第 46 卷 第 4 期
2005 年 10 月

天 文 学 报
ACTA ASTRONOMICA SINICA

Vol.46 No.4
Oct., 2005

恒星大气物理参量的非参数估计方法*

张健楠 [1,2] 吴福朝 [1] 罗阿理 [2] 赵永恒 [2]

(1 中国科学院自动化研究所模式识别国家重点实验室，北京 2728 信箱 100080)
(2 中国科学院国家天文台 北京 100012)

摘要 恒星大气物理参量 (有效温度、表面重力、化学丰度) 是导致恒星光谱差异的主要因素. 恒星大气物理参量的自动测量是 LAMOST 等大规模巡天望远镜所产生的海量天体光谱数据自动处理中一个重要研究内容. 针对测量大样本的恒星光谱数据估计每个恒星的大气物理参量，提出了一种基于变窗宽核函数的估计算法：变窗宽算法是对固定窗宽算法的改进，分为 3 个步骤： (1) 将历史恒星光谱数据进行 PCA 处理，得到光谱的低维特征数据； (2) 利用特征数据与其物理参数的对应关系，建立一种变窗宽的非参数估计模型； (3) 利用该估计模型，直接计算待测恒星光谱的 3 个物理参量 (有效温度、表面重力、金属丰度). 实验结果表明：该方法与固定窗宽估计模型以及在其他文献中报道的方法相比，具有较高的估计精度和鲁棒性.

关键词 **方法：数据分析，方法：统计，恒星：基本参数，** PCA, **非参数估计，窗宽**
中图分类号： P 399; **文献标识码：** A

1 引言

恒星物理参量的自动测量是 LAMOST 海量天体光谱数据自动处理中一个重要研究内容 [1]. 恒星光谱的差异是由其有效温度、重力加速度与化学丰度等物理参量所决定. 传统测量方法要求准确地提取连续谱和计算谱线的等值宽度和深度：如测量表面有效温度的帕邢连续谱斜率测量法，谱线系列的边界跳变法等，测量表面重力的分光测定法，以及测量作为压力指示物的连续谱、氢线，某些强谱线等 [2,3]. MK 恒星光谱分类系统只粗略地确定恒星表面有效温度 [4], 其光度级与表面重力成反比.

随着 LAMOST 、SDSS 、2DF 等大规模巡天望远镜的计划与实施，模式识别与人工智能技术已应用到海量天体光谱数据自动处理中. 近年来，在研究恒星物理参量的自动测量方面，主要集中在神经网络算法与最近邻算法. Bailer-Jones[5,6] 在研究 GAIA 计划将产生的中低分辨率的恒星光谱的数据分析中，采用神经网络算法对模拟光谱的物理参量进行计算，训练与测试所采用的是 Kurucz[7] 的恒星大气模型模拟合成光谱数据，温度范围为 4000—30000 K, 表面重力 $\log g$ 为 2.0—5.0 dex, 化学丰度 [M/H] 为 −3.0—+1.0 dex, 其测试结果在较低分辨率和不同信噪比下均得到很高的精度，例如估计精度在分辨率为 2.5 nm, SNR 为 20 时 $\log \mathrm{Teff} = 0.0033$, $\log g = 0.182$, [M/H]=0.145. Snider 等 [8] 构造

2004-11-10 收到原稿， 2005-06-29 收到修改稿
* 国家 863 计划项目 (2003AA133060) 资助

的后馈神经网络对实测的 F 、 G 型恒星光谱 (参量范围： Teff=4250—6500 K, $\log g$ = 1.0—5.0 dex, [M/H]=−4.0—+0.3 dex) 估计精度为 σ(Teff)=135—150 K, $\sigma(\log g)$ = 0.25—0.30 dex, σ([M/H])=0.15—0.20 dex. Soubiran 等 [9], Katz 等 [10] 对分辨率为 0.1 nm 的 ELODIE 恒星光谱建立了一个由 211 个恒星光谱构成的模板库，采用最近邻法对 Teff 为 4000—6300 K, [M/H] 为 −2.9—+0.35 的恒星光谱确定物理参量，当精度在 SNR=100 时，Teff=86 K, $\log g$ = 0.28 dex, [M/H]=0.16 dex; 当 SNR=10 时， Teff=102 K, $\log g$ = 0.29 dex, [M/H]=0.17 dex.

我们曾采用基于固定窗宽的核函数，建立核估计非参数回归模型对恒星表面有效温度进行估计 [11], 取得较好地估计精度和鲁棒性，不足之处在于窗宽大小是以经验值探试方法确定的，经验值大小与模型样本点的分布有关，当模型样本点的分布改变时，窗宽大小需要重新进行探试确定. 本文采用变窗宽的核估计非参数回归模型，除了对恒星的有效温度 Teff 进行估计外，对重力 $\log g$ 和金属丰度 [M/H] 也进行了估计. 该方法根据 PCA 处理后的历史光谱数据与其物理参量的对应关系建立核函数估计模型，核函数的窗宽随样本点的分布密度而发生变化，由模型本身以及最近邻法确定. 该模型中仅需调节的参数是平滑因子，它的取值范围通常为 [0.1,1.2]. 实验结果表明该模型与固定窗宽估计模型相比，温度估计的精度和鲁棒性更高，模型参数的调节简单. 同时对另外两个参数的估计也超过了前述文献方法给出的精度.

