

doi: 10.3969/j.issn.1000-8349.2020.01.04

The Cannon: 一种基于光谱数据的 恒星参数测量方法

黄轶琦^{1,2}, 钟 靖¹, 侯金良^{1,2}

(1. 中国科学院 上海天文台 星系与宇宙学重点实验室, 上海 200030; 2. 中国科学院大学, 北京 100049)

摘要: 新一代大规模光谱巡天项目产生了近千万条低分辨率恒星光谱, 基于这些光谱数据, 介绍一种名为 The Cannon 的机器学习方法。该方法完全基于已知恒星大气参数 (有效温度、表面重力加速度和金属丰度等) 的光谱数据, 通过数据驱动来构建特征向量, 建立光谱流量特征和恒星参数的函数对应关系, 进而应用到观测光谱数据中, 实现对恒星光谱的大气参数求解。The Cannon 的主要优势为不直接基于任何恒星物理模型, 适用性更广; 由于使用了全谱信息, 即便对于低信噪比光谱也能得到较高可信度的参数结果, 该算法在大规模恒星光谱的数据处理和参数求解方面具有明显的优势。此外, 还利用 The Cannon 得到 LAMOST 光谱数据中 K 巨星和 M 巨星的恒星参数。

关 键 词: 光谱巡天; 恒星参数; 数据驱动

中图分类号: P144

文献标识码: A

1 引 言

近年来, 人们在中低分辨率的光谱巡天观测取得了举世瞩目的成就。郭守敬望远镜 (Large Sky Area Multi-Object Fiber Spectroscopy Telescope, LAMOST) 为代表的望远镜获取的千万量级光谱数据为银河系天文学的研究提供了巨大的数据支持, 也对目前的数据分析和参数测量提出了巨大的挑战。如何快速准确地求得恒星参数, 成为当前天文大数据时代一个亟待解决的问题。

传统的恒星参数测量方法主要基于恒星大气模型, 通过特定参数条件下理论光谱与观测光谱的拟合来确定其恒星参数。然而, 该参数测量方法通常受限于拟合光谱参数所用到的恒星大气模型以及观测光谱的谱分辨率, 一般只适用于中高分辨率恒星光谱, 且参数处

收稿日期: 2019-04-15; 修回日期: 2019-05-24

资助项目: 国家自然科学基金 (U1731129); 973 项目 (2014CB845702)

通讯作者: 钟靖, jzhong@shao.ac.cn

理的效率低下。随后发展起来的模板匹配方法虽然实现了大规模中低分辨率光谱的统一参数测量,但是该方法受光谱模板的限制十分明显,其模板的参数范围、光谱范围以及光谱分辨率直接决定了参数求解的精度和适用的恒星类型。考虑到光谱模板主要来源于实测光谱,其参数空间很难做到均匀,模板光谱的参数间隔较大,需要进行插值来满足与目标光谱的匹配要求。而插值的算法往往具有很大的不确定性,从而导致参数测量结果的不确定度进一步增大。此外,由于不同类型恒星具有不同的表面温度,因此其光谱特征的差异十分明显。用于斯隆数字巡天 (Sloan Digital Sky Survey, SDSS) 的恒星参数测量程序 (Sloan extension for galactic understanding and exploration stellar parameter pipeline, SSPP)^[1, 2] 和法国里昂大学光谱分析软件 (university of Lyon spectroscopic analysis software, ULYSS) 为代表的恒星参数测量方法,主要对原子谱线的测量 (如 H 的巴尔末线、Ca II 三重线等) 进行了优化,并进一步丰富了恒星光谱的参数网格。利用该类测量方法,虽然人们对 A-F-G-K 型恒星取得了较好的测量结果,但由于方法及模板类型的限制,对于更低温的晚型 M 型星以及更高温的 OB 型星的参数测量精度仍然差强人意,甚至无法给出参数估计结果。值得注意的是,随着巡天光谱数据的不断增加,越来越多基于数据的参数估计方法开始应用到大规模巡天数据中,如支持向量机 (support vector machine, SVM)^[3], The Payne^[4, 5], t-SNE (stochastic neighbor embedding, SNE)^[6] 等,在处理特定参数测量问题时都有各自的优势。

随着多目标光纤光谱仪的运用,越来越多的大望远镜参与了光谱巡天计划,中高分辨率光学/红外光谱的观测数据积累也达到了百万量级。虽然高分辨率光谱的波长范围很短,难以作为光谱模板进行全谱拟合,但其相对准确的参数测量结果对于低分辨率光谱仍然具有较高的参考价值。在这样的背景下,一种完全基于数据驱动的方法——The Cannon^[7]应运而生。这一名称来源于天文学家 A J Cannon,她是推动无物理模型恒星分类的先驱。该算法完全从数据出发,不直接基于任何物理模型或假设,且光谱处理速度快,参数测量准确度高,能够很好地将高分辨率光谱的参数结果应用于低分辨率光谱的测量,十分适合于大样本低分辨率光谱的参数求解,以及不同巡天数据之间的交叉定标。

作为一种数据驱动方法,The Cannon 首先需要建立一批已知恒星大气参数的光谱训练样本,并利用计算机的大规模运算能力来构建恒星参数与光谱流量的函数关系,进而将这一关系应用到所有恒星光谱数据中,实现对恒星光谱的大气参数求解。由于这一过程不直接依赖于任何物理假设,而是类似于利用事物特征进行分类命名的方法,故而本文将各种恒星大气参数的集合表述为“恒星标签”,The Cannon 的核心就是基于光谱特征为恒星贴上具有不同大气参数的恒星标签。

2 The Cannon 简介

2.1 第一步:训练样本的建立

为了使用 The Cannon 对恒星参数进行求解,首先需要构建一个已知大气参数值的光谱训练样本。训练样本的光谱质量和参数测量精度直接影响了光谱模型的建立以及后续恒星参数的求解,因而合理选择具有高信噪比流量和高精度参数的光谱作为训练样本尤为重要。

此外, 该训练样本还必须经过一定的预处理, 以方便后续光谱特征的分析 and 建模, 主要包括: (1) 视向速度改正, 移至静止坐标系; (2) 所有光谱波长范围确保一致; (3) 对光谱进行流量归一化处理。

