A型恒星光谱线指数岭回归有效温度的预测分析

薛仁政[1]，陈淑鑫[1*]，黄宏本[2]

1. 齐齐哈尔大学计算机与控制工程学院，黑龙江 齐齐哈尔 161006
2. 梧州学院大数据与软件工程学院，广西 梧州 543002

摘 要 天文光谱线指数数据能够较好地保留着恒星的物理特征信息，为此借助线指数特征数据构建多参数模型，有利于更好地回归分析数据的共变关系及谱线的内在规律。世界上光谱获取率最高的施密特天文望远镜LAMOST发布的观测光谱都已经过标记，利用天文可视化工具分析这些标记的恒星光谱线指数会产生预测因子自相关，多元线性回归时因变量存在共线性，导致方差较大、得到最小二乘回归系数不稳定，虽不影响使用回归的有效性，但较难从回归方程中得到独立预测因子的评估系数。利用LAMOST巡天光谱数据中A型恒星Lick线指数为数据源，选取有效温度 $T_{\rm eff}$ 为7 000～8 500 K，取信噪比大于50的光谱特征值实现回归分析恒星参数 $T_{\rm eff}$ 值，经箱线图呈现DR5星表中，A型光谱86 097条具备 $T_{\rm eff}$ 值大样本光谱数据的整体分布，统计分析26种线指数的特征值后，选取分布相似且带宽为12 Å的kp12，halpha12和hgamma12字段，减少解释线指数变量的数目，优化冗余变量方差膨胀因子(VIF)系数。实验选取两两变量间观测数据集，局部拟合回归散点、同样的数据源使用散点图的总体轮廓生成高密度散点图，利用色差透明性突出显示数据密集区域。结果表明多元线性回归和岭回归算法都能从低分辨率光谱中确定A型恒星的有效温度，但经过共线性数据分析有偏估计实验，使用岭回归分析寻找最佳模型，能更准确地确定恒星有效温度，进而得到预测A型恒星有效温度及谱线回归特性。

关键词 恒星光谱；LAMOST；岭回归；线性模型；Lick线指数

中图分类号：P145.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)08-2624-06

引 言

2008年10月16日作为世界上光谱获取率最高的施密特天文望远镜LAMOST投入使用，增强了我国在国际天文研究领域巡天观测的地位[1]，提升了我国大视场天文学及大数据光学光谱观测研究方面的科研水平。天文大数据中蕴含着海量天体光谱信息[2]，研究者们通过定义光谱线指数来描述光谱的特征，其中Lick线指数的应用最为广泛[3]，已有研究者利用Lick指数对LAMOST光谱分析，例如，2015年国家天文台刘超利用LAMOST星表中线指数分析MK恒星分类CaII K特征值之间的分布[4]，2016年潘景昌等提出利用线指数特征对LAMOST DR2数据中F，G，K和M型恒星光谱聚类分析研究[5]，并基于SVM输入光谱线指数完成恒星分类等。本文通过分析LAMOST已发布的A型光谱线指数，利用多元线性回归算法分析实现估计A型恒星的有效温度。实验选取温度值7 000～8 500 K，信噪比 S/N 大于50的A型恒星数据，经线性拟合分析，最后利用岭回归方法构建共线性数据分析有偏估计回归模型，解决过拟合问题，得出一种预测LAMOST大样本实测光谱有效温度的回归方法。

1 天文光谱线指数

天文光谱线指数值在天文研究领域已取得诸多成果，线指数表示天文光谱中物理特征的数值，保留着多种类型的参数特征数据，1994年Guy Worthey等给出Lick线指数的完整定义及描述[6]，光谱线指数的数值定义光谱中特征谱线的积分星等特征数值、谱线等值宽度(EW)以及半高全宽(FWHM)的光谱线指数组合。

1.1 LAMOST线指数

收稿日期：2019-02-24，**修订日期**：2019-05-16

基金项目：国家自然科学基金项目(U1631239)，国家自然科学基金青年科学基金项目(11803013)，黑龙江省教育厅基本业务专项项目(135109248)，齐齐哈尔市科技计划项目(GYGG-201720)