2 物理参量估计模型与算法

2.1 恒星光谱的 PCA 分析

主分量分析 (Principal Component Analysis,PCA) 的基本原理是要找到一种空间变换方式，让标准化后的原始变量线性组合成若干个矢量，要求它们之间相互正交，且第一个矢量能反映样本间自变量的最大差异，其他矢量所反映的差异程度依次降低. 这些矢量称之为主分量矢量，而样本在主分量矢量上的投影为样本的主分量. 这样，由主分量我们可以得到样本的特征矢量 (特征数据), 以较少数量的特征对样本进行描述，从而达到降低原始数据的维数. 该方法在天文信号处理上得到成功应用：例如， Storrie-Lombardi[12] 和 Bailer-Jones[13] 分别在 PCA 的基础上使用神经网络方法对恒星进行分类； Folkes 等 [14] 用 PCA 做 2dF 红移巡天的光谱分类； Darren 等 [15] 将 PCA 用于 DEEP2 红移巡天. 为减少计算量，本文采用 PCA 方法对恒星光谱数据降维，应用光谱的 PCA 特征数据建立恒星参数的非参数估计模型.

2.2 基于变窗宽的非参数回归模型

设天体的光谱特征数据样本为 X, Y 为 X 的某个物理参数. $m(x)$ 为 Y 对 X 的回归函数，在非参数回归中，通常使用权函数法估计回归函数 $m(x)$[16]. 设：

$$W_{ni} = W_{ni}(x) = W_{ni}(x; X_1, \cdots, X_n), (i = 1, 2, \cdots, n), \tag{1}$$

是选定的 n 个依赖于 x 和 $X_1, \cdots, X_n$ 函数，回归函数 $m(x)$ 的估计为：

$$m_n(x) = \sum_{i=1}^{n} W_{ni} Y_i \,, \tag{2}$$

$m_n(x)$ 为回归函数 $m(x)$ 的权函数估计， W_{ni} 为权函数. 权函数满足自然条件:

$$W_{ni}(x; X_1, \cdots, X_n) \geq 0, \sum_{i=1}^{n} W_{ni}(x; X_1, \cdots, X_n) = 1 \,, \tag{3}$$

这样的权函数称为概率权函数.

在文 [11] 中，我们曾采用固定窗宽的核估计非参数回归模型进行恒星表面有效温度的估计，权函数采用核函数法确定，以经验值探试的方法确定窗宽，实验结果表明窗宽的选取决定了回归函数的估计效果，而且不同类型的样本显示出不同的估计效果和实验中采用统一的窗宽有关. 因此，采用基于变窗宽核估计的非参数回归模型，能够更有效地估计恒星物理参数. 在本文中，我们采用由 k 近邻法所建立的变窗宽核估计的非参数回归模型 [17], 对恒星物理参量进行有效估计.

令测试样本 x 的 k 个近邻的模型样本为 $X_1, \cdots, X_k$,

$$r_k(x) = \max_{1 \leq i \leq k} \|x - X_i\|_2 \,,$$

$r_k(x)$ 即测试样本 x 距 k 个邻近模型样本的最大距离，取总体密度 $f(x)$ 的核估计为:

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_k(x)} K\Big(\frac{x - X_i}{r_k(x)}\Big) \,, \tag{4}$$

其中 $K(\cdot)$ 为定义在 R^d 上的核函数，在本文中 K 为 Gaussian 函数. 由密度函数核估计 (4) 式，定义权函数:

$$W_{ni}(x; X_1, \cdots, X_n) = K\Big(\frac{x - X_i}{h_i}\Big) \Big/ \sum_{j=1}^{n} K\Big(\frac{x - X_j}{h_i}\Big) \,, \tag{5}$$

这样，就得到了回归函数 $m(x)$ 的权函数估计:

$$m_n(x) = \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_i}\Big) Y_i \Big/ \sum_{j=1}^{n} K\Big(\frac{x - X_j}{h_i}\Big) \,, \tag{6}$$

其中 $h_i = \lambda * r_k(x)$, 为可变窗宽， $\lambda > 0$, 为平滑因子，是密度函数 $f_n(x)$ 整体的平滑程度的影响因子，当 k 值固定时 $f_n(x)$ 的整体平滑程度决定于平滑因子 λ. 因此，样本点 x 处的核宽与 $r_k(x)$ 成比例，对应样本点稀疏的邻域有较平坦的核函数；当 h_i 为固定常数时，即为固定窗宽的估计模型.

2.3 变窗宽非参数估计的算法

(1) 对 n 条历史恒星光谱数据进行 PCA 分析，得到恒星光谱的特征矢量，记为 X_i. 将特征矢量与相应的物理参数配对，得到历史光谱的特征样本 (X_i, Y_i), $(i = 1, 2, \cdots, n)$. 对实际观测恒星光谱 X, 进行 PCA 投影得到特征矢量 x.

(2) 计算测试样本 x 到模型样本中 k 个近邻样本的最大距离 $r_k(x)$, 并确定对应窗宽： $h_i = \lambda * r_k(x)$, 其中 λ 在下述范围内取值：温度估计为 $0.5 > \lambda > 0.1$; 表面重力为 $1.0 > \lambda > 0.6$; 化学丰度为 $1.1 > \lambda > 0.7$.