2.2 第二步: 光谱模型的建立

基于训练样本, The Cannon 建立了每个波长点的光谱流量值与恒星参数之间的函数关系。这是一个概率生成模型, 即在恒星光谱的每个波长点生成其流量的概率分布函数, 从而得到该波长点流量的期望值和方差。它满足以下两条基本假设:

(1) 在不考虑观测误差的情况下, 相同大气参数的归一化连续谱在每个波长点的流量值相同 (实际上这只是一个近似, 例如, 具有相同 T_{eff} , $\lg g$, $[\text{Fe}/\text{H}]$ 的恒星, 它们的光谱可能不同, 因为它们的年龄和旋转速度可能不同)。

(2) 在每个波长点的流量值随大气参数的变化是平滑的。简言之, 光谱模型是光谱每个波长点的流量值作用于恒星大气参数的平滑函数。该函数不是唯一的, 能够根据实际需要选择函数的复杂程度, 这也是 The Cannon 灵活度的表现。

假设有 N 个训练样本, 每个样本 n 在波长点 λ 有流量值 $f_{n\lambda}$, 每个样本都有由 K 个参数组成一组恒星标签 ℓ_{nk} , 并表示成一个参数矢量 ℓ_n 。根据基本假设, 每个样本在每个波长点 λ 的流量值 $f_{n\lambda}$ 可以表示为相应恒星标签 $\ell_{nk}(T_{\text{eff}}, \lg g, [\text{Fe}/\text{H}], \dots)$ 的连续性方程, 与广义线性模型相似, 特征向量 θ_λ 表示方程在每个波长点的系数向量。在已知特征向量 θ_λ 的光谱模型中, 给定恒星标签 ℓ_n 的流量值 $f_{n\lambda}$ 一般用线性函数来表示:

$$f_{n\lambda} = \theta_\lambda^T \cdot \ell_n + n, \quad (1)$$

其中, 波长 λ 处的噪声项 (n) 主要由仪器效应和光子计数导致的误差 $\sigma_{n\lambda}$ 和光谱模型在拟合过程中产生的弥散 s_λ 这两部分组成。该噪声模型可以表示为 $n = [s_\lambda^2 + \sigma_{n\lambda}^2] \xi_{n\lambda}$, 其中 $\xi_{n\lambda}$ 是一个均值为 0, 方差为 1 的高斯随机数。

首先考虑最简单的一阶光谱模型, 标签矢量 ℓ_n 为线性, 其参数矢量表示为:

$$\ell_n \equiv \left[1, \ell_{n1} - \bar{\ell}_1, \ell_{n2} - \bar{\ell}_2, \dots, \ell_{nK} - \bar{\ell}_K \right], \quad (2)$$

这里的第一个元素 “1” 意味着拟合过程中允许进行线性平移。 ℓ_k 通常选取训练样本的参数平均值, 从而使得光谱模型能够在参数空间的合理范围内震荡。此模型在每个波长点满足如下对数似然函数方程:

$$\ln p(f_{n\lambda} | \theta_\lambda^T, \ell_n, s_\lambda^2) = -\frac{1}{2} \frac{[f_{n\lambda} - \theta_\lambda^T \cdot \ell_n]^2}{s_\lambda^2 + \sigma_{n\lambda}^2} - \frac{1}{2} \ln(s_\lambda^2 + \sigma_{n\lambda}^2), \quad (3)$$

其中, 矢量 f_λ 是所有样本在波长点 λ 流量的集合。式 (3) 也是恒星标签 ℓ_n 和特征向量 θ_λ 在每个波长点 λ 的概率密度分布函数。

可见, 基于训练样本, 在已知光谱流量 $f_{n\lambda}$ 和恒星标签 ℓ_n 的情况下, 可以解算得到光谱模型的特征向量 θ_λ 及其弥散 s_λ :

$$\theta_\lambda, s_\lambda \leftarrow \arg \max_{\theta_\lambda, s_\lambda} \sum_{n=1}^N N \ln p(f_{n\lambda} | \theta_\lambda^T, \ell_n, s_\lambda^2). \quad (4)$$

相应地, 二阶的光谱模型对应的标签矢量可以表示为:

$$\ell_n \equiv \left[1, l_{n1} - \bar{l}_1, l_{n2} - \bar{l}_2, \dots, l_{nK} - \bar{l}_K, (l_{n1} - \bar{l}_1)(l_{n1} - \bar{l}_1), \right. \\ \left. (l_{n1} - \bar{l}_1)(l_{n2} - \bar{l}_2), \dots, (l_{nK} - \bar{l}_K)(l_{nK} - \bar{l}_K) \right] . \quad (5)$$

在基于训练样本求解特征向量 θ_λ 和弥散 s_λ 时, 二阶标签矢量的求解过程和一阶标签矢量类似, 需要解算的光谱模型都是线性的; 但是, 后续对恒星标签的求解过程会比较复杂, 需要通过非线性拟合来确定恒星参数。

2.3 第三步: 恒星参数的求解

如上所述, 基于训练样本的光谱流量和恒星参数, The Cannon 能够解算得到一个覆盖了一定参数空间范围的光谱模型 (概率生成模型)。该模型可以在给定每个波长点光谱流量的前提下解算其对应的恒星参数。对于目标光谱, 在进行了包括归一化、红移改正和波长范围对齐等预处理操作后, 基于求得的光谱模型特征向量和弥散 $[\theta_\lambda, s_\lambda^2]$, 以及所有波长点的流量值 $f_{m\lambda}$, 最终可通过积分得到它的恒星标签值:

$$\{\ell_{mk}\} \leftarrow \arg \max_{\{\ell_{mk}\}} \sum_{\lambda=1} N_{\text{pix}} \ln p(f_{m\lambda} | \theta_\lambda^T, \ell_m, s_\lambda^2) . \quad (6)$$

3 The Cannon 使用实例

3.1 利用 The Cannon 求解 LAMOST DR2 K 巨星大气参数

Ho 等人^[8]利用 The Cannon 这一数据驱动方法对 LAMOST K 巨星光谱中 4 个恒星参数进行了求解。他们结合 APOGEE 提供的高分辨率光谱恒星参数 (T_{eff} , $\lg g$, $[\text{Fe}/\text{H}]$, $[\alpha/\text{M}]$), 对 4.5×10^5 条 LAMOST DR2 光谱数据中的 K 巨星恒星参数进行了重新的解算, 结果显示对于 K 巨星来说, The Cannon 对恒星参数的测量精度要明显优于 LAMOST 的恒星参数处理程序得到的参数结果 (对于信噪比大于 50 的光谱, The Cannon 的恒星参数测量误差为 $T_{\text{eff}} \approx 70 \text{ K}$, $\lg g \approx 0.1$, $[\text{Fe}/\text{H}] \approx 0.1$, $[\alpha/\text{M}] \approx 0.04$)。

