量窗口：位于用户界面的最下方矩形区域，用来显示选中源的相关信息。

4.3.2 运行实例

打开VO_IMPAT的运行窗口，在位置窗口上输入想要查找天区的赤经赤纬与查找半径，点击“Load DSS”按钮，就可以将对应天区的DSS图像显示在显示窗口内，同时可以将其他波段的星表信息添加到DSS底图上。下面以M87 为例，

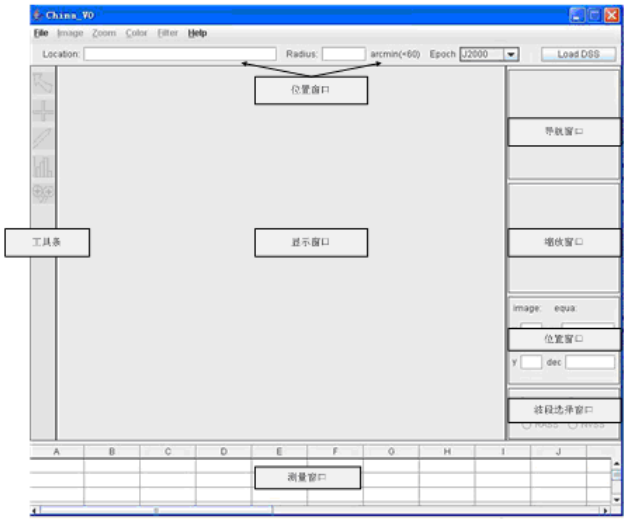


图 4.2: VO_IMPAT界面布局图。

简单描述VO_IMPAT的使用过程。

（1）在位置窗口内的Location区输入“12 30 49+12 23 37”，在Radius区内输入“20”，点击Load DSS按钮。此时对应天区查找半径为20′的DSS图与相应光学波段的USNO星表数据显示出来。其中Location的赤经赤纬坐标可以以空格分隔、冒号分隔、也可以写成度的形式（例如：187.704+12.394）。如图4.3所示。

（2）点击工具条上的Select按钮，将鼠标移到显示窗口，点击任意一个源，关于这个源的相应信息就显示在测量窗口内，例如：赤经、赤纬、历元、星等、...。拖拽测量窗口的上下左右的滚动条可以查看测量窗口的全部信息。