(3) 选择 Gaussian 函数 $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$ 作为核函数，构造物理参数的变窗宽估计模型：

$$m_n(x) = \sum_{i=1}^{n} W_{ni}Y_i = \frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{1}{2}\left\|\frac{x-X_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n} \exp\left(-\frac{1}{2}\left\|\frac{x-X_i}{h_i}\right\|^2\right)}. \tag{7}$$

3 实验结果与分析

3.1 实验过程

实验采用恒星大气模拟模型合成光谱库 Lejeune97 中的光谱数据 [18]. 选取库中 7 个光谱集共 3198 条光谱，表面有效温度 Teff: 2000—50000 K; 表面重力加速度 log g: −1.0—6.0 dex; 化学丰度 [M/H]: −0.3—+0.3 dex. 所有样本的截取波段均为 380—780 nm, 并将光谱分辨率插值为 1 nm. 3198 条光谱分为两部分， 1599 条光谱做训练集构造非参数回归算法模型，另外 1599 条光谱作测试样本，它们分别涵盖了 3 种物理参量范围内所有光谱. 为检验模型的抗噪性，每条光谱分别加信噪比 SNR 为 5 、 10 、 15 、 20 、 30 的高斯白噪声.

实验过程：

• 模型光谱的 PCA 分析

在实验中采用 PCA 处理后的前 3 个主分量构成的光谱特征数据，其方差贡献率之和大于 98%.

• 构造估计模型

按 (7) 式构建恒星物理参量的估计模型，其中样本的温度参量采用其实际温度的对数，即 $Y_i = \log_{10}(\text{Teff})$, 模型输出值 $m_n(x)$ 也是温度的对数值.

• 误差计算

误差采用绝对误差，其中温度的误差计算为：设第 i 个估计样本的实际温度为 T, 模型估计值为 m, 该样本的估计误差为 $\varepsilon_i = |\log_{10}T - m|$, 则平均绝对误差 $\varepsilon = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$.

3.2 实验结果

采用两种模型：变窗宽估计模型和固定窗宽估计模型对测试样本集分别进行物理参量的估计，其中变窗宽估计模型中 λ 取值 0.1, 近邻个数 $k = 1$, 固定窗宽的非参数估计模型中窗宽采用文 [11] 的经验值探视取 0.001. 两种方法对表面温度、表面重力和化学丰度估计的平均绝对误差分别见表 1 、表 2, 两种方法的对比曲线见图 1 的 (a) 、 (b) 和 (c).