作为一个低分辨率 ($R \approx 1800$) 光谱巡天望远镜, LAMOST 的第二次数据释放 (DR2^[9]) 包含约 410 万条光谱和 220 万颗恒星的大气参数 (T_{eff} , $\lg g$, $[\text{Fe}/\text{H}]$), 其中 K 巨星约 50 万颗^[10]。LAMOST 1D 光谱由 LAMOST 2D 流水线处理程序统一处理, 恒星参数来自于 LASP 恒星参数处理程序^[11, 12]。以 SDSS DR9 公共星的流水线参数为标准, LAMOST 大气参数的测量偏差和弥散为 $T_{\text{eff}} \approx (-91 \pm 111) \text{ K}$, $\lg g \approx (0.16 \pm 0.22)$, $[\text{Fe}/\text{H}] \approx (0.04 \pm 0.15)$ ^[13]。

阿帕奇顶点天文台星系演化实验 (the Apache Point Observatory galactic evolution experiment, APOGEE) 是一个高分辨率 ($R \approx 22500$) 和高信噪比 ($S/N \approx 100$) 的近红外 ($1.514 \sim 1.696 \mu\text{m}$) 光谱巡天计划^[14, 15]。APOGEE 巡天使用位于美国新墨西哥州阿帕奇波因特天文台的 2.5 m 斯隆望远镜进行观测, 拥有 300 条光纤, 主要观测目标是位于银河系核球、银盘和银晕的红巨星 (K/M 巨星)。其释放的 DR12^[16, 17] 包含约 1×10^5 颗红巨星光谱的高精度大气参数和化学丰度值。

3.1.1 LAMOST 光谱预处理

利用 The Cannon 进行恒星参数求解之前, Ho 等人^[8]首先对 LAMOST 光谱进行了预处理: 利用每条 LAMOST 光谱头文件中的红移值把光谱移至静止坐标系, 然后将所有光谱的波长范围调整为 $3905 \sim 9000 \text{ \AA}$, 最后将每条光谱进行归一化。归一化光谱为:

$$\bar{f}(\lambda_0) = \frac{\sum_i [f_i \sigma_i^{-2} \omega_i(\lambda_0)]}{\sum_i [\sigma_i^{-2} \omega_i(\lambda_0)]}, \quad (7)$$

这里, f_i 是波长点 i 的流量, σ_i 是波长点 i 的测量误差, 权重 $\omega_i(\lambda_0)$ 由高斯函数得到:

$$\omega_i(\lambda_0) = e^{-\frac{(\lambda_0 - \lambda_i)^2}{L^2}}, \quad (8)$$

其中, L 是自由参数, 表征了高斯平滑的范围。考虑到 LAMOST 的分辨率以及 K 巨星光谱的主要特征, 该工作的 L 取 $50 \text{ \AA}$, 因此其平滑范围远大于一般原子谱线宽度。

3.1.2 The Cannon 的应用及结果检验

为了从 LAMOST DR2 和 APOGEE DR12 的 11057 个公共源中挑选出可靠的训练样本, Ho 等人^[8]根据描述的不可靠条件或者是有 ASPCAPFLAG 标记的星, 剔除了 677 个不可靠参数的样本, 剩下 10380 个源。随后, Ho 等人^[8]把这 10380 个源全部作为训练样本并重新测量它们的恒星参数值, 并剔除了参数值与 APOGEE 标准值的差值大于 4σ 的 428 个源 (小于 0.5%), 然后把剩下的 9952 个源作为最终的训练样本。

图 1 显示了 9952 个源在 $T_{\text{eff}}\text{-}\lg g$ 空间的分布情况。黑点表示所有 LAMOST DR2 中的源, 图 1a) 中的彩色点表示具有 LAMOST 参数的训练样本, 图 1b) 中的彩色点表示具有 APOGEE 参数值的训练样本。由于分辨率和光谱信噪比的限制, LAMOST 在参数空间的弥散度比 APOGEE 更大, 尤其对于巨星。

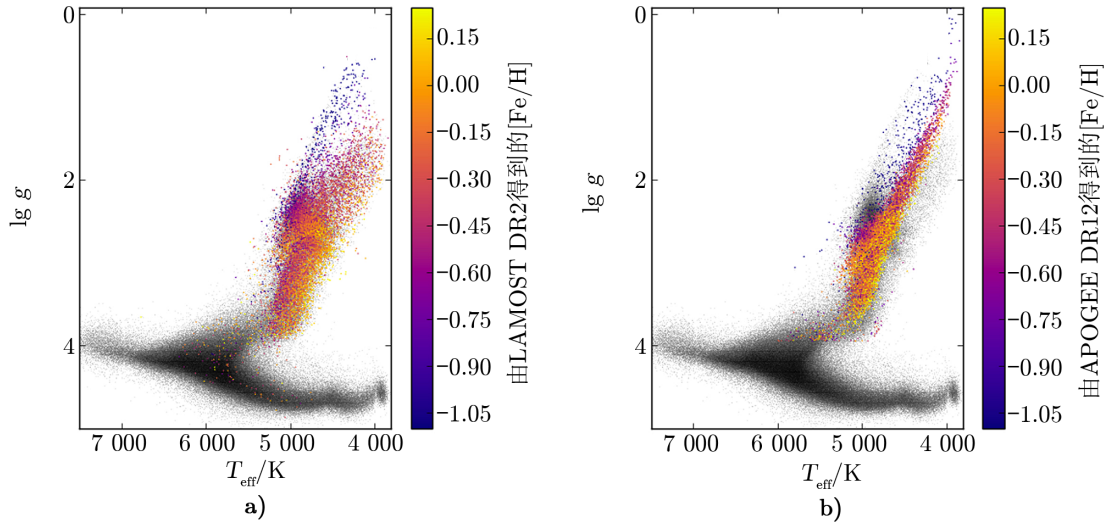


图 1 9952 个公共源在 $T_{\text{eff}}\text{-}\lg g$ 空间的分布情况^[8]

