基于叠加的河外星系中性氢(HI)谱线数据处理方法

郭元旗^{1,2,3}, 朱明^{1,2*}, 艾美^{1,2}

1. 中国科学院国家天文台, 北京 100012;

2. 中国科学院FAST重点实验室, 北京 100012;

3. 中国科学院大学, 北京 100049

*联系人, E-mail: mz@nao.cas.cn

收稿日期: 2018-05-08; 接受日期: 2018-05-24; 网络出版日期: 2018-08-21

国家重点基础研究发展计划(编号: 2017YFA0402600)和国家自然科学基金(批准号: U1531246)资助项目

摘要 中性氢(HI)巡天是探测宇宙气体大尺度结构的重要手段. 在对河外星系进行HI巡天的过程中, 由于信号本身非常弱, 再加上望远镜灵敏度的限制, 导致观测数据中大量星系的HI谱线信号被湮没在噪声之中, 无法直接观测到. 但是利用谱线叠加技术就可以通过叠加一些已知光学红移和位置的星系来提高谱线的信噪比. 本文所要介绍的就是将一个星系群中所有成员星系的HI谱线进行叠加得到合成HI谱线的方法. 通过这种方法获得的叠加谱线就是该星系群中总的HI辐射. 叠加后的谱线由于rms水平的降低而获得比单一成员星系的谱线更高的信噪比, 从而使一些在原来单条谱线中并不明显的HI信号在叠加后的合成谱线中显现出来, 达到 $S/N > 4.5$ 的水平. 谱线叠加过程涉及了大量数据的批量处理, 因此我们开发了相关的自动化软件, 用来处理美国Arecibo望远镜ALFALFA中性氢巡天观测的数据, 成功得到了80多个选自SDSS观测星表的星系群样本的叠加谱线. 预计未来我国的FAST望远镜巡天将产生大量的中性氢谱线数据, 这一软件将会在大批量处理FAST河外星系HI谱线叠加方面发挥重要作用.

关键词 中性氢, 谱线, 叠加, 河外星系

PACS: 95.75.-z, 95.75.Pq, 95.80.+p, 98.38.Gt, 98.65.Cw

1 引言

氢是星际物质中含量最丰富的元素, 研究宇宙中氢元素的分布有助于我们更深刻地理解宇宙的大尺度结构及星系的演化过程. 星系群是成员星系个数比较少的星系团, 被认为是宇宙中一种中级演化状态的物体. 最近的观测研究表明, 星系群也缺乏HI气体^[1,2]. 在星系掉入星系团之前, 在星系群

环境中, 就已经被“提前处理”(Pre-Processing)了^[3-5]. 在星系群中HI星系已经开始分布在星系群边缘, 而且随着成员数的增加成员星系HI的缺乏越明显. 但是星系群中心一般是缺乏热气体的, 所以热气体剥离的模型并不能解释星系群HI的缺乏问题. 星系群中HI气体缺乏的原因, 星系形态的转变及演化是目前天体物理领域中最基本的科学问题之一.

目前世界上最大的河外中性氢巡天项目是

引用格式: 郭元旗, 朱明, 艾美. 基于叠加的河外星系中性氢(HI)谱线数据处理方法. 中国科学: 物理学 力学 天文学, 2018, 48: 119801
Guo Y Q, Zhu M, Ai M. Extragalactic HI spectra data processing method based on Stacking (in Chinese). Sci Sin-Phys Mech Astron, 2018, 48: 119801, doi: [10.1360/SSPMA2018-00181](https://doi.org/10.1360/SSPMA2018-00181)

Arecibo望远镜的ALFALFA(Arecibo Legacy Fast Arecibo L-band Feed Array)巡天^[6,7], 其巡天覆盖了约7000平方度的北部天区, 占全天区面积的1/5, 是目前最大的河外中性氢巡天项目, 其结果是现阶段最全面的北天区HI源表, 并且与光学巡天项目SDSS(Sloan Digital Sky Survey)^[8]的北天区有很好的重合, 使多波段交叉研究更为方便. 尽管ALFALFA是目前世界上范围最大的HI巡天, 但是由于其灵敏度的限制, 以及其科学目标的设定是搜寻最大天区范围的HI信号, 因此其巡天的深度就相对较浅. 在相对高红移的区域($z > 0.03$)中性氢的探测率较低.