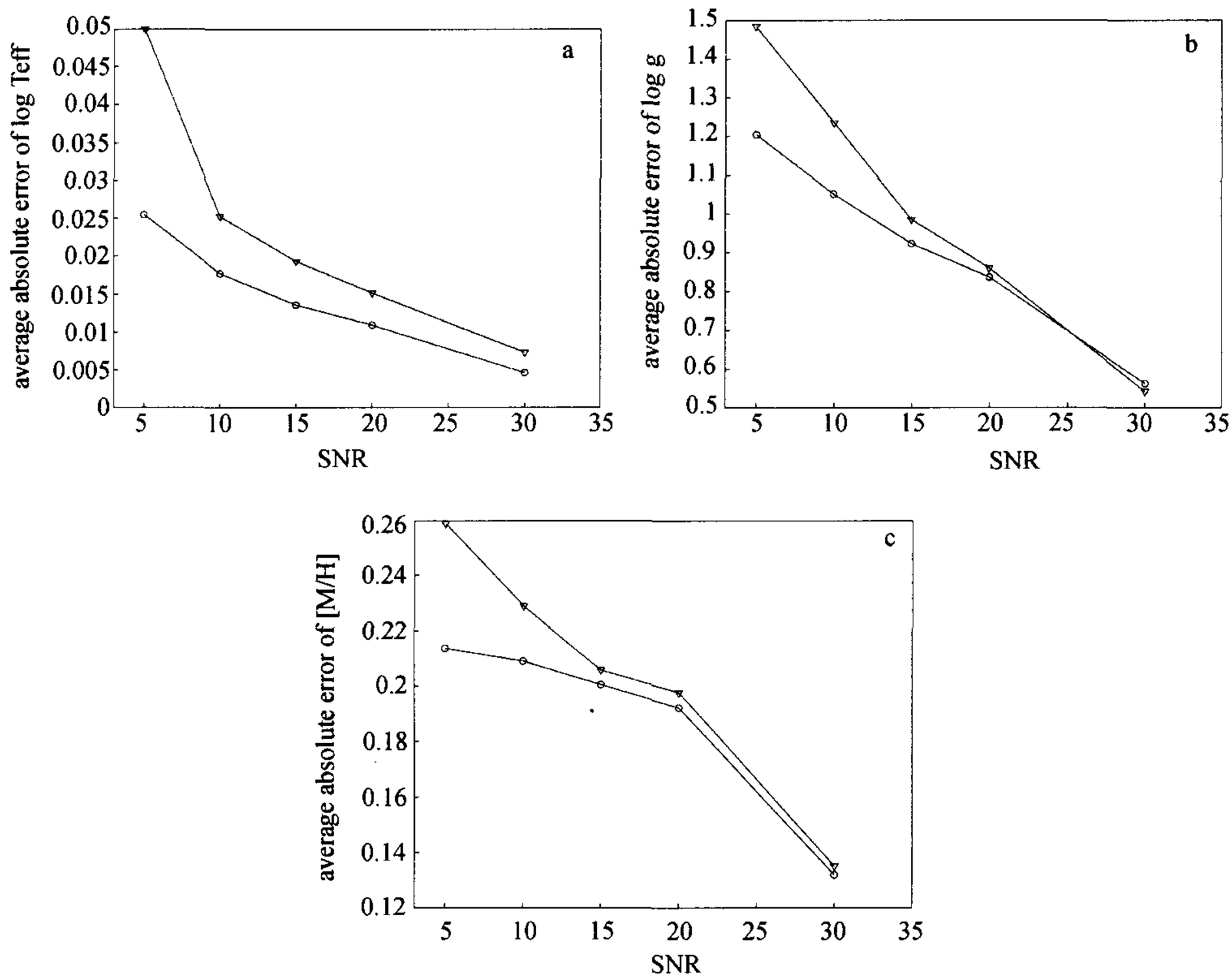

图 1 固定窗宽 (三角标记) 与变窗宽 (圆圈标记) 的误差曲线对比图: (a) 表面温度, (b) 表面重力, (c) 化学丰度

Fig. 1 The error curves from the two estimators (line with triangle:fixed window-width; line with circle: variable window-width): (a) log Teff; (b) log g; (c) [M/H]

表 1 变窗宽估计模型计算测试样本恒星物理参量的平均绝对误差 (SNR— 信噪比)

Table 1 The average absolute errors of the estimated fundamental parameters from non-parameter estimator with variable window-width (SNR—the signal-to-noise ratio)

parameter	SNR=30	SNR=20	SNR=15	SNR=10	SNR=5
log Teff	0.0048	0.0111	0.0137	0.0178	0.0255
log g	0.5632	0.8374	0.9237	1.0524	1.2053
[M/H]	0.1323	0.1924	0.2007	0.2092	0.2138

表 2 固定窗宽非参数估计模型计算测试样本恒星物理参量的平均绝对误差 (SNR— 信噪比)

Table 2 The average absolute errors of the estimated fundamental parameters from non-parameter estimator with fixed window-width (SNR—the signal-to-noise ratio)

parameter	SNR=30	SNR=20	SNR=15	SNR=10	SNR=5
log Teff	0.0075	0.0153	0.0194	0.0253	0.0500
log g	0.5429	0.8610	0.9863	1.2360	1.4850
[M/H]	0.1355	0.1979	0.2060	0.2290	0.2589

对变窗宽非参数估计模型，本文作了进一步实验：将测试样本按实际温度划分为不同的类型， O: ≥30000 K; B: 30000 K—10000 K; AF: 10000 K—5800 K; GK: 5800 K—4000 K; M: <4000 K, 计算各种类型所估计的参数值的平均绝对误差. 实验结果见表 3 至表 5 与图 2 至图 4.

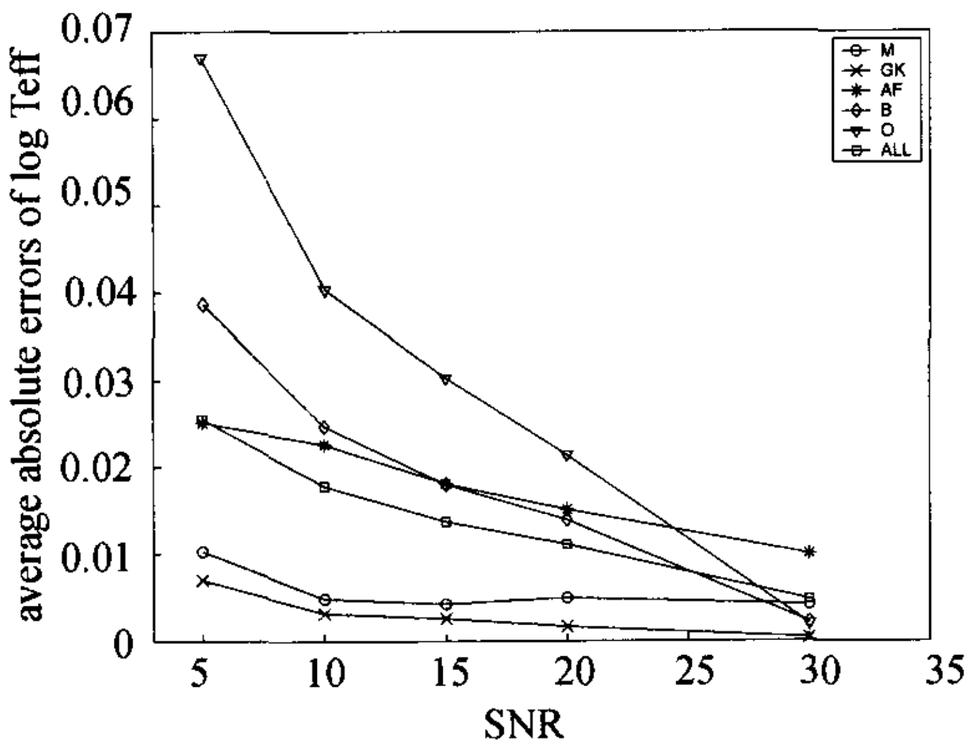

图 2 M 、 GK 、 AF 、 B 、 O 的温度平均绝对误差曲线图

Fig. 2 The average absolute errors of log Teff for M, GK, AF, B, O and the whole samples at different SNR