为了得到 K 巨星的光谱模型, Ho 等人使用了二阶标签矢量:

$$\ell_n \equiv \left[1, T_{\text{eff}}, \lg g, [\text{Fe}/\text{H}], [\alpha/\text{M}], A_k, T_{\text{eff}} \cdot \lg g, T_{\text{eff}} \cdot [\alpha/\text{M}], T_{\text{eff}} \cdot A_k, \right. \\ \lg g \cdot [\text{Fe}/\text{H}], \lg g \cdot [\alpha/\text{M}], \lg g \cdot A_k, [\text{Fe}/\text{H}] \cdot [\alpha/\text{M}], [\text{Fe}/\text{H}] \cdot A_k, \\ \left. [\alpha/\text{M}] \cdot A_k, T_{\text{eff}}^2, \lg g^2, [\text{Fe}/\text{H}]^2, [\alpha/\text{M}]^2, A_k^2 \right]. \quad (9)$$

该标签矢量主要包括有效温度 T_{eff} 、表面重力加速度 $\lg g$ 、金属丰度 $[\text{Fe}/\text{H}]$ 、 α 元素丰度 $[\alpha/\text{M}]$ 以及 k 波段消光 A_k 这 5 个恒星参数。

基于训练样本的 LAMOST 流量和 APOGEE 恒星参数, 通过对数似然函数方程可以得到每个波长点的特征向量和弥散 $[\theta_\lambda, s_\lambda^2]$ 。图 2 展示了光谱模型中每个波长点的一阶特征向量和弥散值。该一阶特征向量直观显示了每个波长点对不同恒星参数的敏感程度。

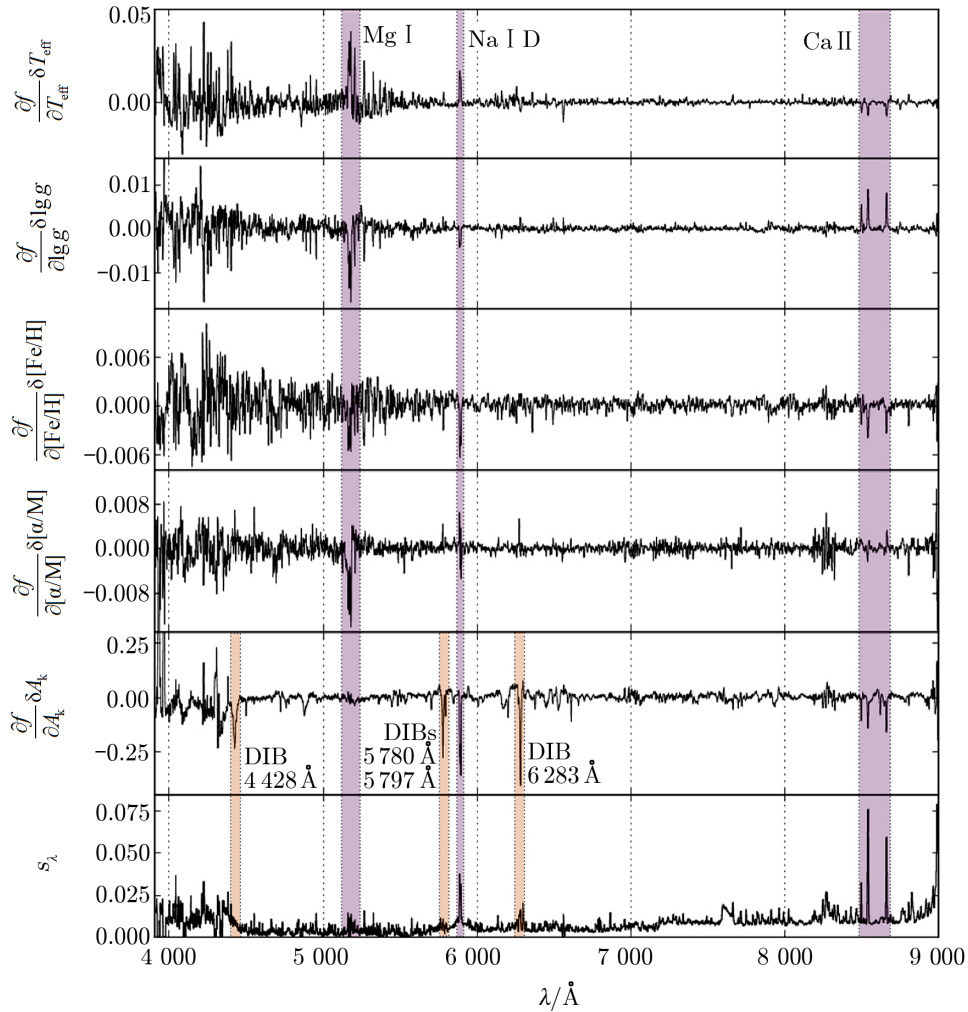


图 2 通过 9952 个训练样本得到的光谱模型^[8]

在已知光谱模型特征向量和弥散 $[\theta_\lambda, s_\lambda^2]$ 的前提下, 将光谱模型应用到所有待求的 LAMOST 光谱中, 最终得到 LAMOST K 巨星的恒星参数值。由于光谱模型使用了恒星参数的二阶项, 需要进行非线性拟合, Ho 等人^[8]采用了 scipy 库中的 Levenberg-Marquardt 算法来实现。

为了检验光谱模型的可信度, Ho 等人^[8]首先对训练样本进行了检验。具体过程如下: 把 9952 个源均分成 8 组, 编号为 0–7, 分 8 次实验, 每次拿走一组记为 k , 再用剩下的 7 组训练出一个光谱模型, 并用它来求解 k 组的恒星参数值。如此求解 8 次, 最终得到全部训练样本的参数值。将 The Cannon 得到的恒星参数值 (T_{eff} , $\lg g$, $[\text{Fe}/\text{H}]$, $[\alpha/\text{M}]$, A_k) 与 ASPCAP 得到的恒星参数值进行对比, Ho 等人^[8]发现所有参数的系统偏差和弥散均在合理范围之内, 如图 3 所示。