（3）点击工具条上的Tag按钮，可以将选中的源编号，并将该源的对应信息显示在测量窗口。如图4.4所示。

（4）要想画线，可以点击工具条上的Draw按钮，在显示窗口点击拖拽鼠标，就可以画出线。可以利用此工具在图像上进行标记、注释。

（5）VO_IMPAT提供了四种直方图模式，分别是线性、归一化、分段化和均衡化。工具条中的Histogram按钮实现的线性直方图的方法。如图4.5所示。

（6）VO_IMPAT提供了两种放大模式：一种是对局部区域进行放大；另一种是对整张图放大。局部放大通过工具条中的Magnify实现。点击Magnify按钮，

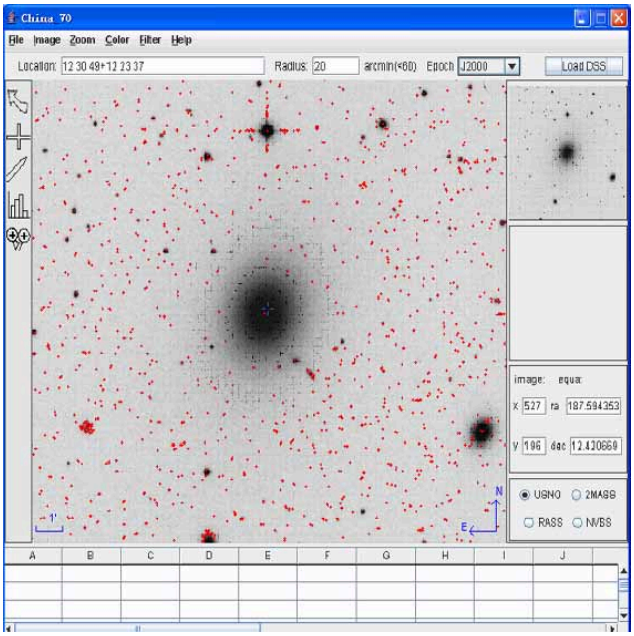


图 4.3: 以M87为例，显示DSS底图以及叠加的USNO星表。

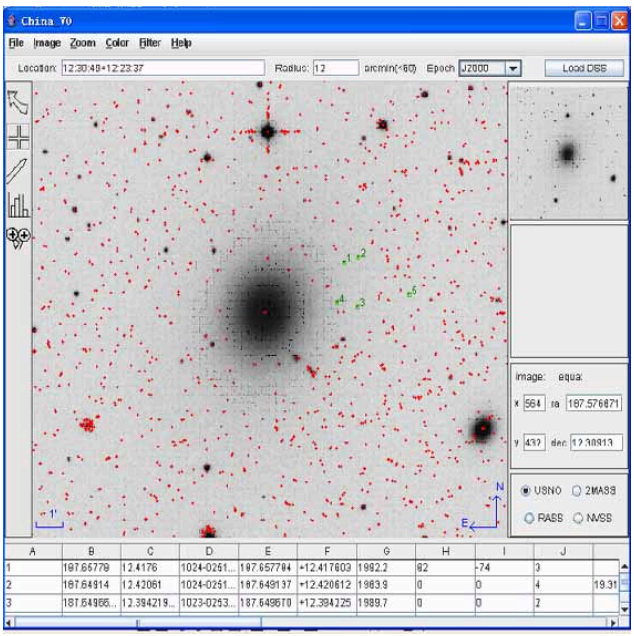


图 4.4: 以M87为例，选中目标。

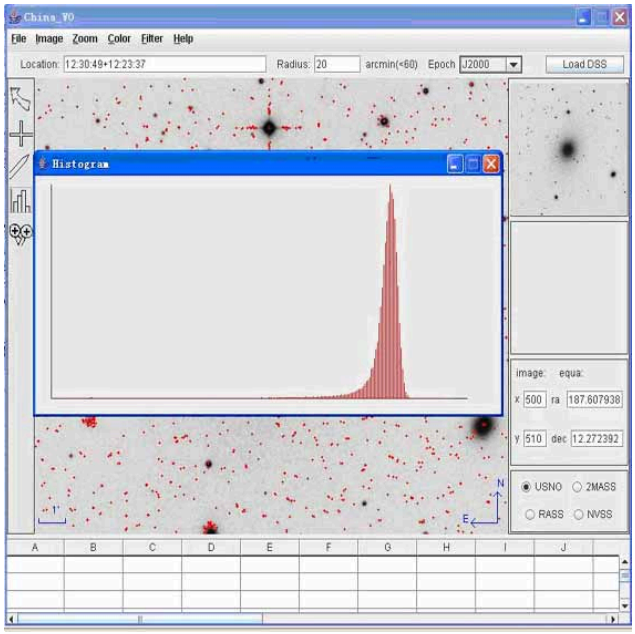


图 4.5: 以M87为例，直方图。

在显示窗口中点击任意想放大的区域，以该点为中心的一个小区域将被放大，放大后的图像同时显示在显示窗口和放大缩小窗口。如图4.6所示。

（7）为了实现多波段数据的融合，可以通过选择波段窗口内的不同波段数据库来实现。图4.7中叠加了光学波段的USNO、红外波段的2MASS和射电波段的NVSS。在图中“.”代表来自于USNO的数据，“×”代表2MASS，“◇”代表NVSS。

（8）VO_IMPAT允许用户上传本地数据，包括FITS图像和星表数据。

（9）VO_IMPAT可以将图像保存成FITS或者JPG格式，也可将星表保存成CSV或VOTable格式。

（10）在菜单栏Image中，可以对图像进行处理，包括对比度、亮度、旋转、等高线等。

（11）Color菜单栏中提供了反色功能，以及其他伪彩色处理。

（12）在Filter菜单栏中提供了对图像进行尖锐化和平滑化的处理。

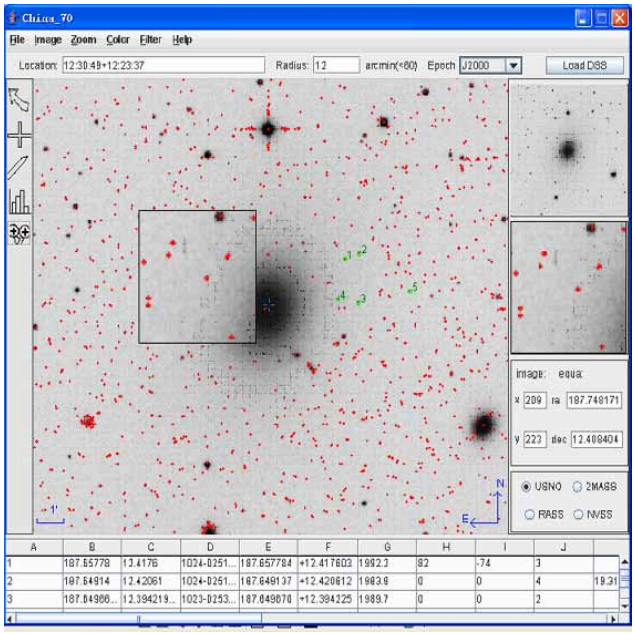


图 4.6: 以M87为例，等高线图。

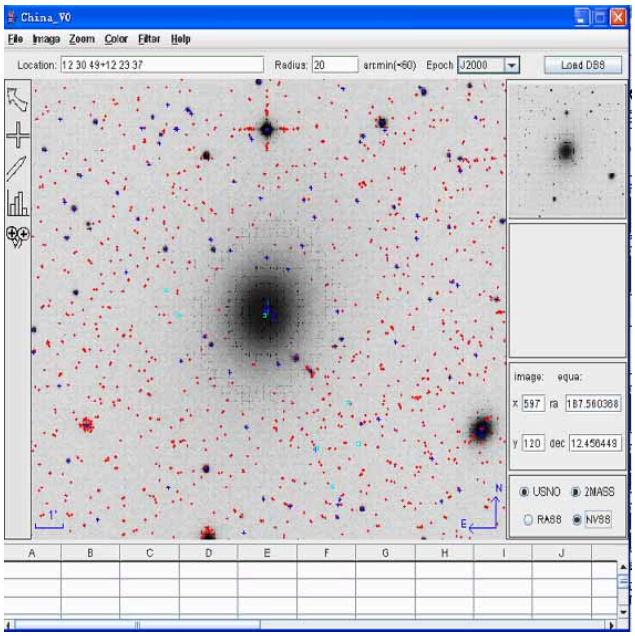


图 4.7: 以M87为例，四波段数据融合。

4.4　小结

我们开发的VO_IMPAT图像处理分析工具已经实现了多波段数据的融合。尽管功能还不尽完善，但是其是中国虚拟天文台开发的第一个图像处理工具，为中国虚拟天文台探索新的工具提供了宝贵的经验和重要的参考。在下一步工作中，我们将继续维护它的稳定性，并不断地提高它的功能的全面性和健壮性，以及在现有的基础上增加新的功能。VO_IMPAT可以帮助天文学家融合多波段数据来发现稀有的天体或现象，也可用于科普教育上。