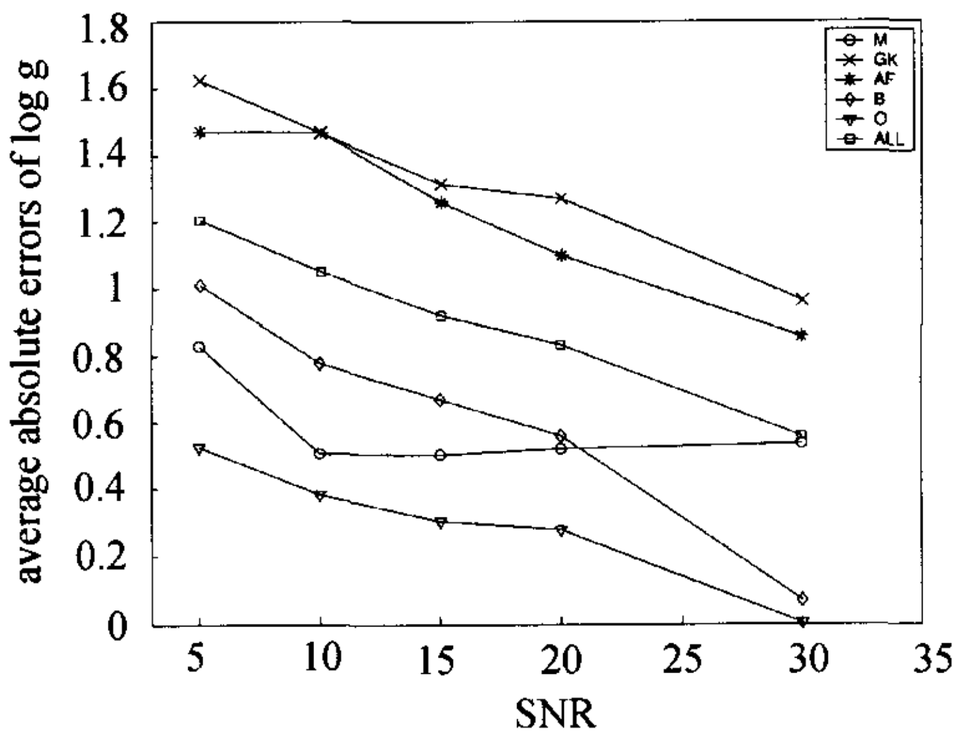

图 3 M 、 GK 、 AF 、 B 、 O 的表面重力平均绝对误差曲线图

Fig. 3 The average absolute errors of log g for M, GK, AF, B, O and the whole samples at different SNR

为了测试平滑因子 λ 对模型估计精度的影响，在信噪比 SNR=30 时，我们计算了 3 种物理参量在不同平滑因子 λ 的平均绝对误差，图 5 为 λ 在 0.07—1.4 区间内取值时的误差变化曲线：其中 $\lambda > 0.1$ 时，步长为 0.1; $\lambda < 0.1$ 时，步长为 0.01.

为了测试近邻个数 k 对模型估计精度的影响以及它的最优取值，我们进行了如下实验：在信噪比 SNR=30 时，我们将 λ 固定为 0.5 和 0.8 (分别对应于图 5 中温度误差曲线最低点，重力与化学丰度误差曲线最低点)，计算了不同 k 取值下 3 种物理参量的平均绝对误差. 图 6 为 k 由 1 增加至 30, $\lambda = 0.5$ 时温度的平均绝对误差曲线，与 $\lambda = 0.8$ 时表面重力与化学丰度平均绝对误差曲线.

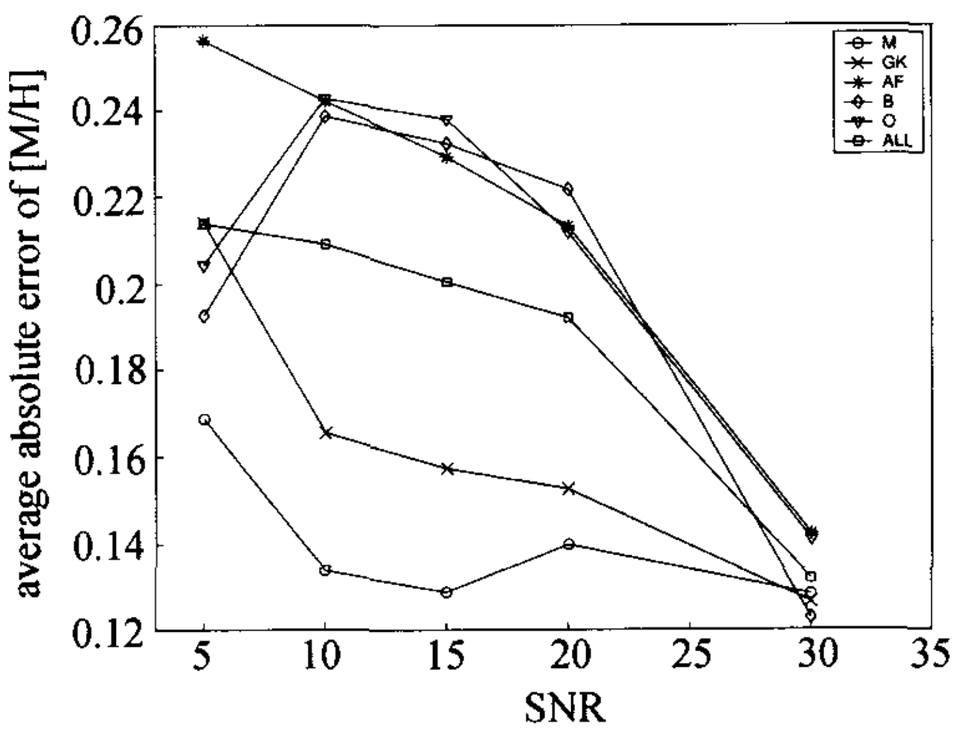

图 4 M 、 GK 、 AF 、 B 、 O 的化学丰度平均绝对误差曲线图

Fig. 4 The average absolute errors of [M/H] for M, GK, AF, B, O and the whole samples at different SNR