由于HI谱线中心频率约为1420.4 GHz处于射电波段, 因此世界上很多射电望远镜(如Parkes, Arecibo, GBT等)都将中性氢巡天作为主要的科学目标. 已经建成并正在调试的500 m口径球面射电望远镜(FAST)是目前世界上最大的也是最先进的单口径射电望远镜, 它超高的灵敏度、灵活的天线面板及19波束接收机阵列将有利于观测更暗弱天文目标、搜索更大的天区范围, 能够对邻近宇宙中HI做出更完备的探测, 并获得高质量的天文观测数据, 对HI观测而言将大大提升目前HI巡天数据的质量. FAST的科学目标除了研究邻近宇宙的HI在宇宙中的分布之外, 还有一个非常重要的科学目标, 就是探测星系团/群尺度下的HI信号, 因为FAST波束相对较大(2.95角分)^[9], 在观测高红移的HI信号时波束之内可能含有不止一个星系, 很可能是星系群/团尺度的一团HI星系的叠加, 高灵敏度的特征使FAST具有探测世界上最高红移的HI星系团的能力^[10].

谱线叠加(Stacking)技术是将星系团(群)内每个成员星系的HI谱线按照一定方法进行叠加, 叠加后HI信号强度不变, 但是噪声会减弱. 它的基本原理就是利用统计学中随机变量叠加总体均方根减小的特点, 而HI信号为非随机信号, 叠加之后它们的值不会减小, 所以可以显现出来. 通过叠加方法能够更准确地估计星系团(群)的整体HI质量. 从巡天数据利用率方面讲, 实际上在河外HI巡天数据中大多数都是噪声信号, 需要探测的信号被湮没其中, 如果这些数据只用于提取探测到HI信号的谱线, 那么

利用率将会很低, 因此对其可利用价值的评估也是一个重要的研究方向, 特别是现在随着数据存储和计算能力的显著提升, 大数据的分析和挖掘也可以在这方面充分发挥作用. HI谱线叠加就属于一种对挖掘巡天数中有效信息的技术. 基于谱线叠加过程中的关键和重点部分, 文章将从谱线叠加样本的选取、叠加过程和结果及叠加的自动化3个方面来做详细介绍.

2 样本选取

在做谱线叠加之前首先需要准确测量星系群中每个成员星系中心的红移, 以保证在叠加时HI的中心是对齐的, 这样才能准确获得每个星系的HI含量. 本文中的样本选自SDSS观测的Mr18数据^[11,12], 从成员星表(<http://lss.phy.vanderbilt.edu/groups/dr7/>)中可以找到这些星系群的每个成员的星表数据, 见表1. 星表中提供了样本星系群每个成员的赤经、赤纬、红移等信息, 图1显示了这些星系群的红移速度区间, 横坐标的范围是7200–12800 km/s, 该范围是ALFALFA巡天数据的第3个(共4个)通道范围, 也是HI谱线最可能存在的速度区间; 纵坐标代表叠加星系群的序号, 颜色条内的数字代表星系群的成员数量, 从下至上与表2中左侧的数据一一对应.

图1 (网络版彩图)样本星系群的红移范围
Figure 1 (Color online) Redshift ranges of sample galaxy groups.

表 1 叠加样本星表

Table 1 Sample catalogue for stacking

GroupID	RA	DEC	z	Mr	g-r	fibcol	comp	r_edge	Cube name
712	18.786665	15.719010	0.03761	-18.559	0.335	0	0.949	0.23	gridbf_0116+15c.sav
712	18.718223	15.711505	0.03810	-18.041	0.750	0	0.949	0.25	gridbf_0116+15c.sav
712	18.570020	15.592507	0.03828	-20.654	0.915	0	0.949	0.50	gridbf_0116+15c.sav
7625	164.382846	8.298944	0.03498	-20.717	0.540	1	0.977	0.13	gridbf_1100+09c.sav
7625	164.500182	8.417871	0.03427	-19.412	0.363	0	0.977	0.16	gridbf_1100+09c.sav
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表 2 样本叠加结果