VO_IMPAT目前是以jar包的形式存放在http://services.chinavo.org/vo_impat下。用户需要安装Java运行环境（Java Runtime Environment 1.5），AXIS 1.2以及JAI 1.1.3。双击安装目录下的start.sh就可以使用VO_IMPAT。希望天文学家使用后，将宝贵的意见反馈给我们，以帮助我们进一步提高和完善VO_IMPAT。

第五章　总结与展望

本论文的重要工作是基于测光红移算法的研究以及虚拟天文台工具的开发和应用。应用和对比了各种测光红移算法，并开发了简单的测光红移工具和图像处理工具。

在第一章中，我们从红移的概念出发，谈到了测光红移的由来、背景和意义，随后概述了几种比较典型的预测测光红移的算法，简要地介绍了这些算法的原理、预测红移的效率及它们的优缺点。

在第二章中，我们介绍了SDSS和2MASS巡天的基本情况及释放的数据产品，并详细地介绍了SDSS星系样本和类星体样本的常用星表参数。

第三章是本论文的核心。我们探讨了三种预测测光红移的经验训练集法（CMR、MPR和SVMs）和一种事例学习方法（KR）。详细地介绍了四种算法的基本原理和应用。尤其对SVMs和KR尝试了不同的样本集和各种输入参数，并将四种方法的预测结果与他人的工作进行了全面的比较。从各种算法的自身原理出发，仔细地剖析了它们用于预测测光红移的优缺点。并且简单介绍了我们开发的红移测量工具。

在第四章中介绍了基于VO环境下开发的图像处理分析工具VO_IMPAT。阐述了该工具开发的背景、目的、底层数据库、基本流程、用户界面和具体实现方案。

归根结底，预测测光红移属于数据挖掘的回归任务，因而，任何数据挖掘的回归算法原则上都可以用来预测红移。这些算法也可以作为回归算法的补充。

随着大规模巡天和天文观测技术的飞速发展，天文数据在急速增长。这些激增的数据背后隐藏着许多重要的信息。人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。这正是数据挖掘（Data Mining，简称DM）产生的原动力。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，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。确切地说，数据挖掘，又称数据库中的知识发现（Knowledge Discovery in Database，简称KDD），是指从大型数据库或数据仓库中提取隐含的、未知的、非平凡的及有潜在应用价值的信息或模式。它是数据库研究中的一个很有应用

价值的新领域，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。数据挖掘其实是知识发现的核心部分，而知识发现是在积累了大量数据后，从中识别出有效的、新颖的、潜在的、有用的及最终可以理解的知识，人们利用这些知识改进工作，提高效率和效益。数据挖掘是信息发展到一定程度的必然产物，是利用积累数据的一个高级阶段。用数据库管理系统来存储数据，借助机器学习的方法来分析数据，挖掘大量数据背后的知识，这两者的结合促成了数据库中的知识发现的产生。知识发现的基本流程如图5.1所示。 图5.1描