表 3 变窗宽非参数估计模型估计恒星表面温度的平均绝对误差 (SNR— 信噪比)

Table 3 The average absolute errors of the estimated log Teff from non-parameter estimator with variable window-width (SNR—the signal-to-noise ratio)

testing subset	number of sample	SNR=30	SNR=20	SNR=15	SNR=10	SNR=5
M	284	0.0042	0.0050	0.0043	0.0049	0.0103
GK	270	0.0004	0.0017	0.0026	0.0032	0.0070
AF	527	0.0100	0.0151	0.0181	0.0226	0.0251
B	422	0.0022	0.0139	0.0180	0.0247	0.0387
O	96	0.0020	0.0214	0.0303	0.0404	0.0670

表 4 变窗宽非参数估计模型估计恒星表面重力的平均绝对误差 (SNR— 信噪比))

Table 4 The average absolute errors of the estimated log g from non-parameter estimator with variable window-width (SNR—the signal-to-noise ratio)

testing subset	number of sample	SNR=30	SNR=20	SNR=15	SNR=10	SNR=5
M	284	0.5414	0.5256	0.5069	0.5091	0.8307
GK	270	0.9677	1.2727	1.3162	1.4688	1.6240
AF	527	0.8612	1.1020	1.2611	1.4707	1.4716
B	422	0.0736	0.5643	0.6720	0.7809	1.0112
O	96	0.0052	0.2830	0.3073	0.3858	0.5271

表 5 变窗宽非参数估计模型估计恒星化学丰度的平均绝对误差 (SNR— 信噪比)

Table 5 The average absolute errors of the estimated [M/H]from non-parameter estimator with variable window-width (SNR—the signal-to-noise ratio)

testing subset	number of sample	SNR=30	SNR=20	SNR=15	SNR=10	SNR=5
M	284	0.1286	0.1402	0.1290	0.1340	0.1687
GK	270	0.1268	0.1530	0.1576	0.1656	0.2141
AF	527	0.1427	0.2134	0.2293	0.2423	0.2564
B	422	0.1231	0.2219	0.2324	0.2387	0.1927
O	96	0.1414	0.2120	0.2381	0.2429	0.2044

3.3 实验分析

(1) 实验结果表明核函数窗宽的取值影响估计效果，采用变核宽的非参数估计算法比固定核宽的非参数估计算法提高了恒星物理参量的估计精度与鲁棒性. 尤其是对表面有效温度的估计，具有更好的估计精度与抗噪性.

(2) 从图 5 可见，当平滑因子 $1.0 > \lambda > 0.1$ 时，温度误差曲线较平缓，而表面重力与化学丰度误差曲线呈现为平缓的凹形. 3 种物理量的最优估计分别对应于不同的 λ, 温度： $\lambda = 0.5$; 表面重力： $\lambda = 0.8$; 化学丰度： $\lambda = 0.9$.

(3) 从图 6 可以看出，随着近邻样本数 k 的增大， 3 种物理参量的平均估计误差均随之增大. 当 $k = 1$ 时，估计误差最小. 因此，建立估计模型时可以直接将 k 固定为 1, 从而仅需调整 λ 的值.