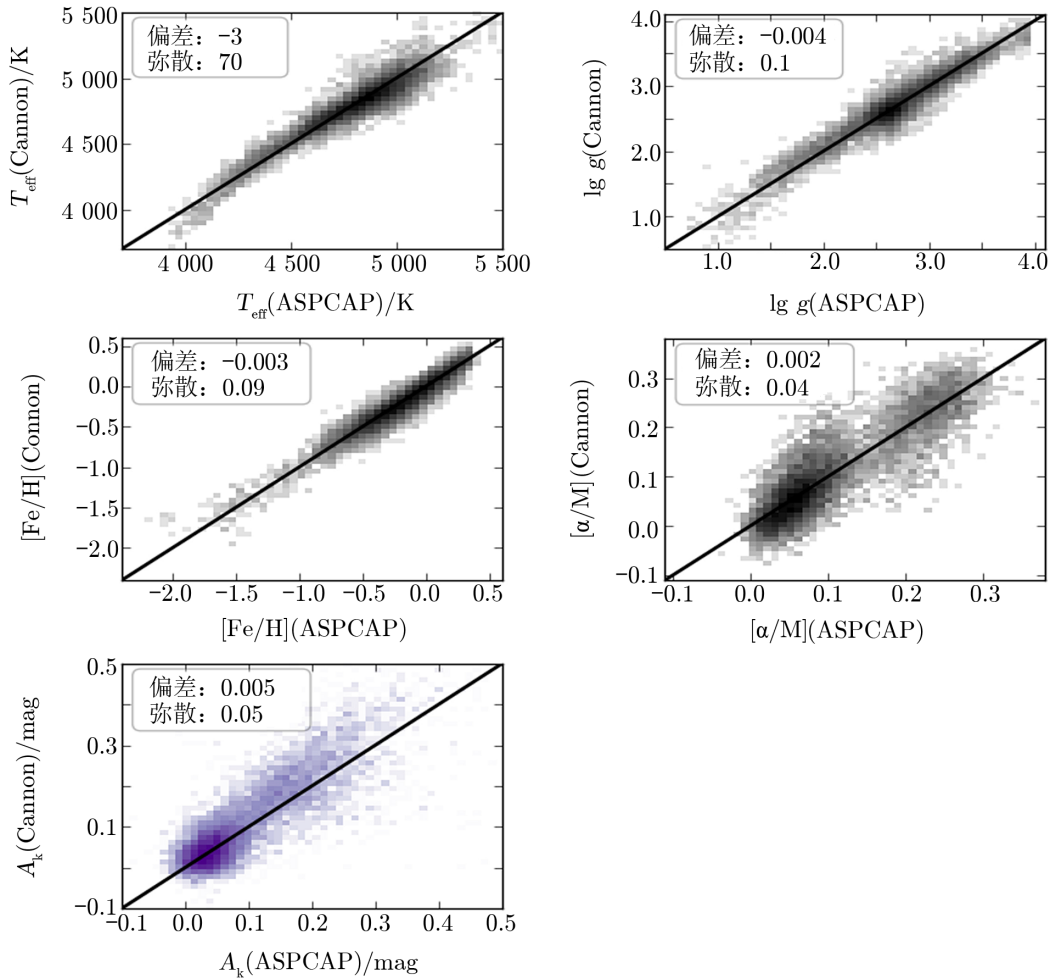


图 3 The Cannon 得到的恒星参数值与 ASPCAP 得到的恒星参数值对比^[8]

此外, Ho 等人^[8]还进一步比较了不同方法所求得的 K 巨星样本参数结果的分布情况, 结果如图 4 所示。对于同一批公共星样本, 3 幅图比较了 LASP, The Cannon 和 ASPCAP 三种方法所得到的测量结果在赫罗图上的分布。虽然使用的是 LAMOST 低分辨率光谱, 但 The Cannon 的结果明显好于同样基于 LAMOST 光谱的 LASP 结果, 其红团簇星的分布更接近采用 APOGEE 光谱的 ASPCAP 结果。

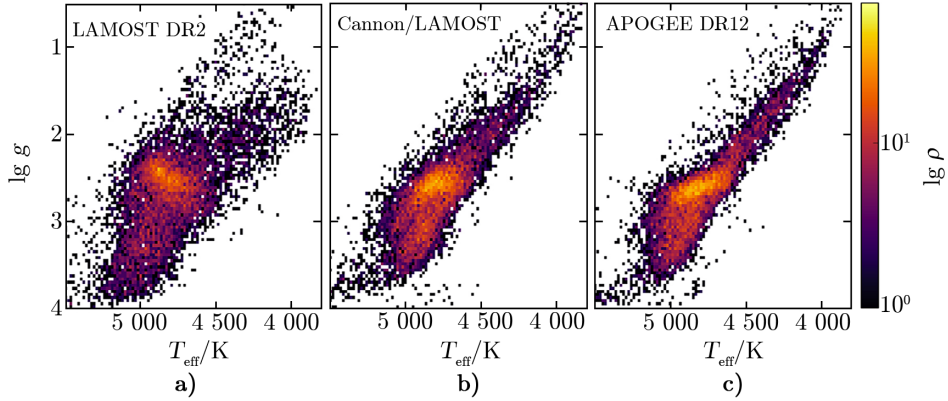


图 4 同一批公共星样本的来自三种方法的测量结果在赫罗图上的分布^[8]

为了验证 The Cannon 的优势, Ho 等人^[8]还详细讨论了 The Cannon 与 LASP 所得各参数结果的对比情况, 结果均显示基于同样的 LAMOST 光谱, The Cannon 结果比 LASP 结果明显提高。

3.2 利用 The Cannon 求解 LAMOST M 巨星大气参数

相比于 K 巨星, M 巨星的表面温度更低, 光度更高, 相同条件下的可探测距离大大超越其他类型恒星。如果能够可靠地获得包含大气参数、视向速度和距离信息在内的大样本 M 巨星统计数据, 其遥远的示踪范围将能够进一步扩展人们基于现有 A-F-G-K 型恒星样本对银河系外盘及银晕的认识, 为深入研究银河系结构和演化提供关键性依据: 譬如, 银河系的外盘结构和星族特征、银河系暗物质晕的密度轮廓和整体质量、已知星流的化学动力学演化以及银河系晕中星流结构的发现和证认等。

LAMOST DR5 数据发布了超过 5×10^5 条 M 型星数据, 其中包括约 4×10^4 条 M 巨星光谱数据^[19]。但是, 与具有恒星参数测量值的 A-F-G-K 型恒星星表相比, LAMOST 发布的 M 型星星表并没有列出大气参数 (T_{eff} , $\log g$, $[\text{Fe}/\text{H}]$) 的测量结果, 而只包括位置、光谱类型、 $\text{H}\alpha$ 等值宽度、多条分子谱线的谱指数, 以及用来标定 M 矮星金属丰度的 ζ 指数等少数光谱观测量。由于缺乏 M 巨星基本恒星参数的测量, 无法构建相应的增值星表, 这一 M 巨星样本难以被有效地应用到银河系相关的各类前沿研究中。