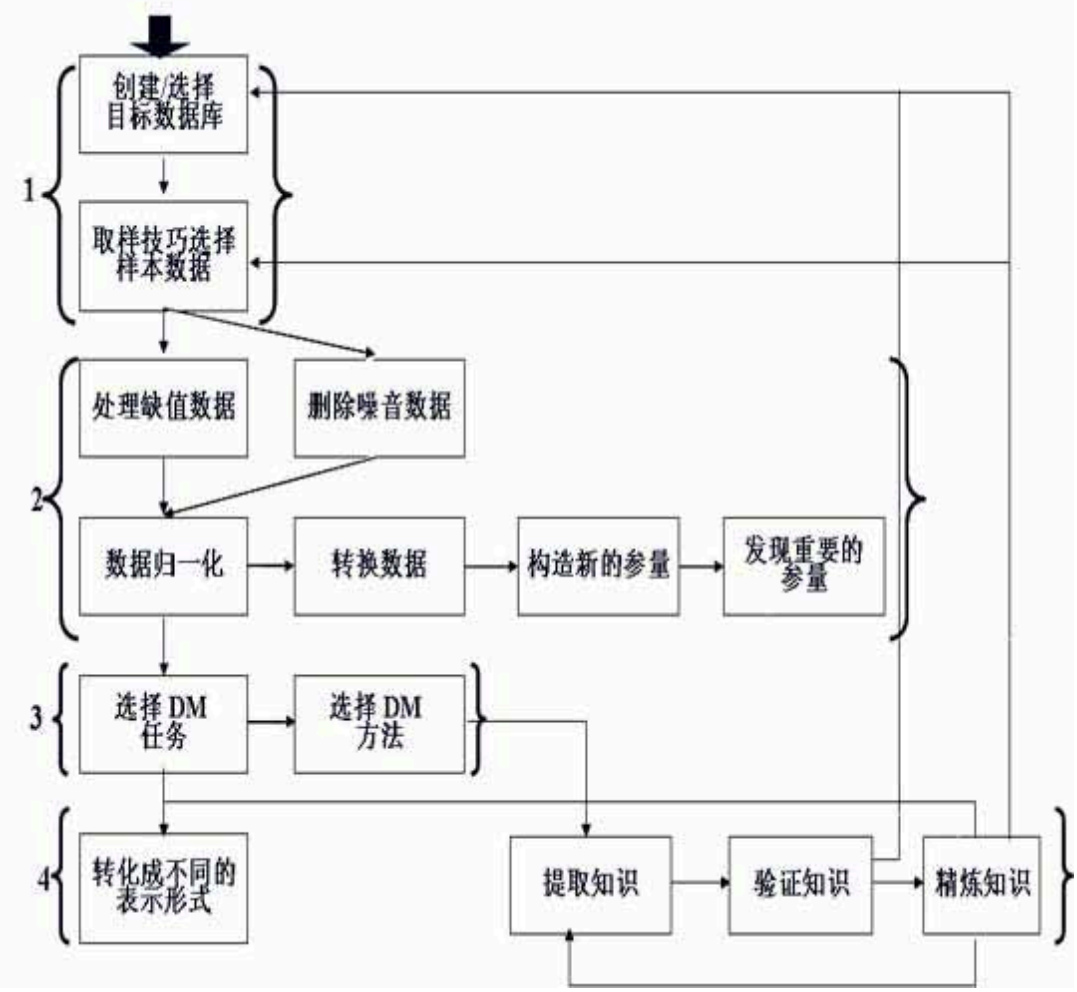


图 5.1: 知识发现的过程和步骤。

述了知识发现的四个步骤，即：数据选择、数据预处理、数据挖掘、结果解释与

评估。

（1）在数据选择阶段，搜索所有与挖掘对象有关的内部和外部数据信息，并从中选择一个数据集或在多数据集的子集上聚集，挑出适合于数据挖掘应用的数据。

（2）数据预处理是指去除噪声或无关数据，去除空白数据域，考虑时间顺序和数据变化等。研究数据的质量，为进一步的分析做准备，并确定将要进行的挖掘操作类型。在数据预处理阶段还包括数据转换。找到数据的特征表示，用维变化或转化方法减少有效变量的数目或找到数据的不变式。将数据转化成一个分析模型，这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

（3）在数据挖掘阶段要对所得到的经过转化的数据进行挖掘。用KDD过程中的准则，选择某个特定数据挖掘算法（如汇总、分类、回归、聚类等）用于搜索数据中的模型。除了完善从选择合适的挖掘算法外，其余一切工作都能自动地完成。然后搜索或产生一个特定的感兴趣的模型或一个特定的数据集。

（4）结果的解释与评估是指解释某个发现的模式，去掉多余的不切题意的模式，转化成某个有用的模式，以使用户明白。其使用的分析方法一般应由数据挖掘操作而定，通常会用到可视化技术。将分析所得到的知识集成到业务信息系统的组织结构中去，获得这些知识的作用或证明这些知识。用预先、可信的知识检查和解决知识中可能的矛盾。

数据挖掘是虚拟天文台重要的组成部分。虚拟天文台不仅要给用户提供大的数据资源、快速的数据查找服务，还会为天文学家提供多种多样的工具和软件包。VO_IMPAT和红移测量工具就是基于这种目的而开发的。接下来我们将在原有的简单的红移测量工具基础上集成更多种测光红移算法，该工具的流程图如5.2所示。

在此工具中，我们将提供不同波段数据的选择和预测测光红移算法的选择。用户可以根据自己课题的需要选取数据。例如：选取单波段数据或多波段数据。如果用户选择了多波段数据，首先需要对数据进行交叉证认。该工具可以集成交叉证认工具来实现多波段数据的融合。当然，用户也可以上传自己的数据，利用该工具所提供的算法来预测测光红移。在数据预处理阶段，可参照图5.1所示的数据预处理步骤进行。对数据进行“净化处理”，包括去除离散数据，对数据进行归一化，去噪声，对数据形式进行转化等。根据数据的特点选择合适的测光

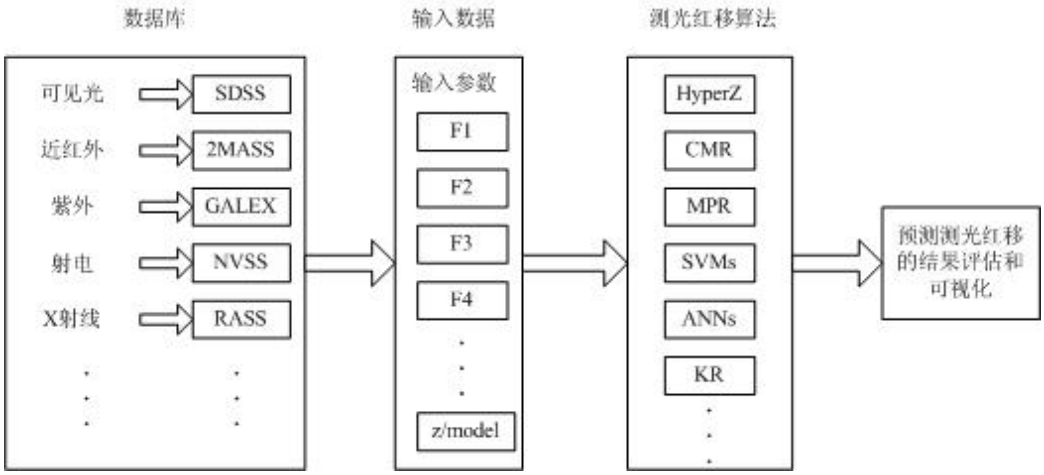


图 5.2: 测光红移工具流程图。

红移算法。对于既想得到测光红移又想得到星系类型的用户，可以选择模板匹配方法。如果要得到高预测精度的红移值，则可以选择ANNs或KR。对高红移样本，最好选择模板匹配或贝叶斯方法。最后可以对预测结果进行分析与可视化（如：测光红移与光谱红移对比图、红移分布的直方图等）。另外，此测光红移工具还具有一定的拓展性，不仅可以用于预测星系和类星体的测光红移，而且还可以根据用户需求预测恒星的物理参量，如温度、表面重力加速度、金属丰度等。随着各种工具的开发、集成和应用，虚拟天文台的功能将越来越完善。各种工具的自动化将给天文界提供更优质的服务。随着虚拟天文台的成长和壮大，它将成为天文学家从事科研工作得心应手的好帮手，也将成为整个天文界不可或缺的一环。