Table 2 Stacking results of the galaxy group samples

GroupID	All				Non-detections only			
	n	rms	S/N	Flux (mJy)	n	rms	S/N	Flux (mJy)
712	12	0.39	7.5	2.95788	8	0.46	4.86	2.243116
7625	9	0.49	7.72	3.746579	8	0.54	5.75	3.091659
12598	13	0.44	4.76	2.093297	10	0.48	2.67	1.277299
17438	12	0.46	5.04	2.33943	7	0.66	2.49	1.641643
20436	13	0.4	2.98	1.176955	11	0.42	1.69	0.713974
7700	18	0.38	4.25	1.598822	16	0.4	3.29	1.318477
12711	8	0.49	5.18	2.544294	4	0.73	3.34	2.452332
21006	8	0.52	4.99	2.59079	6	0.56	2.69	1.518187
18625	20	0.34	6.95	2.335173	16	0.4	4.31	1.739831
15033	18	0.32	5.52	1.79294	15	0.34	2.15	0.733977
17839	31	0.25	6.72	1.70882	22	0.35	5.04	1.748733
18627	35	0.18	9.48	1.702439	25	0.18	4.25	0.773212
19146	13	0.37	4.28	1.592881	10	0.38	3.27	1.256236
17951	9	0.49	7.38	3.631604	6	0.59	3.96	2.354738
17810	8	0.49	7.85	3.809288	5	0.77	4.67	3.615241
5424	24	0.35	5.07	1.788119	18	0.45	3.66	1.657524
11328	14	0.48	7.13	3.406174	10	0.52	4.85	2.501155
19207	8	0.47	3.96	1.858633	7	0.53	4.12	2.199398
19245	17	0.33	8.32	2.71346	10	0.4	4.76	1.921971
2263	16	0.44	4.12	1.815943	12	0.48	3.56	1.722382
20304	21	0.31	5.6	1.719694	13	0.43	2.81	1.212723
19166	22	0.29	7.89	2.297163	16	0.36	5.23	1.888763
7699	11	0.4	4.79	1.909557	8	0.45	2.75	1.243163
11438	9	0.4	6.46	2.598069	6	0.53	2.48	1.323128
1932	42	0.26	12.52	3.294856	32	0.31	3.32	1.021857
2128	8	0.53	6.47	3.452277	6	0.69	3.36	2.323686
16825	11	0.41	8.64	3.52137	7	0.4	6.58	2.610449
17813	19	0.49	3.66	1.792725	18	0.44	3.93	1.724891
18636	148	0.09	10.97	0.981177	126	0.09	5.26	0.487228
11261	15	0.37	7.21	2.655709	12	0.31	6.32	1.988306
7814	22	0.33	6.71	2.215841	20	0.33	5.11	1.677576
20024	19	0.33	6.02	1.994175	15	0.36	3.98	1.427598
12418	5	0.95	2.45	2.329706	5	0.95	2.45	2.329706
20211	28	0.28	5.54	1.566638	21	0.35	3.82	1.332446
22080	9	0.45	4.32	1.92699	7	0.54	2.81	1.509111
12537	15	0.38	3.2	1.223588	13	0.41	3.63	1.475935
20728	8	0.54	2.46	1.339878	5	0.53	2.95	1.550805
4068	8	0.49	4.53	2.228721	7	0.51	2.97	1.505672
9428	10	0.35	3.48	1.217501	8	0.44	3.11	1.376402
12133	24	0.31	4.09	1.273838	20	0.32	4.07	1.288244
17470	24	0.25	5.8	1.451648	20	0.27	4.07	1.086918
2332	48	0.24	3.53	0.855237	45	0.26	2.83	0.741343