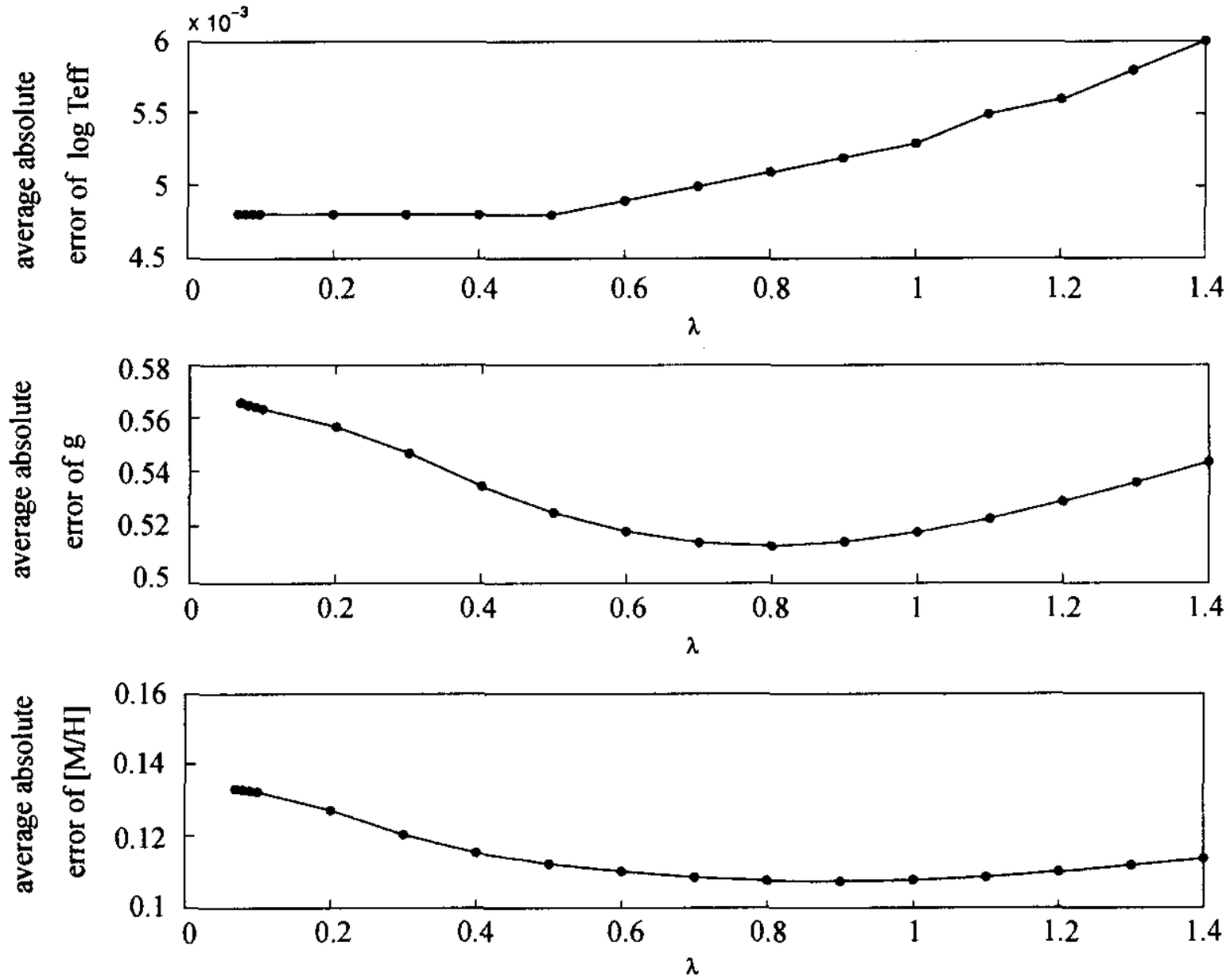

图 5　3 种物理量在不同平滑因子 λ 下的估计误差曲线图

Fig. 5　The error curves of log Teff, log g and [M/H] at different λ

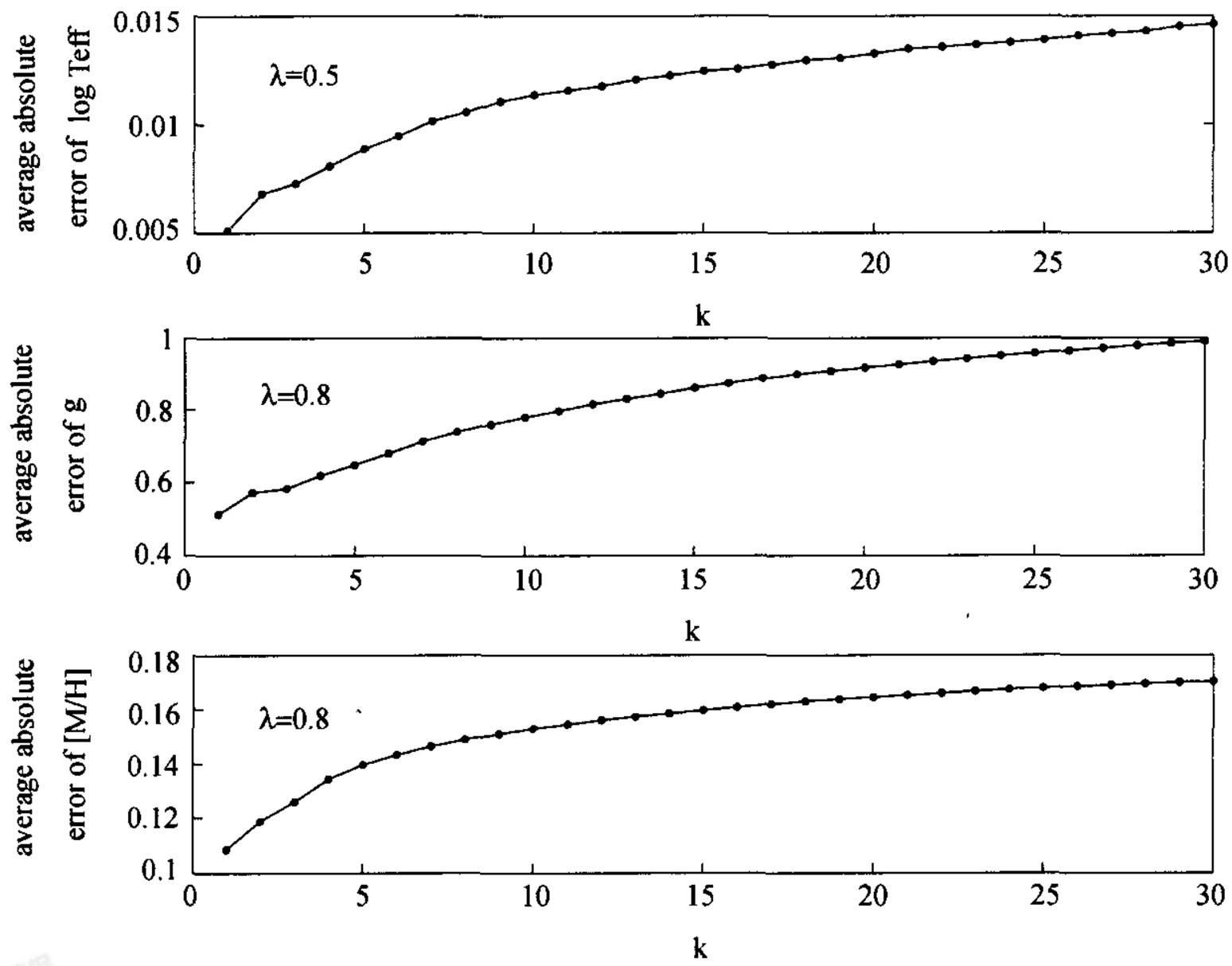

图 6　3 种物理量在不同近邻样本数 k 下的估计误差曲线图

Fig. 6　The error curves of log Teff, log g and [M/H] at different k

(4) 3 种物理参量的估计效果各有差异，表面有效温度的估计效果最好，化学丰度的估计效果次之，表面重力估计效果最差，这与采用 PCA 处理后的恒星光谱作为模型输入量有关.