APOGEE 仅对 K/M 巨星进行了高分辨率的光谱观测, 因此其参数测量结果具有较高的准确性和可靠性。由 Ho 等人^[8]对 LAMOST DR2 数据中 K 巨星恒星参数进行求解的工作可知, The Cannon 能够很好地结合 LAMOST 光谱与 APOGEE 参数, 对晚型恒星进行

有效的恒星参数求解。通过搜寻 LAMOST 和 APOGEE 的 M 巨星公共星, 我们将能够建立具有 LAMOST 光谱和 APOGEE 恒星参数的 M 巨星训练样本, 构建光谱模型, 然后利用 The Cannon 求解所有 LAMOST DR5 光谱中的 M 巨星恒星参数。

我们将 LAMOST DR5 中的 M 巨星光谱^[18] (42 151 颗 M 巨星) 与 APOGEE DR14 的恒星参数星表进行了交叉认证, 共匹配到 2 909 个样本。由于 M 型星光谱主要表现为红端的分子谱线, 因此谱线特征受光谱信噪比和天光发射线的影响较大。为了尽可能确保光谱特征和恒星参数的可靠性, 我们从公共源中挑选了温度范围在 3 580 ~ 4 220 K, ASPCAP 等于 0 或 129 (尽量扩大样本数量), LAMOST 光谱信噪比和 APOGEE 光谱信噪比都高于 100 的 717 个源作为训练样本, 其参数分布如图 5 所示。

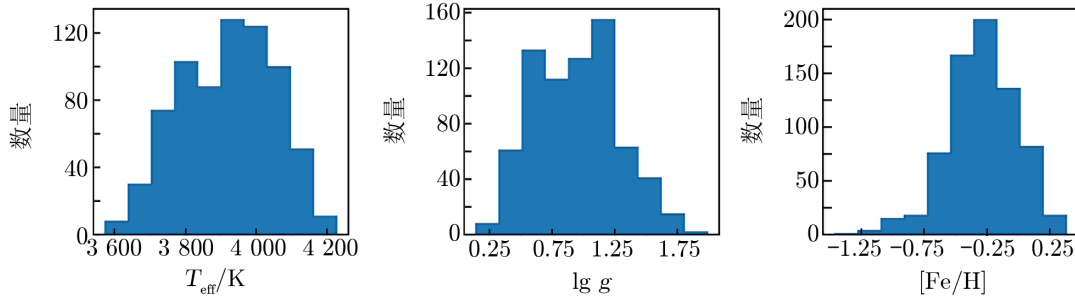


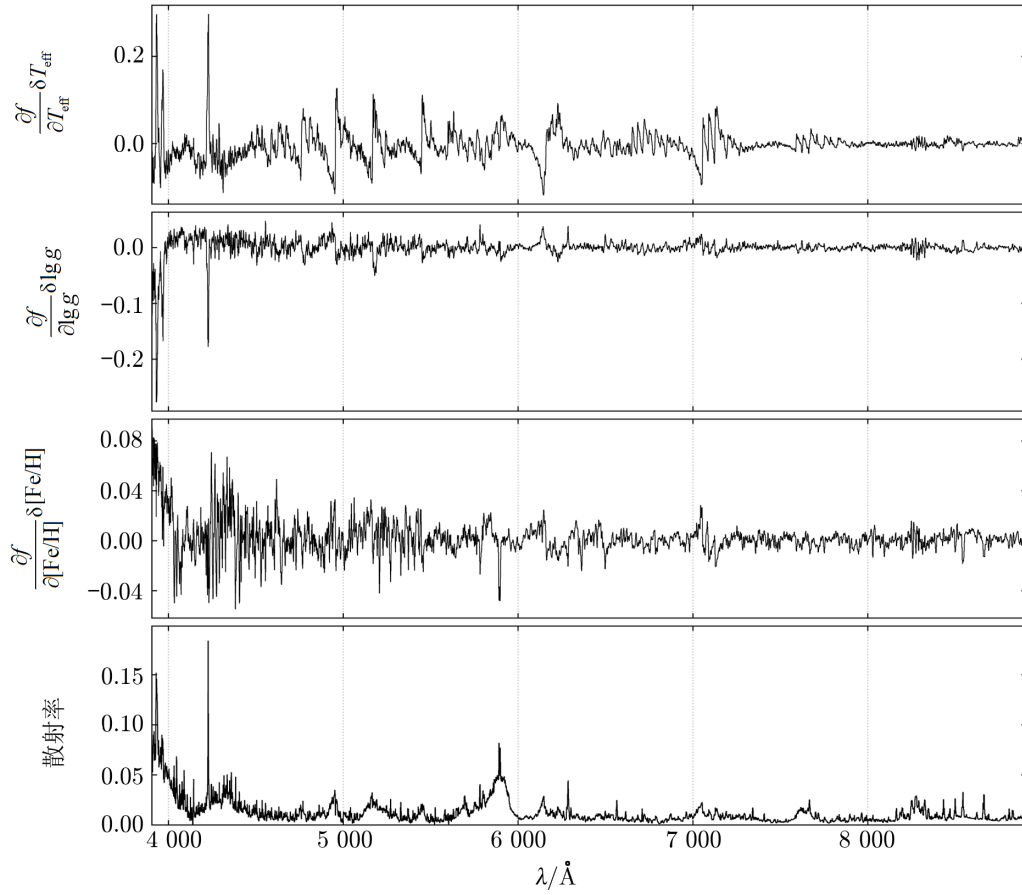
图 5 训练样本的参数分布

相比于 K 巨星光谱, M 巨星光谱缺少原子谱线, 且光谱在蓝端流量很低, 对 α 元素丰度的测量精度十分有限。为此, 我们只测量了恒星有效温度 T_{eff} 、表面重力加速度 $\lg g$ 和金属丰度 $[\text{Fe}/\text{H}]$ 这 3 个基本大气参数, 给定的恒星标签矢量为:

$$\ell_n \equiv \left[1, T_{\text{eff}}, \lg g, [\text{Fe}/\text{H}], T_{\text{eff}} \cdot \lg g, \lg g \cdot [\text{Fe}/\text{H}], T_{\text{eff}} \cdot [\text{Fe}/\text{H}], T_{\text{eff}}^2, \lg g^2, [\text{Fe}/\text{H}]^2 \right] . \quad (10)$$