参考文献

[1] Gwyn, S. D., Photometric Redshifts of Galaxies. [master thesis], University of Victoria, Canada, 1995.

[2] Puschell, J. J., Owen, F. N., Laing, R. A. Near-infrared photometry of distant radio galaxies - Spectral flux distributions and redshifts estimates. Astrophysical Journal. 1982, 257, L57-L61.

[3] Loh, E. D., Spillar, E. J. Photometric redshifts of galaxies. Astrophysical Journal. 1986, 303, L54-L61.

[4] Baum W. A. Photoelectric determinations of redshifts beyond 0.2 c. Astronomical Journal. 1957, 62, 6 - 7.

[5] Baum W. A. Photoelectric Magnitudes and Red-Shifts. Proceedings from IAU Symposium no. 15, Macmillan Press, New York, 1962, p.390.

[6] Koo D. C. Optical multicolors - A poor person's Z machine for galaxies. Astronomical Journal. 1985, 90, 418 - 440.

[7] Bruzual A. G. Spectral evolution of galaxies. I - Early-type systems. Astrophysical Journal. 1983, 273, 105 - 107

[8] Cowie, L. L., Lilly, S. J., Gardner, J., McLean, I. S. A cosmologically siginificant population of galaxies dominated by very young star formation. Astrophysical Journal. 1988, 332, L29 - L32.

[9] Guhathakurta, P., Tyson, J. A., Majewski, S. R. A redshift limit for the faint blue galaxy population from deep U band imaging. Astrophysical Journal. 1990, 357, L9 - L12.

[10] Steidel,C. C., Hamilton D. Deep imaging of high redshift QSO fields below the Lyman limit. II - Number counts and colors of field galaxies. Astrophysical Journal. 1993, 105, 2017 - 2030.

[11] Puschell, J. J., Owen F. N., Laing, R. A. Near-infrared photometry of distant radio galaxies - Spectral flux distributions and redshift estimates. Astrophysical Journal. 1982, 257, L57 - L61.

[12] Guiderdoni, B. Photometric redshifts and evolution of distant galaxies. High redshift and primeval galaxies. Proceedings of the Third IAP Workshop, Paris, France. 1987, L271 - L278.

[13] Koo, D. C. Multicolor photometry of field galaxies to B = 24. Astrophysical Journal. 1986, 311, L651 - L679.

[14] Lilly, S. J., Cowie, L. L., Gardner, J. P. A deep imaging and spectroscopic survey of faint galaxies. Astrophysical Journal 1991, 369, L79 - L105.

[15] Subbarao M. U., Connolly, A. J., Szalay, A. S., Koo, D. C. Luminosity Functions From Photometric Redshifts. I. Techniques. Astrophysical Journal. 1996, 112, L929.

[16] Connolly, A. J., Szalay, A. S., Koo D., et al. Superclustering at Redshift Z = 0.54. Astrophysical Journal. 1996, 473, L67.

[17] Koo, D. C. Multicolor photometry of the red cluster 0016+16 at Z = 0.54. Astrophysical Journal. 1981, 251, L75 - L79.

[18] Fontana, A., DOdorico, S., Poli, F., et al., Photometric Redshifts and Selection of High-Redshift Galaxies in the NTT and Hubble Deep Fields. The Astronomical Journal. 2000, 120, 2206- 2219.

[19] Poli, F., Giallongo, E., Menci, N., et al., The Evolution of the Galaxy Sizes in the New Technology Telescope Deep Field: A Comparison with Cold Dark Matter Models. The Astronomical Journal. 1999, 527, 662-672.

[20] Giallongo, E., Menci, N., Poli, F., et al., Comparing the Evolution of the Galaxy Disk Sizes with Cold Dark Matter Models: The Hubble Deep Field. The Astrophysical Journal. 2000, 530, 73-76.

[21] Blake, C., Collister, A., Bridle, S., et al., Cosmological baryonic and matter densities from 600000 SDSS luminous red galaxies with photometric redshifts. Monthly Notices of the Royal Astronomical Society. 2007, 374, 1527-1548.

[22] Padmanabhan, N., White, M., Eisenstein, D. J., A robust estimator of the small-scale galaxy correlation function. Monthly Notices of the Royal Astronomical Society. 2007, 376, 1702-1706.

[23] York D. G., et al. The Sloan Digital Sky Survey: Technical Summary. The Astronomical Journal. 2000, 120, 1579-1587.

[24] Le Fèvre, et al. The Messenger. 2003, 111, 18.

[25] David M., et al. Science Objectives and Early Results of the DEEP2 Redshift Survey. Procedding of SPIE. 2003, 4834, 161.

[26] Bruzual A. G., Charlot S. Spectral evolution of stellar populations using isochrone synthesis. Astrophysical Journal. 1993, 405, 538-553.