作者简介：薛仁政，1979年生，齐齐哈尔大学讲师 e-mail：27744950@qq.com

*通讯联系人 e-mail：shuxinfriend@126.com

光谱数据是天体物理学研究的基础和证认依据，我国国家天文台运行着大天区多目标光纤光谱望远镜(LAMOST)截止到 2018 年 7 月，LAMOST 已经积累了六年的巡天数据(http://dr5.lamost.org/)，DR5 数据集共获得9 017 844个光谱。LAMOST 巡天光谱数据按 MK 分类标准系统进行光谱型分类，波长覆盖范围从 3 690～9 100 Å，步长为 1 Å(总采样点数 N=5 491)，分辨率为 1 800，在用模板光谱来自约 100 万条的大量先导巡天实测恒星光谱数据。

LAMOST 发布 DR5 数据 v1 版中 A 型恒星提供的光度类型比 DR1 目录中包含了更多的线指数信息，DR5 星表中共计439 914条 A 型光谱，其中86 097条光谱数据具备 $T_{\rm eff}$ 值，如图 1 所示经箱线图呈现大样本之间的不同，反映线指数统计量整体分布，从 26 种线指数特征值中选取分布相似，且带宽为 12 Å 的 kp12，halpha12 和 hgamma12 字段，减少解释变量的数目，增加方差膨胀因子(VIF)系数，在第 3.1 节分析 VIF 冗余变量获得更好的预测效果。后文实验选取信噪比 S/N 大于 50，且温度在 7 000～8 500 K 范围的 A 型恒星数据线性拟合分析恒星大气物理参数的有效温度值。

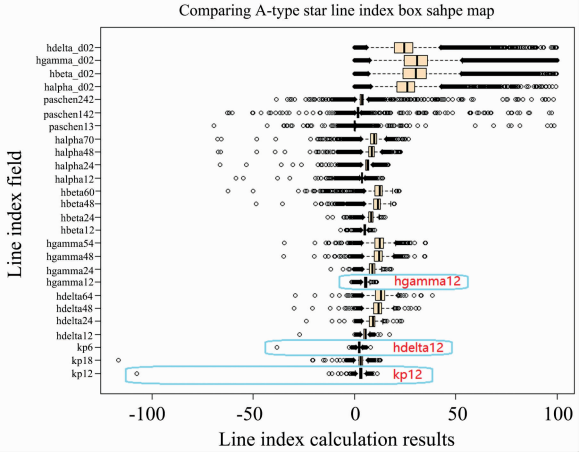

图 1　A 型恒星线指数 26 种特征值分析箱线图

Fig. 1　Boxplot analysis of 26 eigenvalues of A-type stellar line index

1.2　构造数据模型

LAMOST 发布的观测光谱都已经过标记，先前研究所构建的回归模型大部分都是假定自变量和因变量之间呈线性关系[7]，对于任何回归问题的预测因子都可能产生自相关，虽然并不影响回归使用的有效性，但很难或者不可能从回归方程中得到独立预测因子的评估系数。后文提出的方案包括以下步骤：首先，利用天文可视化工具对 LAMOST 线指数数据统计分析；其次，用 Lick 线指数对 $T_{\rm eff}$ 测量进行多元线性回归；最后，采用岭回归寻找最佳模型，得到多元线性回归训练预测的模型。结果表明多元线性回归和岭回归算法都能准确地从低分辨率光谱中确定 A 型恒星的有效温度。

2　线性模型分析

多个不同的解释变量显示相似的变量信息时，可能导致方差非常大，使估计准确性变差，需要解决变量间的共线性问题。当线指数的特征变量和恒星参数呈非线性关系时，需保留线指数的多个类型的参数数据，本节结合响应变量与解释变量之间的关系，用散点图表示，并进行多线性回归分析，较好地解释变量相互关联性问题。