(5) 本文所提出的方法与文献中现有方法所报道的结果相比，除表面重力外无论是温度还是化学丰度均具有较高的估计精度和抗噪性. 更重要的是本文的估计模型较其他方法相比，模型参数的选择更为简单，因而更容易实现.

4 结束语

本文采用非参数估计方法，为恒星光谱的大样本数据的物理参量 (表面有效温度、表面重力、化学丰度) 的估计，建立了一种基于变窗宽的非参数估计模型. 确定这种模型的关键在于选择近邻样本点的个数与平滑因子，对此本文在大量实验的基础上讨论了它们的取值范围. 实验结果表明，本文的模型比基于固定窗宽的估计模型具有更好地估计精度和鲁棒性，此外对于表面有效温度和化学丰度的估计与现有其他模型相比也具有较高的估计精度. 目前的方法还局限在估计简化的恒星大气模型产生的光谱的参数，而实际光谱由于受到观测和仪器的影响 (不仅仅是信噪比) 产生了各种畸变，而恒星的演化关系决定了文中估计的这 3 个参数之间存在着联系. 只有考虑了影响光谱的各种因素和演化模型，对大样本恒星光谱的物理参数估计才能完全自动化. 本文的工作为达到这一目标打下了一个很好的基础，下一步我们将在此基础上针对实际光谱进一步改进算法，以期完全解决这一困扰大样本巡天光谱数据的自动分析问题.

参 考 文 献

1 LAMOST 项目可行性研究报告. http://www.lamost.org/feasible.htm

2 黄润乾. 恒星物理. 北京：科学出版社， 1998: 8

3 Gray D F, 著. 黄磷，李宗伟，蒋世仰等译， The Observation and Analysis of Stellar Photospheres(恒星光球的观测和分析). 北京：科学出版社， 1981: 383

4 Kurtz M J. Progress in Automation Techniques for MK Classification. In: F Garrison, ed. The MK Process and Stellar Classification. 1984: 136

5 Bailer-Jones C A L. Ap&SS, 2002, 280: 21

6 Bailer-Jones C A L. A&A, 2000, 357: 197

7 Kurucz R L, In: Barbuy B, Renzini A, ed. Stellar populations of galaxies. Kluwer, Dordrecht, 1992 255

8 Snider S, Prieto C A, Hippel TV, et al. ApJ, 2001, 562: 528

9 Soubiran C, Katz D, Cayrel R. A&AS, 1998, 133: 221

10 Katz D, Soubiran C, Cayrel R, et al. A&A, 1998, 338: 151

11 Jiannan ZHANG, Fuchao WU, Ali LUO, et al. Non-parameter estimation algorithm to determine stellar effective temperature accepted by Spectroscopy and Spectral Analysis

12 Storrie-Lombardi M C, Irwin M J, von Hippel T, et al. Vistas in Astronomy, 1994, 38(3): 331

13 Bailer-Jones C A L, Mike Irwin, Ted von Hippel. arXiv:astro-ph/9803050,5 Mar 1998

14 Simon Folkes, Shai Ronen, et al. MNRAS, 1999, 308: 459

15 Darren S Madgwick, Alison L Coil. arXiv:astro-ph/0305587 v2 12 Sep 2003

16 陈希孺，方兆本，李国英. 非参数统计教程. 上海：科学技术出版社， 1989, 281

17 Silverman B W. Density estimation for statistics and data analysis. Published in Monographs on Statistics and Applied Probability. London: Chapman and Hall, 1986

18 Lejeune T, Cuisinier F, Buser R. A standard stellar library (Lejeune+ 1997). http://vizier.u-strasbg.fr/viz-bin/ftp-index?J/A+AS/125/229

Non-parameter Estimation Algorithm to Determine Stellar Fundamental Parameters

ZHANG Jian-nan[1,2] WU Fu-chao[1] LUO A-li[2] ZHAO Yong-heng[2]

(1 *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

(2 *National Astronomical Observatories, Chinese Academy of Sciences, Beijing* 100012)

Abstract The three fundamental parameters of stellar atmosphere, which are the effective temperature, the surface gravity, and the metallic abundance, determine the continuum and spectral lines in the stellar spectrum. A new algorithm is proposed for estimating the three parameters in this paper. It is composed of three steps: (1) the sample data of stellar spectrum, whose three fundamental parameters are known, is reduced to a lower dimensional feature space using PCA (Pnincipal Component Analysis) technique; (2) a non-parameter estimator with variable window-width is set up from the correspondence between the feature data and their fundamental parameters data; (3) this estimator is used to compute the fundamental parameters of a stellar spectrum. Experiments indicate that the estimator is more efficient and robust than the fixed window-width estimator.

Key words Methods: data analysis, Methods: statistical, Stars: fundamental parameters, PCA, Non-parameter Estimation, Window-width