续表2

groupID	All				Non-detections only			
	<i>n</i>	rms	S/N	Flux (mJy)	<i>n</i>	rms	S/N	Flux (mJy)
7633	9	0.37	6.43	2.351087	8	0.42	3.11	1.30571
10707	13	0.37	3.63	1.338088	10	0.4	2.06	0.82748
10587	24	0.3	2.68	0.810906	20	0.32	2.41	0.771191
18257	9	0.6	3.62	2.171828	8	0.72	2.85	2.042455
20153	21	0.41	3.42	1.398543	18	0.37	3.3	1.21079
20157	9	0.53	5.63	2.959677	4	0.68	5	3.408736
22019	23	0.27	8.07	2.17156	20	0.28	7.28	2.045453
9146	28	0.28	5.18	1.469565	24	0.29	3.15	0.905679
20036	11	0.37	5.45	2.042352	7	0.5	2.16	1.078041
3620	19	0.39	4.51	1.743309	16	0.44	4.63	2.013705
265	79	0.27	2.07	0.547808	45	0.27	3.27	0.892152
11270	9	0.53	3.02	1.597493	6	0.5	3.15	1.571024
10716	8	0.48	2.81	1.345636	4	0.69	2.52	1.725082
13702	13	0.39	4	1.560439	12	0.41	3.16	1.282947
19258	35	0.22	2.99	0.646485	33	0.22	2.45	0.544517
11271	103	0.17	5.79	0.9888	97	0.16	4.97	0.797373
19958	14	0.46	4.42	2.024507	13	0.52	3.13	1.62618
2336	49	0.2	6.06	1.205255	39	0.23	5.76	1.32008
11464	18	0.37	4.19	1.563987	17	0.36	4.18	1.52288
17454	56	0.23	3.21	0.736183	49	0.26	1.98	0.51067
2268	14	0.45	4.19	1.874693	13	0.47	3.32	1.558657
464	21	0.37	4.65	1.713174	11	0.64	3.32	2.130146
17450	54	0.21	4.79	0.991786	39	0.21	2.98	0.629373
9168	15	0.3	5.76	1.7525	15	0.3	5.76	1.7525
2005	59	0.21	3.04	0.645955	57	0.22	2.74	0.600717
12515	62	0.38	2.02	0.76618	52	0.25	2.31	0.578949
19709	8	0.56	3.52	1.963122	7	0.59	3.36	1.968304
9500	21	0.4	3.09	1.223215	20	0.4	2.78	1.115509
16375	26	0.28	2.31	0.648369	25	0.28	2.31	0.65576
7621	9	0.48	3.58	1.709358	8	0.6	3.34	1.991474
11339	27	0.31	4.17	1.274169	25	0.32	3.85	1.249572
19710	32	0.23	4.64	1.088107	28	0.28	2.7	0.741813
2760	70	0.2	2.63	0.537588	67	0.21	2.16	0.449184
9505	11	0.42	4.97	2.073069	9	0.43	3.7	1.591577
7335	10	0.47	4.44	2.067779	7	0.52	3.57	1.860524
13680	18	0.27	5.44	1.487734	16	0.3	5.06	1.503051
13622	13	0.48	2.41	1.162217	10	0.44	2.42	1.064289
17843	54	0.2	4.66	0.955085	54	0.2	4.66	0.955085
13870	25	0.31	2.97	0.911998	24	0.32	2.23	0.706298
9530	20	0.37	2.2	0.82291	20	0.37	2.2	0.82291
5447	37	0.32	2.33	0.75049	34	0.32	3.11	1.002576
1910	44	0.2	5.11	1.042256	44	0.2	5.11	1.042256

可以看到所选的80多个样本星系群的红移区间都在这一速度范围内基本均匀分布, 具有较好的代表性.

3 谱线叠加过程

叠加谱线是从ALFALFA射电巡天^[6,13]数据中提取的. 该巡天数据采用IDL(Interactive Data Language, <http://www.harrisgeospatial.com/ProductsandTechnology/Software/IDL.aspx>)的程序文件发布类型—SAV格式. 其中每个SAV文件包含了天空中约 $2.4^\circ \times 2.4^\circ$ (144×144 像素)的天区范围、1024个频率通道、两个偏振方向的数据块, 以及其他一些校准信息. SAV文件名中包含了数据块中心位置的赤经赤纬信息, 如“gridbf_1532+07c.sav”, 其中“1532”和“07”分别表示赤经15时32分和赤纬 07° 的位置, “gridbf”代表数据经过了基线和平场处理, “c”表示该数据块的通道范围是2048–3071 (7200–12800 m/s).