2.1　谱线多重共线性

多重共线性分析可定量解释模型中包含的多个变量函数，基于 A 型恒星参数建立的回归模型能够有效预测 $T_{\rm eff}$ 数据特征之间相关方法，与典型线性回归不同，使用多重线性回归来实现分析 Lick 线指数与 $T_{\rm eff}$ 之间的关系，特别是连续光谱中存在着校准和消光等较多的不确定性因素，后文运用预测方法有效地利用谱线指数从天文光谱中提取 $T_{\rm eff}$ 特征。多元线性回归方程模型如式(1)所示

$$T_{{\rm eff},i}=\beta_0+\beta_1X_{i,1}+\beta_2X_{i,2}+\cdots+\beta_pX_{i,p}+\varepsilon_i \quad (1)$$

式(1)中，i=1,2,…,N；回归误差 ε_i；方差 σ^2；预测因素的数量级为 p；每个独立变量的值 X_{ip}；N 是测试数据 $N(0,\sigma^2)$，满足 $E(\varepsilon_i)=0$ 和 $\mathrm{Var}(\varepsilon_i\mid X)=\sigma^2$，预测因子系数 β_1，…，β_{p-1}，β_p 常数项 β_0 是估计与最小二乘方法。利用拟合函数能执行完整的线性模型分析，输出值与最小二乘估计 β_1 和 β_0 值如式(2)所示。

$$S(\beta)=\sum_{i=1}^{n}(T_{{\rm eff},i}-\beta_0-\beta_1X_{i1}-\beta_2X_{i2}-\cdots-\beta_pX_{i2p})^2=\|T_{\rm eff}-\chi^{\alpha}\|^2 \quad (2)$$

2.2　线性拟合观测数据

依据 1.1 节分析结果，当 LAMOST 观测样本量较大，所绘制数据点非常集中时，很多数据点重合叠加，不利于直观展示数据的局部规律和趋势以及线指数特征值之间的相关性特征，本文实验选取相应比例的局部数据集拟合回归。实验将观测数据两两变量间以散点呈现在二维平面的数据点分布，如图 2—图 4 所示被分析量恒星有效温度 $T_{\rm eff}$ 与线指数之间相关关系。实验中用模型回归线与观测数据的拟合程度来表示因变量与所有自变量之间的总体关系，经函数拟合回归曲线如图 2(a)、图 3(a)和图 4(a)数据点重叠集中，分别包含蓝色线、绿色线和红色线显示线性回归趋势。由于数据点的重叠使得因变量和自变量之间的关系难以识别，不利于直观地显示观察变量之间的相关特征，同样的数据源使用统计透明性如图 2(b)、图 3(b)和图 4(b)所示任意坐标上重叠点的数量，使用散点图的总体轮廓生成高密度散点图，利用色差突出显示数据密集区域，将不同 Counts 数据点分箱，用灰度深浅表示箱中数据点的个数，明晰散点图的整体轮廓，数据的散点映射表示核密度估计。该函数自动在一定范围内设置数据点，显示数据点被划分成几个框，灰色的数据用来表示框中数据点的数量。

从图 2—图 4 中散点分布趋势显示 $T_{\rm eff}$ 与 kp12，hdelta12 和 hgamma12 变量之间的负线性相关性是非常明显的，如表 1 所示两两变量间所得到协方差矩阵为对称矩阵，表中计算各列的方差值，其中以主对角线为对称轴对应相等的矩阵，列出的运行结果可得因变量可变性的百分比，后续章节利用回归方程误差度量线性模型反映拟合程度真实关系，后文岭回归预测模型中协同因子是最关键的相关关系。