[27] Coleman, G. D., Wu, C. C., Weedman, D. W. Colors and magnitudes predicted for high redshift galaxies. Astrophysical Journal Supplement Series. 1980, 43, 393-416.

[28] Brunner, R. J., Connolly, A. J., Szalazy, A. S., Toward More Precise Photometric Redshifts: Calibration Via CCD Photometry. Astrophysical Journal. 1997, 482, 21.

[29] Wang, Y., Bahcall, N., Turner, E. L., A Catalog of Color-based Redshift Estimates for $Z < 4$ Galaxies in the Hubble Deep Field. The Astronomical Journal. 1998, 116, 2081-2085.

[30] Budavari, T., et al., The Ultraviolet Luminosity Function of GALEX Galaxies at Photometric Redshifts between 0.07 and 0.25. The Astrophysical Journal. 2005, 619, 31-34.

[31] Firth, A. E., Lahav, O., Somerville, R. S., Estimating photometric redshifts with artificial neural networks. Monthly Notice of the Royal Astronomical Society. 2003, 339, 1195-1202.

[32] Tagliaferri, R., Lecture Notes in Computer Science. 2003, 2859, 226.

[33] Brunner, R. J., et. al Toward More Precise Photometric Redshifts: Calibration Via CCD Photometry. Astrophysical Journal. 1997, 482, L21.

[34] Wadadekar, Y. Estimating Photometric Redshifts Using Support Vector Machines. The Publications of the Astronomical Society of the Pacific. 2005, 117, 79-85.

[35] Yang, Y. B., Yuan, Q. R., Wen, L., Detecting Clusters of Galaxies in SDSS: I Photometric redshifts of Galaxies. The Astronomy Journal. 2005, submitted.

[36] Wang, D. Software kits for measuring photometric redshifts. Proceeding of SPIE. 2006, 6274, 13.

[37] Connolly, A. J., et. al Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry. Astronomical Journal. 1995, 110, 2655.

[38] Csabai, I., et al. The Application of Photometric Redshifts to the SDSS Early Data Release. The Astronomical Journal 2003, 125, 580.

[39] Benitez, N. Bayesian Photometric Redshift Estimation. The Astrophysical Journal. 2000, 536, 571.

[40] Vapnik, V. N. The Nature of Statistical Learning Theory. New York: SpringerVerlag. 1995.

[41] Wang D., Zhang Y. Z., Liu C., Zhao Y. H., Two novel approaches for photometric redshift estimation based on SDSS and 2MASS databases. Chinese Journal of Astronomy & Astrophysics. 2007, accepted.

[42] Ball, N. M., et al. Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. Monthly Notices of the Royal Astronomical Society. 2004, 348, 1038-1046.

[43] Collister, A. A., Lahav, O. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. The Publications of the Astronomical Society of the Pacific. 2004, 116, 345-351.

[44] Vanzella, E., et al Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS. Astronomy and Astrophysics. 2004, 423, 761-776.

[45] The Sloan Digital Sky Survey, http://www.sdss.org/.

[46] Apache Point Observatory, http://www.apo.nmsu.edu/.

[47] Stoughton, C., Lupton, R. H., Bernardi, M., et al., Sloan Digital Sky Survey: Early Data Release. Astronomical Journal, 2002, 123, 485-548.

[48] Abazajian, K., Adelman-McCarthy, J. K., Agueros, M. A., et al., The First Data Release of the Sloan Digital Sky Survey. Astronomical Journal, 2003, 126, 2081-2086.

[49] Abazajian, K., Adelman-McCarthy, J. K., Agueros, M. A., et al., The Second Data Release of the Sloan Digital Sky Survey. Astronomical Journal, 2004, 128, 502-512.

[50] Abazajian, K., Adelman-McCarthy, J. K., Agueros, M. A., et al., The Third Data Release of the Sloan Digital Sky Survey. Astronomical Journal, 2005, 129, 1755-1759.

[51] Adelman-McCarthy, J. K., Agueros, M. A., Allam, S. S., et al., The Fourth Data Release of the Sloan Digital Sky Survey. Astrophysical Journal Supplement Series, 2006, 162, 38-48.

[52] Adelman-McCarthy, J. K., Agueros, M. A., Allam, S. S., et al., The Fifth Data Release of the Sloan Digital Sky Survey. submitted to Astrophysical Journal Supplement Series, 2007.

[53] Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al., Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. Astronomical Journal, 2002, 124, 1810-1824.

[54] Eisenstein, D. J., Annis, J., Gunn, J. E., et al., Spectroscopic Target Selection for the Sloan Digital Sky Survey: The Luminous Red Galaxy Sample. Astronomical Journal, 2001, 122, 2267-2280.

[55] Richards, G. T., Fan, X. H., Newberg, H. J., et al., Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample. Astronomical Journal, 2002, 123, 2945-2975.

[56] Lupton, R. H., Gunn, J. E., Szalay, A. S. A Modified Magnitude System that Produces Well-Behaved Magnitudes, Colors, and Errors Even for Low Signal-to-Noise Ratio Measurements. Astronomical Journal, 1999, 118, 1406-1410.

[57] Petrosian, V. Surface brightness and evolution of galaxies. Astrophysical Journal, 1976, 209, L1-L5.

[58] Schlegel, D. J., Finkbeiner, D. P., Davis, M. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. The Astrophysical Journal, 1998, 500, 525.

[59] 高建云, 陈力, 王家骥, 侯金良, 赵君亮, 2MASS的科学意义和成果概览. 天文学进展, 2004, 22, 4.