3.1 谱线提取

谱线提取是指从射电望远镜观测数据中提取特定位置的谱线数据. 首先需要将每个星系赤经、赤纬的坐标位置转换为相应SAV数据块中的像素位置, 这里确定每个星系所在的数据块是关键步骤, 将在第3节做详细介绍. 每条谱线需要选择多少像素点由像素点的分辨率和目标源的大小决定. 根据Arecibo望远镜多波束的排列及扫描方式^[6]可知, 它记录的数据经过grid处理之后每像素分辨率约为 $1' \times 1'$, 且每个像素至少经过主波束约 $3.3' \times 3.8'$ 的椭圆天区范围, 因此它的原始数据为每个波束内实测的流量值, 并没有对每个波束亮度衰减^[14]做出修正. 然而由于我们的样本星系的轮廓范围基本都在10个角秒的量级, 远小于望远镜的波束范围, 需要计算每一个像素的实际流量值. 为了消除望远镜波束存在的衰减效应, 首先要根据望远镜波束在天空中的亮度分布 B 计算目标源附近各个像素点的亮度, Arecibo望远镜波束的亮度分布可用以下公式表示:

$$B(x, y) = e^{\left[-\frac{1}{2}\left(\frac{x}{\sigma_x}\right)^2 - \frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right]}, \quad (1)$$

其中, $x = y = 1'$, $\sigma_x = (2\sqrt{2\ln 2})^{-1} \times 3.3'$, $\sigma_y = (2\sqrt{2\ln 2})^{-1} \times 3.8'$.

根据Arecibo波束的大小, 我们选择了16个像素组成的正方形天区作为一个星系的范围, 将它们的实测积分流量密度^[15] $S_v(x, y)$ 进行求和, 同时将这16个像素对应位置的望远镜亮度值进行求和, 这样就可以得到目标位置像素点的真实流量值:

$$S_v = \frac{\sum_{-1}^1 \sum_{-1}^1 S_v(x, y)}{\sum_{-1}^1 \sum_{-1}^1 B(x, y)}. \quad (2)$$

以上是一个通道的流量密度计算方法, 同样的方法计算出所有通道的流量密度, 就可以得到关于星系视向速度(km/s)和流量密度(mJy)的HI谱线. 对于有两个极化方向的数据而言, 将两个极化方向上的谱线分别提取, 并将两条谱线的平均值作为最终提取的一条谱线.

3.2 谱线叠加方法

因为每个星系群中各个星系的红移一般不同, 见图1, 所以谱线叠加首先要解决的问题就是消除叠加谱线之间的速度差, 使每条谱线的HI中心处于速度为0 km/s的位置, 这样才能使相应位置的HI谱线得到最大限度的叠加, 也最大限度的消除了噪声的影响.

叠加的基本方法(以一个星系群为例):

1) 根据每个星系的光学红移值算出每条谱线HI中心所在位置的速度;

2) 去除1)中计算的所有谱线的红移速度, 也就是将HI中心位置的速度都平移至0 km/s, 如图2横坐标所示;

3) 以2)计算的谱线速度为准, 将所有经过平移的每条谱线的最小值(负值)称为该谱线的“左端”, 最大值(正值)称为该谱线的“右端”. 选所有左端中的最大值定义为叠加后新谱线的左端(最小值)、所有右端中的最小值定义为叠加后新谱线的右端(最大值);

4) 将每条谱线按照3)中定义的新谱线的左端和右端进行截取, 去除多余部分, 则每条谱线剩余的长

图 2 某个星系群的成员示例星系谱线. 其中外侧的横坐标显示的是谱线的原始(日心)速度, 其范围是固定的7200–12800 km/s; 内侧的横坐标是以HI中心的位置速度为0 km/s重新标定的. 纵坐标是谱线的流量值, 用mJy表示. (a) 某个星系的谱线, 标题的groupID和n分别指星系群的编号和其成员的编号; (b) 叠加谱线, 图中分别显示了目标谱线数量、rms和信噪比; (c) 谱线(b)经过平滑处理后的结果, steps表示平滑通道的个数