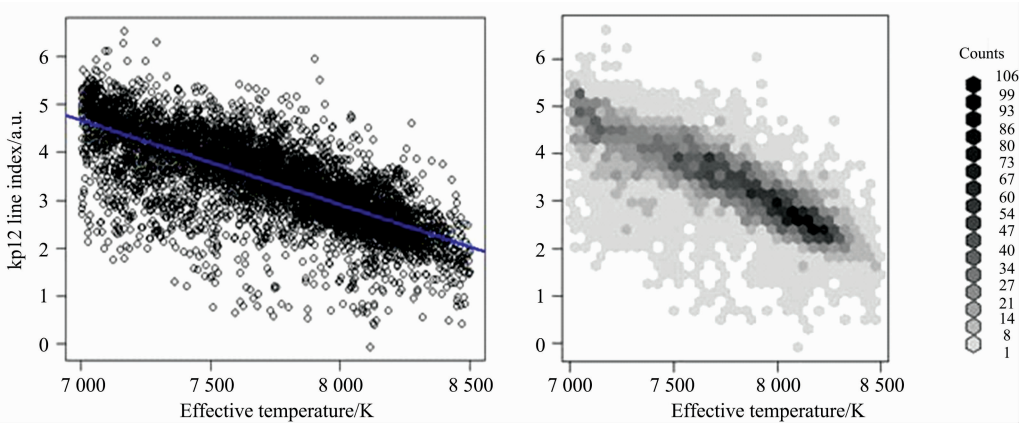

图 2 A 型恒星有效温度 T_{eff} 与 kp12 线指数分析

(a)：线性回归散点图(蓝色)；(b)：高密度散点图

Fig. 2 A-type stellar effective temperature T_{eff} and kp12 line indices

(a)：Scatter plot with linear regression (in blue)；(b)：High density scatter plot

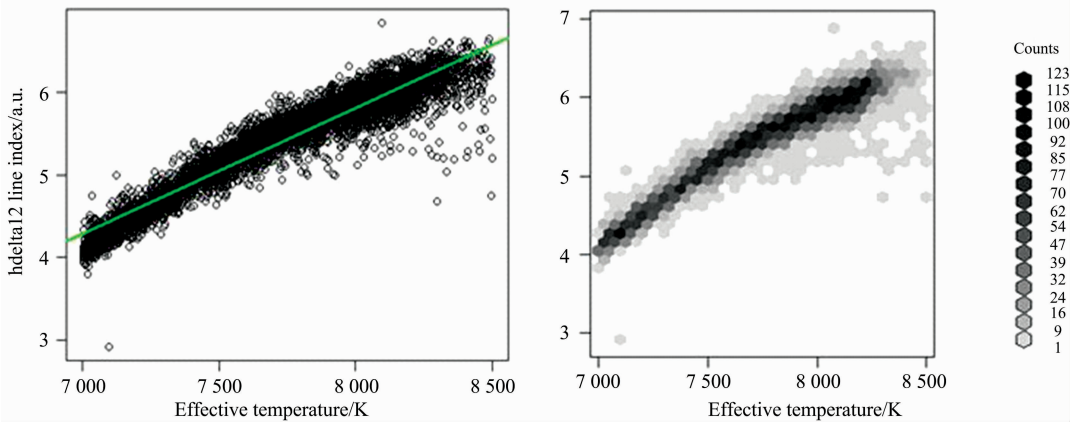

图 3 A 型恒星有效温度 T_{eff} 与 hdelta12 线指数分析

(a)：线性回归散点图(绿色)；(b)：高密度散点图

Fig. 3 A-type stellar effective temperature T_{eff} and hdelta12 line indices

(a)：Scatter plot with linear regression (in green)；(b)：High density scatter plot

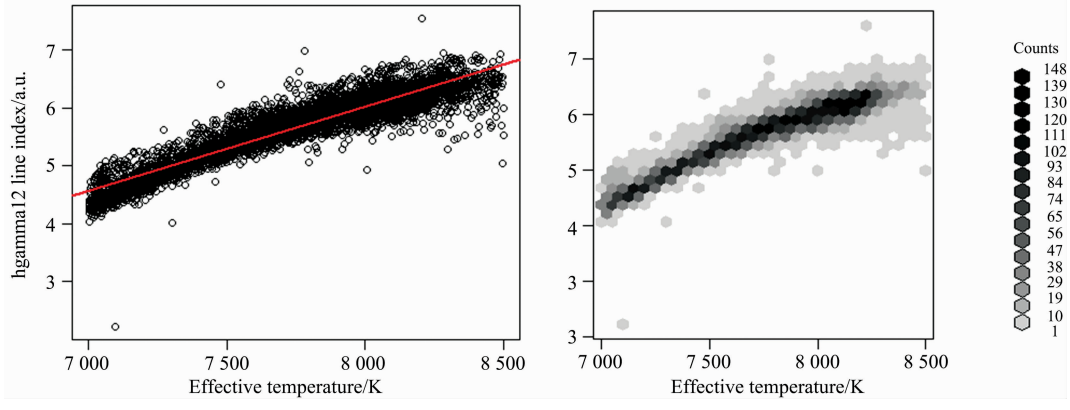

图 4 A 型恒星有效温度 T_{eff} 与 hgamma12 线指数分析

(a)：线性回归散点图(红色)；(b)：高密度散点图

Fig. 4 A-type stellar effective temperature T_{eff} and hgamma12 line indices

(a)：Scatter plot with linear regression (in red)；(b)：High density scatter plot

表 1　线指数特征值与 T_{eff} 参数线性相关系数值

Table 1　Linear correlation coefficient between line index eigenvalues and T_{eff}

相关系数	T_{eff}	kp12	hdelta12	hgamma12
T_{eff}	1.000	0.743	0.954	0.936
kp12	0.743	1.000	−0.633	−0.601
hdelta12	0.954	−0.633	1.000	0.981
hgamma12	0.936	−0.601	0.981	1.000

3　预测大气参数