[60] 张彦霞, 多波段天体物理中的自动分类方法研究. [博士学位论文], 北京：中国科学院国家天文台, 2003.

[61] 张健楠, 恒星光谱大气物理参量自动估计研究. [博士学位论文], 北京：中国科学院自动化研究所, 2006.

[62] Yang, Y. B., Yuan, Q. R., Wen, Z. L., et al., Detecting Clusters of Galaxies in SDSS: I. Photometric Redshifts of Galaxies. The Astrophysical Journal, submitted.

[63] Hsieh, B. C., Yee, H. K. C., Lin, H., Gladders, M. D. A Photometric Redshift Galaxy Catalog from the Red-Sequence Cluster Survey. The Astrophysical Journal Supplement Series, 2005, 158, 161-177.

[64] Fukugita, M., Ichikawa, T., Gunn, J. E., et al., The Sloan Digital Sky Survey Photometric System. Astronomical Journal, 1996, 111, 1748.

[65] Sowards-Emmerd, D., Smith, J. A., McKay, T. A., et al., A Catalog of Photometry for Las Campanas Redshift Survey Galaxies on the Sloan Digital Sky Survey System. Astronomical Journal, 2000, 119, 2598-2604.

[66] 叶阿忠, 非参数计量经济学. 南开大学出版社, 2005.

[67] Suchkov, A. A., Hanisch, R. J., Margon, B., A Census of Object Types and Redshift Estimates in the SDSS Photometric Catalog from a Trained Decision Tree Classifier. Astronomical Journal, 2005, 130, 2439-2452.

[68] Strateva, I., Ivezic, Z., Knapp, G. R., et al. Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data The Astronomical Journal, 2001, 122, 1861-1874

[69] Deng K., Moore A., Proceedings of the Twelfth International Joint Conference. on Artificial Intelligence San Francisco: Morgan Kaufmann, 1995, 1233.

[70] VO_IMPAT, http://services.china-vo.org/vo_impat/.

[71] The Digitized Sky Survey, http://archive.stsci.edu/dss/index.html.

[72] The United States Naval Observatory, http://ftp.nofs.navy.mil /projects/pmm/catalogs.html.

[73] The Large Sky Area Multi-Object Fiber Spectroscopic Telescope, http://www.lamost.org/en/.

[74] The National Radio Astronomical Observatory, Very Large Array, Sky Survey, http://www.cv.nrao.edu/nvss/.

[75] A Five hundred meter Aperture Spherical Telescope, http://www.bao.ac.cn/bao/LT/.

[76] The Two Micron All-Sky Survey, http://www.ipac.caltech.edu/2mass/gallery/.

[77] The NASA Infrared Processing and Analysis Center Infrared Science Archive, http://www-astro.physics.ox.ac.uk/ wjs/pscz.html.

[78] The ROSAT X-Ray All-Sky Survey, http://www.xray.mpe.mpg.de/cgi-bin/rosat/rosat-survey.

[79] 何香涛, 陈阳, 李丹丹等，多波段巡天和LAMOST观测目标. 天文学进展, 2004, 22, 3.

[80] 崔辰州, 中国虚拟天文台系统设计. [博士学位论文], 北京：中国科学院国家天文台, 2003.

[81] The National Virtual Observatory, http://www.us-vo.org/.

[82] Virtual Observatory of China, http://www.china-vo.org.

[83] International Virtual Observatory Alliance, http://www.ivoa.net/.

[84] VOPLOT, http://vo.iucaa.ernet.in/ voi/voplot.htm.

[85] Skymouse, http://skymouse.lamost.org/.

[86] fv, http://heasarc.nasa.gov/lheasoft/ftools/fv/.

[87] FTOOLS, http://heasarc.gsfc.nasa.gov/ftools/.

[88] FITSIL, http://heasarc.gsfc.nasa.gov/fitsio/fitsio.html.

[89] ALADIN, http://aladin.u-strasbg.fr/.

[90] DS9, http://hea-www.harvard.edu/RD/ds9/.

[91] SAOImage, http://tdc-www.harvard.edu/software/saoimage.html.

[92] Beijing Astronomical Data Center, http://badc.lamost.org/cnweb/.

[93] The Digitized Sky Survey , http://archive.stsci.edu/dss/index.html.

发表文章目录

[1] **Wang Dan**, Zhang Yan-Xia , Liu Chao, Zhao, Yong-Heng. Kernel Regression for Determining Photometric Redshifts From Sloan Broadband Photometry. 2007, MNRAS, in press.

[2] **Wang Dan**, Zhang, Yan-Xia, Liu, Chao, Zhao, Yong-Heng. Two Novel Approaches for Photometric Redshift Estimation Based on SDSS and 2MASS Databases. 2007, ChJAA, in press.

[3] **Wang Dan**, Zhang, Yan-Xia, Cui, Chen-Zhou, Zhao, Yong-Heng. Software Kits for Measuring Photometric Redshifts. 2006, Proceeding of SPIE, Vol.6274, 13.

[4] **Wang Dan**, Zhang, Yan-Xia, Zhao, Yong-Heng. Survey Methods for Photometric Redshifts. 2006, Proceeding of ADASS XVI, in press.

[5] **Wang Dan**, Zhao, Yong-heng. VO_IMPAT: Image Processing and Analysis Toolkit for the Virtual Observatory of China, 2006, Journal of Astronomical Research & Technology, Vol.3, 295-303.

[6] **Wang Dan**, Zhang, Yan-Xia, Liu, Chao, Zhao, Yong-Heng. Estimating Photometric Redshifts With Multivariable Polynomial Regression. 2006, Advances in Space Research, submitted.