Figure 2 The spectrum of an example galaxy. The outside x -axis is original heliocentric velocity in km/s, which ranges from 7200 to 12800 km/s. The inside x -axis is the normalized velocity, of which the center of the HI spectrum is set to 0 km/s. The y -axis is the flux density in mJy. (a) The HI profile of the example galaxy, the title “groupID” and “ n ” represent the ID of the group and the galaxy’s member number in the group respectively; (b) the stacked spectrum, with the numbers of stacked objects, rms noise level and signal-to-noise ratio listed at the upper right corner; (c) the HI profile after smoothing from the profile in (b), “steps” represents the number of channels smoothed.

度再次相同, 并且每条谱线的HI所在的位置也是对齐的;

5) 按照式(3)和(4)叠加所有通道, 最后得到完整的叠加谱线, 如图2(b).

由于每条谱线的rms不同, 对叠加谱线的贡献量也不同, 因此需要计算每条谱线的权重 w :

$$w = \frac{1}{\text{rms}^2}. \quad (3)$$

叠加谱线中某点的流量值可以表示为

$$S = \frac{\sum_{i=0}^N S_i \cdot w_i}{\sum_{i=0}^N w_i}, \quad (4)$$

其中, $i = 1, 2, \dots, N$ 为一次叠加的谱线条数. 根据式(4)可以计算出叠加谱线的每一个通道的流量.

谱线平滑是指将相邻的若干通道流量求平均

值,从而达到进一步降低rms,提高谱线信噪比的目的.图2(c)是图2(b)经过5个通道的平滑之后得到的,可以看到谱线的rms明显降低了、信噪比明显提高了.需要说明的是虽然谱线平滑会降低谱线的分辨率,使谱线失去部分细节,但是谱线叠加更多关注的是叠加后星系群整体的HI轮廓,适当的牺牲分辨率并不会影响叠加结果.对于ALFALFA巡天数据而言,其通道分辨率约为5.5 km/s,数据平滑时一般不超过50 km/s,也就是最多平滑至9个通道.

3.3 叠加结果及分析

图3显示了部分叠加结果,并且已经过平滑处理,其中左侧是一个星系群中所有星系叠加的结果,右侧是一个星系群中ALFALFA未探测到的星系叠加的结果.图中谱线平滑的通道根据HI的宽度和通道分辨率来确定.从图中右侧叠加结果可以看出,即使用ALFALFA没有探测到HI的谱线进行叠加,仍然有HI信号出现,而且目标星系数量越多,叠加后得到的谱线信噪比也相对越高.左右两侧谱线的对比可以更明显地看出谱线叠加的效果.所有样本星系群的叠加结果及分析见表2及第4节.

4 谱线叠加过程的自动化

要想通过谱线叠加获得较好的HI谱线轮廓,就需要进行大量的谱线叠加,一次叠加至少要上百条谱线,才会有较满意的结果^[16],因此叠加过程需要实现自动化处理.另外,如果需要对大量的星系群或星系团的谱线进行叠加,也需要做批量处理的工作.

在确定了需要叠加的星系群(团)之后,首先需要根据星系群(团)的ID从成员星表中提取出所有星系群(团)的星表信息,组成叠加样本星系表,即表1中前9列的内容.

有了叠加样本星表之后,下一步要做的就是将星表中的每个星系与其所在的星系巡天数据块对应起来.在第2节中提到数据块的名字中包含其中心位置的赤经、赤纬信息,每个数据块的大小又是由144×144个像素组成,因此可以通过分别遍历星

系位置和数据块名的方法找到每个星系与它所在的数据块,将数据块名加入叠加样本星表,如表1中第10列所示.这样在提取某个星系(或星系群)的谱线时只需要载入相应的数据块即可,而不需要将整个叠加天区的数据都加载,从而可以灵活地对任意星系群的谱线进行叠加.

实现谱线叠加过程的自动化包括对谱线叠加的自动化和对星系群处理的批量化两个步骤.其中谱线叠加的自动化就是按照前面介绍的谱线叠加方法来实现自动提取谱线、对齐HI、计算和叠加流量等;对星系群的批量化处理过程是将表1作为输入星表,并实时载入巡天数据文件,通过程序循环来控制.谱线叠加的自动化流程如图4.