基于上述对 LAMOST 观测数据模型的分析，建立多线性回归预测恒星参数的数据模型，预测模型体现输出的恒星有效温度 T_{eff} 即被解释变量与线指数输入的多个特征变量的线性和非线性关系。

3.1　线指数共线性

结合上节提及的共线性问题是多个线指数的特征值变量给出相似的分析，LAMOST 数据绘制散点图呈现所有变量的散点图表示响应变量与解释变量之间的关系相关性，利用方差膨胀因子 VIF 确定解释变量的共线性程度。实验利用多线性共同标准方差膨胀因子 VIF$=1/(1-r_j^2)$，其中 r_j^2 表示多个其他相关变量的回归系数，线指数通过 X_j 变量计算 VIF，得到 hgamma12 的 VIF 值为 $3.288\ 479\times10^6$ 远远超过 VIF 的最大限度影响因子 $r_j^2>0.9$，故存在多重共线性，多线性分析会影响估计量的准确性。依据存在非线性的因素，建立多线性回归模型变量的相关系数，得到与有效温度相关的皮尔森相关系数矩阵，建立模型残差为 1.213，调整可决系数为 0.993，优化模型线指数特征值结果。如式(3)表述线性组合在两组随机变量 $X'X$ 中选取若干个相关关系的指标，表示原来的两组变量的综合关系。后文实验采用岭回归估计在变量 $X'X$ 中增加正常矩阵 $kI(k>0)$，则 $X'X+kI$ 更接近真实的回归值，符合参数 k 值如图 5 所示，正规方程最优解时当 $k\to0$ 时 $\hat{\beta}(0)$得到原来的最小二乘估计，训练线性回归模型为式(3)。

$$\hat{\beta}(k)=(X'X+kI)^{-1}X'y \tag{3}$$

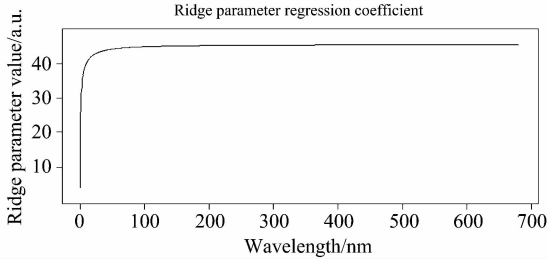

图 5　A 型恒星线指数系数线性回归估计分析图

Fig. 5　Linear regression analysis diagram of A-type star line exponential coefficient

3.2　岭回归分析

从上文得到线指数值实现多元线性回归时系数矩阵与其转置矩阵相乘得到的矩阵不能求逆，且方差较大使得光谱特征变量间存在共线性造成最小二乘回归不稳定。为此本节通过 Ridge 岭回归解决最小二乘法的无偏性，没有抛弃任何特征缩小回归系数获得可靠的回归系数预测大气有效温度参数预测模型 teff$=12\ 770+\beta_1$ kp12$+\cdots+\beta_{26}$ hdelta_d02 式中 β 系数值，如表 2 列出各特征显示模型准确地从低分辨率光谱中确定 A 型恒星的有效温度。