[7] **Wang Dan**, Zhang, Yan-xia, Zhao, Yong-heng. The Methods of Estimating Photometric Redshifts. 2007, Progress in Astronomy, submitted.

[8] Liu, Chao, **Wang Dan**, Liu, Bo, Gao, Dan, Cui, Chen-Zhou, Zhao, Yong-Heng. An Astronomical Data Mining Application Framework for Virtual Observatory. 2006, Proceeding of SPIE, Vol.6274, 15.

致　　谢

论文收笔之际，回忆起在科学院的求学生活，我感慨万千。回想与大家在一起的时时刻刻，点点滴滴，感慨、感激之情难以言表。

首先，由衷感谢我的恩师赵永恒研究员。赵老师知识渊博，待人诚恳，兢兢业业，对学术前沿有敏锐的洞察力，让我钦佩不已。初见导师之时，我不甚惊讶如此年轻竟然已经在学术上取得了许多骄人成绩。有幸师从赵老师，是我人生道路上的一个机遇。自此，每当我在科研的道路上遇见困难，难以超越，困惑不解之际，赵老师总能快速找出我的问题关键所在，并提出切实的解决方案和具体的建议，使我顿时就有“山重水复疑无路，柳暗花明又一村”的感觉。整篇论文的字字句句都凝集着赵老师的汗水和心血。赵老师以其宽广的胸襟，谦虚诚恳地待人方式及对科研工作的执著和脚踏实地的作风深深地影响着我，催我不断进步，赵老师还以宽泛、和谐、民主的方法引导着我步入科学的殿堂。三年转眼即逝，我将顺利完成我的学业，可赵老师已添了丝丝白发。在此，郑重地道声“赵老师，您辛苦了!”，并表达我对恩师由衷的感恩和敬意。

在课题研究过程中，还时常得到张彦霞副研究员具体而细致的指导，张老师以其丰富的实践经验和丰厚的理论基础，为我提供了及时的帮助，常常针对具体问题与我探讨至深夜；在论文撰写过程中，也得到了张老师的全面指教，我撰写的每篇论文都凝聚了她点点滴滴的汗水，从文章的构架、逻辑和语言的修饰上都浸入了张老师的智慧；在日常生活中更是得到了张老师的诸多鼓励和热心帮助，她就像大姐姐一样给予我无私的关怀和无限的快乐，在三年博士求学生活即将结束之际，我发自内心的道一声“谢谢”!

感谢国家天文台赵刚研究员。正是赵刚老师引领我步入天文台的大门。他的仁者风范和平易近人的态度感染和影响着我。

感谢国家天文台的胡景耀研究员、武向平研究员、邓李才研究员、魏建彦研究员、周旭研究员、马骏研究员、彭勃研究员、吴宏研究员、邹振隆研究员，虽然与各位老师的接触不多，但是他们严谨的治学态度值得我学习。

感谢北京大学的吴学兵教授，北京师范大学的何香涛教授、李宗伟教授和姜碧沩教授，中科院研究生院的邓祖淦教授，他们对我的课题给予了很多帮助，

与他们的讨论让我受益匪浅。

感谢研究生办公室的杜红荣老师和艾华老师。正是她们的无私奉献和对学生工作的热情，让我们有种家的感觉。感谢赵景芝老师的热心帮助。

感谢崔辰州博士三年来的关心和指导。感谢LAMOST罗阿理博士的热心帮助和鼓励。同时还要感谢石火明、张昊彤、张健楠、贾磊、陈建军和施建荣的关心。

感谢虚拟天文台的刘超同学。虽然刘超晚我半年入学，但是我油然地叫他“师兄”。刘超同学数学功底扎实，程序经验丰富，科学研究不气馁，生活中乐于助人，在此感谢师兄对我的大力支持。

感谢LAMOST项目总部的老师：苏洪钧老师, 王钢老师、陈英老师、袁晖老师、孙盛慈老师、曹淑蕴老师、李颀老师、门力老师、冯磊老师的关心和帮助。

感谢LAMOST324实验室的兄弟姐妹们，深厚的同窗情谊和活跃的实验室文化让我永生难忘。大家的“开心果”郑征、田海俊，湖南“辣妹子”杨阳，“帅哥”孙华平、罗宇，憨厚踏实的老乡尹红星，“天文活字典”胡晨、杨帆，“心理专家”高丹、路勇，以及我们心目中永远的“超兄”刘超。同时还要感谢吴悦、李丽丽、孙世卫、薛元、邹思成、王淑青、朱黎楠。

感谢天文台的好友陈杰、刘玉娟、刘国卿、夏芳、李会贤。

感谢在国外求学的吴潮博士和杨雁宾博士。

感谢山东大学威海分校的韩圣浩校长、梁作堂老师、管立老师、张鹏老师、杨田林老师、胡启萍老师、吴爱玲老师、王爱芳老师、胡绍明同学。在威海工作学习一年，他们给予我很大的帮助和鼓励。

谨将此文献给我最挚爱的父母。谢谢父母三十年的养育之恩，正是他们这份无私的爱，以及对我的信任和支持，才有了我今天的自信和成绩。此刻，道出那句埋藏于心底多年的话：爸妈，我爱您们！同时也将此文献给远在新西兰的干爹干妈，感谢他们对异国他乡的我给予的关爱和照顾。

最后，我要将此文献给已逝去的亲人和朋友，还有鼓励我考博士，陪伴我复习的知心朋友。

谨祝大家：平安、健康、幸福！