按照 The Cannon 的操作步骤, 我们首先将训练样本的光谱进行视向速度改正、波长范围对齐, 以及流量归一化处理; 然后进一步应用 The Cannon, 基于训练样本计算了每个波长点的特征向量 $[\theta_\lambda, s_\lambda^2]$, 得到的光谱模型如图 6 所示。在已知 LAMOST M 巨星光谱模型基础上, 我们将 The Cannon 应用于 42 151 颗 LAMOST M 巨星光谱, 最终求得 M 巨星的基本大气参数 ($T_{\text{eff}}, \lg g, [\text{Fe}/\text{H}]$)。

为了检验 LAMOST 和 APOGEE 的训练样本得到的光谱模型的可信度, 我们对训练样本进行了类似 K 巨星的检验, 具体是把 717 个源分成 7 组, 编号为 0—6, 分 7 次实验, 每次拿走一组记为 k , 再用剩下的 6 组训练出一个光谱模型, 并用它来求解 k 组的恒星参数值。如此求解 7 次, 最终得到全部训练样本的新参数值。我们将由 The Cannon 得到大气参数值和由 ASPCAP 得到的大气参数进行了对比, 结果如图 7 所示。相比于 K 巨星的 The Cannon 结果, M 巨星的参数测量弥散度更大。这主要是因为 M 巨星表现为大量的分子吸收谱线, 在光谱归一化过程中很难真正将连续谱剔除干净; 此外, 由于缺乏蓝端的原子谱线, 其视向速度改正的精度也明显低于 K 巨星, 从而导致参数测量精度有所下降。考虑到



注：灰色点线表示整数波长的网格。

图 6 训练样本得到的光谱模型

LAMOST 低分辨率的限制，M 巨星参数测量精度与 LAMOST 标准化恒星参数处理程序 LASP 对 A-F-G-K 型星的测量精度接近，也进一步验证了 The Cannon 处理此类低分辨率光谱参数问题的有效性。

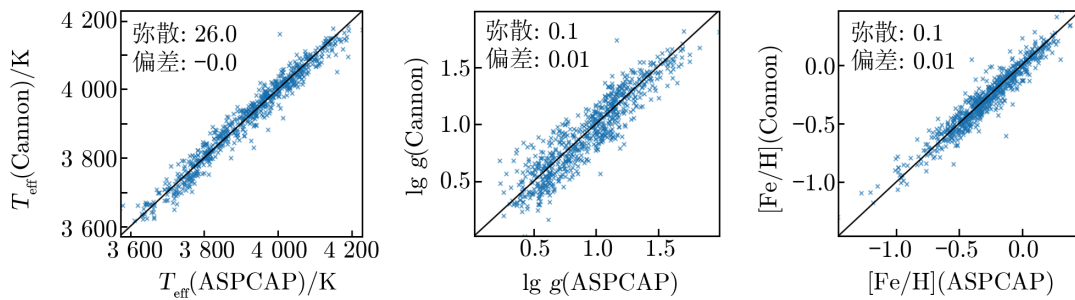


图 7 The Cannon 得到的大气参数与 ASPCAP 得到的大气参数的对比

图 8 显示了 LAMOST 和 APOGEE 的 2909 颗公共星样本的赫罗图分布, 图 8a) 为 The Cannon 的结果, 图 8b) 为 ASPCAP 的测量结果。The Cannon 结果的参数分布基本符合预期, 与 ASPCAP 给出的 M 巨星参数类似。可见在 3600 K 附近的低温端结果与 APOGEE 结果出现了一定程度的偏差, 出现此问题的原因可能有 2 个: (1) APOGEE 的 M 巨星参数在该参数空间区域附近出现了截止, 由于边界效应导致 3600 K 附近训练样本的数量和测量精度均有所欠缺, 从而使训练结果的可靠性下降; (2) 此温度范围内的恒星分子带进一步加强, 如何正确构建伪连续谱成为一个难题, 从而导致特征向量的不确定度增大。

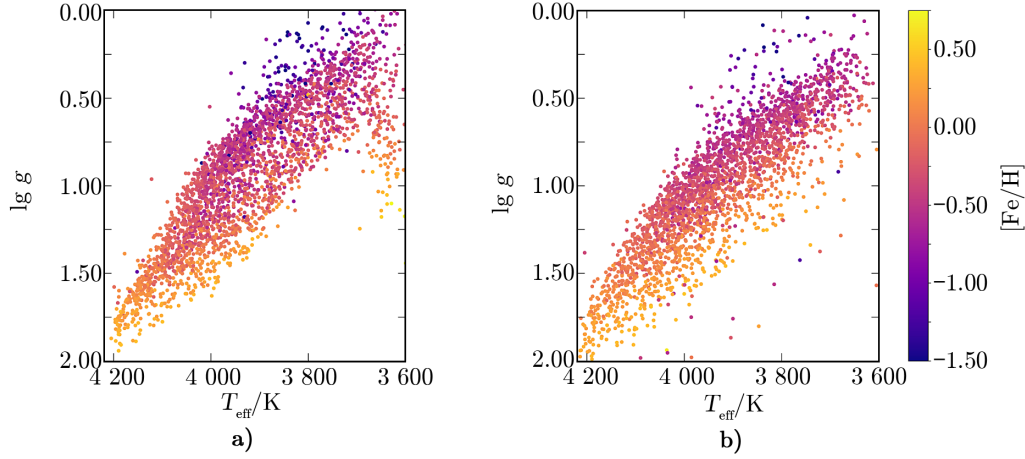


图 8 2909 个公共源的赫罗图分布

最后, 我们对所有 42151 颗 M 巨星的温度进行了外部检验, 将 Zhong 等人^[19]给出的 M 巨星光谱类型 M0—M6 与利用 The Cannon 得到的温度进行对比, 结果符合预期, 如图 9 所示。