表 2　岭回归模型线指数特征值与 T_{eff} 参数线性相关系数值

Table 2　Line index characteristic value and T_{eff} parameter linear correlation value with ridge regression model

特征值	系数 β	特征值	系数 β
kp12	-2.981×10^7	hdelta12	1.506×10^8
kp18	8.222×10^6	hdelta24	-3.067×10^7
kp6	3.503×10^7	hdelta48	-1.167×10^8
hbeta12	-4.225×10^8	hdelta64	4.939×10^7
hbeta24	3.096×10^8	hgamma12	2.523×10^8
hbeta48	-5.391×10^7	hgamma24	-1.186×10^7
hbeta60	2.567×10^7	hgamma48	-6.411×10^8
halpha12	-1.239×10^8	hgamma54	5.244×10^8
halpha24	2.038×10^8	paschen13	9.443×10^5
halpha48	-3.923×10^8	paschen142	-6.710×10^6
halpha70	2.888×10^8	paschen242	-2.306×10^7
halpha_d02	2.529×10^4	hgamma_d02	-4.984×10^3
hbeta_d02	-2.138×10^4	hdelta_d02	3.635×10^6

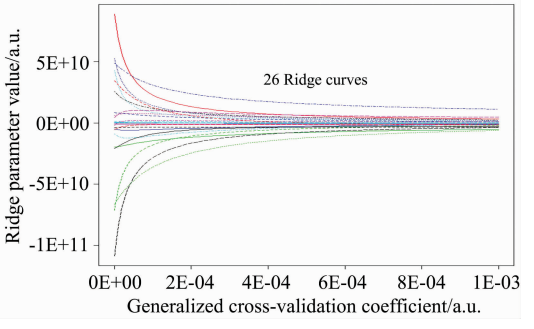

图 6　A 型恒星有效温度 T_{eff} 与 26 种线指数特征值岭回归分析图，修正后的估计值 HKB 为 $1.921\ 567\times10^{-5}$ 和 L-W 为 330.336 5

Fig. 6　Ridge regression analysis chart of T_{eff} and 26 kinds of eigen values with the line index of A-type stars, which modified HKB estimator was $1.921\ 567\times10^{-5}$ and modified L-W estimator was 330.336 5

当变量间存在共线性且方差很大，得到不稳定的最小二乘回归系数。为此系数矩阵 X 与其转置矩阵相乘得到的矩阵不能求得其逆矩阵，实验通过 ridge regression 函数引入参数 lambda，解决上述问题，利用第 1.1 节中列出 26 种特征值选

择岭回归参数 k，从优化模型运行结果得岭回归参数值为 0.014 7，各自变量的系数显著明显提高，岭回归模型的 lambda 值代入线性回归模型，得到 T_{eff} 有偏的估计，也可采用优化广义交叉验证 GCV 方法自动选取得到最佳岭回归的参数 k 值如图 6 所示，经岭回归计算变量的相关性分析，合理简化 LAMOST 发布的线指数变量值，输入由该组变量的数值预测有效温度以增强预测模型的可信度。

4 展 望

线指数作为描述天文光谱较有效的数据特征方式，若将每个波长采样点视作一个维度，则需降维天文高维光谱数据，进而获取更好的数据分布和知识信息。天文光谱直接从谱线特征中获取恒星有效温度 T_{eff} 值具有很好的研究价值，特别是通过训练信息丰富的线指数值得出 A 型恒星特征与 T_{eff} 之间的关系模型，利用 LAMOST 发布的光谱和相应的恒星参数来获得这种关系，尤其提供晚期 A 型恒星的 T_{eff} 更为准确。本文依据光学巡天光谱数据的 Lick 线指数特征值，运用 LAMOST 发布 DR5 实测数据计算预测有效温度 T_{eff}，实验中偏回归系数对 A 型恒星线指数数据绘制散点图其中 X_1，X_2，X_3 分别为 kp12，hdelta12，hgamma12 变量与有效温度多重共线性分析相互关联时，产生多重共线性引起系数的噪声波动，降低其显著性。岭回归预测模型既解决过拟合问题，也给出大样本实测光谱数据预测有效温度 T_{eff} 的方法，进而正确预测未来恒星演化的发展趋势，为后续研究分析 A 型恒星演化规律提供必要的论证模型。

References

[1] Luo Ali, Zhao Yongheng, Zhao Gang, et al. Research in Astron. Astrophys, 2015, 15(8): 1095.