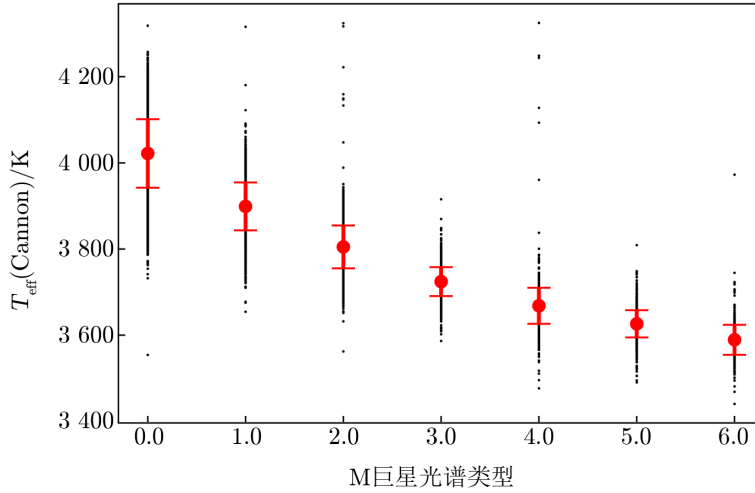


图 9 42151 颗 M 巨星 The Cannon 温度结果与光谱类型^[19]的对比

4 结 论

本文介绍了 The Cannon 的基本原理, 然后利用 The Cannon, 并结合 APOGEE 参数对 LAMOST 光谱中 K/M 巨星进行参数求解。由前人求解 K 巨星的结果可知, The Cannon 对于 LAMOST 光谱中 K 巨星的测量精度比传统的模板匹配方法 (LASP) 明显提高。此外, 我们对于 M 巨星参数求解的结果也进一步表明, The Cannon 的适用范围更广, 尤其对于传统方法很难给出参数测量值的低分辨率晚型星光谱, The Cannon 也能给出符合预期的参数测量结果。

The Cannon 的优势在于其不直接依赖于恒星物理模型来重构恒星光谱和推导恒星参数。由于使用了全谱流量信息来构建光谱模型, 很大程度上避免了由于某些谱线信噪比低导致参数测量精度下降等问题。该方法尤其适用于低分辨率光谱的大气参数求解, 把高分辨率光谱的参数测量结果与低分辨率光谱在观测深度和光谱数量上的优势有效地结合了起来, 进一步提高了低分辨率光谱的测量精度, 扩大了对不同类型恒星参数测量的适用范围, 具有十分广阔的应用前景。

从 The Cannon 的工作原理可知, 其适用范围完全受训练样本限制, 训练样本的参数范围和样本数目直接影响了测量光谱的参数范围和测量精度。一般而言, 训练样本的恒星参数来自于比待求解光谱样本具有相似或者更高分辨率的光谱测量结果。由于中高分辨率的光谱相对难以获得, 有效获取具有公共星观测的训练样本成为 The Cannon 在低分辨率光谱测量中进一步推广运用中遇到的最大困难。不过, 随着 20 ~ 30 m 级望远镜的建成使用, 未来 6 ~ 8 m 级望远镜将更多承担中高分辨率的光谱巡天工作, 大量的中高分辨率光谱将为 The Cannon 的应用提供更好的数据支持。

参考文献:

- [1] Lee Y S, Beers T C, Sivarani T, et al. AJ, 2008, 136: 2022
- [2] Lee Y S, Beers T C, Sivarani T, et al. AJ, 2008, 136: 2050
- [3] Liu C, Bailer-Jones C A L, Sordo R, et al. MNRAS, 2012, 426: 2463
- [4] Ting Y-S, Rix H-W, Conroy C. ApJ, 2017, 849: L9
- [5] Ting Y-S, Conroy C, Rix H-W, et al. arXiv: 1804.01530, 2018
- [6] Kos J, Bland-Hawthorn J, Freeman K, et al. MNRAS, 2018, 473: 4612
- [7] Ness M, Hogg D W, Rix H-W, et al. ApJ, 2015, 808: 16
- [8] Ho A Y Q, Ness M K, Hogg D W, et al. ApJ, 2017, 836: 5
- [9] Luo A-L, Zhao Y-H, Zhao G, et al. VizieR Online Data Catalog. <http://vizier.u-strasbg.fr/viz-bin/VizieR>, 2016
- [10] Liu C, Deng L-C, Carlin J L, et al. ApJ, 2014, 790: 110
- [11] Wu Y, Singh H P, Prugniel P, et al. A&A, 2011, 525: A71
- [12] Wu Y, Luo A-L, Li H-N, et al. Research in Astronomy and Astrophysics, 2011, 11: 924
- [13] Luo A-L, Zhao Y-H, Zhao G, et al. Research in Astronomy and Astrophysics, 2015, 15: 1095
- [14] Eisenstein D J, Weinberg D H, Agol E, et al. AJ, 2011, 142: 72

- [15] Majewski S R, Schiavon R P, Frinchaboy P M, et al. AJ, 2017, 154: 94
 [16] Alam S, Albareti F D, Allende Prieto C, et al. ApJS, 2015, 219: 12
 [17] Holtzman J A, Shetrone M, Johnson J A, et al. AJ, 2015, 150: 148
 [18] Zhong J, Lepine S, Li J, et al. VizieR Online Data Catalog (other). <http://vizier.u-strasbg.fr/viz-bin/VizieR>, 2015
 [19] Zhong J, Li J, Carlin J L, et al. ApJS, 2019, 385: 9

The Cannon: a Data Driving Approach to Stellar Label Determination

HUANG Yi-qi^{1,2}, ZHONG Jing¹, HOU Jin-liang^{1,2}

(1. Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: A new generation of large sky area spectroscopic survey project has produced nearly 10 million low-resolution stellar spectra. Based on these spectroscopic data, this paper introduced a machine learning algorithm named The Cannon. The algorithm is based on the training spectra of known stellar atmospheric parameters (effective temperature, surface gravity and metallicity) to establish the relation between spectral flux characteristics and stellar parameters. Then it is applied to the observed spectral data to calculate atmospheric parameters. The main advantage of The Cannon is that it is not based on any physical model of stars strictly and has a wider applicability. Moreover, because of the use of full spectrum information, even for spectra with low signal-to-noise ratio (SNR), it can obtain more reliable parameters. This algorithm has great advantages in data processing and parameter measuring of large quantity of stellar spectroscopy. In addition, this paper presents two examples of using The Cannon to obtain stellar parameters of K and M giants using the LAMOST spectra.

Key words: spectroscopic surveys; stellar parameters; data-driving