[2] ZHAO Yong-heng(赵永恒). Scientia Sinica: Physica, Mechancia&Astronomica(中国科学：物理学力学天文学), 2014, 44(10): 1041.

[3] CUI Chen-zhou, YU Ce, XIAO Jian, et al(崔辰州，于 策，肖 健，等). Chinese Science Bulletin(科学通报), 2015, 60(5-6): 445.

[4] Liu Chao, Cui Wenyuan, Zhang Bo, et al. Research in Astron. Astrophys, 2015, 15(8): 1137.

[5] WANG Guang-pei, PAN Jing-chang, YI Zhen-ping, et al(王光沛，潘景昌，衣振萍，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(8): 2646.

[6] Guy Worthey, Faber S M, et al. The Astrophysical Journal Supplement Series, 1994, 94: 687.

[7] Chen Shuxin, Sun Weimin, Yan Qi. Research in Astron. Astrophys, 2018, 18(6): 73.

Line Index of A-Type Stellar Astronomical Spectrum Predict Effective Temperature by Ridge Regression Model

XUE Ren-zheng[1], CHEN Shu-xin[1*], HUANG Hong-ben[2]

1. School of Computer and Control Engineering, Qiqihar University, Qiqihar 161006, China

2. School of Data Science and Software Engineering, Wuzhou University, Wuzhou 543002, China

Abstract Line index is widely used in describing the features of spectral lines for astronomical objects because it retains the main physical characteristic information of these objects. Based on line index, a multi-parameter model for regression analysis could be used to uncover co-variation relationship of data and the inherent laws of spectral lines. The observed spectra released by LAMOST, which has the highest spectra acquisition capability, provide us with real data for establishing a robust regression model. The multivariate linear regression was applied to get the co-linearity of the dependent variables, however, it resulted in large variance. It is unstable to obtain the least squares regression coefficient sometimes. Especially, it's difficult for the multivariate linear regression to obtain the evaluation coefficient of independent predictor from the regression equation. In this paper, we use the A-type stellar Lick line index in the LAMOST survey data as the data source. Selecting the spectra with effective temperature (T_{eff}) from 7 000 to 8 500 K, and the signal-to-noise ratio higher than 50 to realize the regression analysis. After a set of linear biased estimation experiment for A-type stars, the method of ridge regression training was employed. In the catalogue of LAMOST data release 5 (DR5), 86 097 A-type spectra have provided the T_{eff} value. After statistical analysis of the eigenvalues of 26 line indices, the kp12, halpha12 and hgamma12 with similar distribution and bandwidth of 12 Å were selected to reduce the data redundance. The number of variety was optimized for the redundant variable variance expansion factor (VIF) coefficient. Two regression experiments selected the same observation dataset to locally fit the regression scatter, using the overall contour of the scatter plot to generate a high-density scatter plot, highlighting the data-intensive region with the color difference transparency. The results show that both the multiple linear regression and the ridge regression algorithm can determine the effective temperature (T_{eff}) of the A-type star through the low-resolution spectrum, but the co-linearity data analysis has some biased es-

timation. The ridge regression model can more accurately predict the effective temperature of A type stars from the low resolution spectra.

Keywords Stellar spectra; LAMOST (Large sky area multi-object fiber spectroscopy telescope); Ridge regression; Linear model; Lick line index

(Received Feb. 24, 2019; accepted May 16, 2019)

* Corresponding author

敬告读者——《光谱学与光谱分析》已全文上网

从 2008 年第 7 期开始在《光谱学与光谱分析》网站(www.gpxygpfx.com)“在线期刊”栏内发布《光谱学与光谱分析》期刊全文，读者可方便地免费下载摘要和 PDF 全文，欢迎浏览、检索本刊当期的全部内容；并陆续刊出自 2004 年以后出版的各期摘要和 PDF 全文内容。2009 年起《光谱学与光谱分析》每期出版日期改为每月 1 日。

《光谱学与光谱分析》期刊社