图4中 m 是人为指定的叠加星系的起始位置,即表1中任意一个星系群的起始位置, n 是当前叠加星系群的成员个数, e 是人为指定的叠加星系的结束位置.首先进行当前星系群的叠加并显示相应结果,之后判断是否需要继续叠加,如果 $m+n < e$,则 $m = m + n$,即跳转到下一个星系群的起始位置继续进行下一个星系群的叠加并显示相应结果,当 $m+n \geq e$ 则叠加完成,退出循环,最后显示如图3和图5等结果信息.从流程图中可以看出谱线叠加自动化过程中需要有两次外部数据的导入,一次是叠加星表数据,另一次是巡天数据.需要注意的是后者数据导入过程较慢,应避免在循环中多次导入相同的数据,以免影响程序执行速度.

以上实现自动化叠加谱线的关键步骤在于找到每条谱线所在的数据块.由于ALFALFA的数据块并不是严格地按照 $2.4^\circ \times 2.4^\circ$ 来划分的,而且有时候数据块名中的赤经、赤纬也并不是严格的数据块中心,这就导致用数据块名来判断数据的位置会出现边缘数据判断不准的情况,如果每条谱线按9个或者16个像素来提取,那么就有很大概率无法从一个数据块中提取一条完整的谱线,从而给数据提取过程带来困难.数据块分割是望远镜巡天数据处理过程中重要的一步,数据块的分割有利于数据的存储、传输及发布.但数据块分割的好坏将直接影响到后期对其中数据提取的难易程度,因此在做分割算法的时候在硬件条件允许的基础上应当充分考虑到后期使用的可行性和便利性.

图 3 三组不同成员数量星系群的谱线叠加结果对比. (a) 所有星系群成员的谱线叠加结果; (b) 用ALFALFA中未探测到HI的星系群成员的谱线叠加结果. 每条谱线分别显示了叠加谱线的群ID、平滑通道数、成员星系数量、rms及信噪比

Figure 3 The comparison of the stacked spectra of three different groups. The left column is the stacked spectra using all galaxies in a group, the right column is the stacked spectra using only those galaxies that were not detected by ALFALFA. For each spectrum, the group ID, smoothing steps, object numbers, rms noise level and S/N ratio are listed.

图4 自动化谱线叠加过程的流程图

Figure 4 The flow chart of the automated stacking process for HI spectral data.

在观测数据方面, FAST相对于美国Arecibo来说在像素分辨率和通道数量上都有极大提高, 因此可以做更精确更合理的数据块分割; 同时我们将针对FAST巡天数据的每个天空位置和其所在的数据块提供相关的查找算法, 确保数据定位尤其是数据块边缘位置的准确性, 为后期的数据提取提供便利。

5 谱线均方根(rms)的计算及验证

在计算rms时需要避开谱线中有HI信号和干扰

信号的通道, 取谱线中剩余通道的值来计算. 对于单条谱线来说, 它的计算公式为

$$\text{rms} = \sqrt{\frac{1}{M}(F_0^2 + F_1^2 + F_2^2 + \dots + F_{m-1}^2)}, \quad (5)$$

其中, M 代表计算时在每条谱线中所选的通道数, F 是指流量值.

对于不加权的 N 条谱线叠加后, 它的rms值应该满足统计关系:

$$\text{rms}_N = \frac{\overline{\text{rms}}}{\sqrt{N}}, \quad (6)$$

其中, $\overline{\text{rms}}$ 是所有样本星系群的平均rms.

图5是所有样本星系群的rms与叠加谱线条数关系图, 红色虚线 $\overline{\text{rms}}/\sqrt{N}$ 是随机噪声与叠加数量 N 的理论值, “+”是由式(5)计算的每个星系群rms的真实值. 从图中可以看出除了个别值偏差较大外, 大部分真实值与理论值很接近, 这就说明在谱线叠加过程中, 随机噪声是满足随着叠加数量 N 的增加呈 $1/\sqrt{N}$ 减少的, 为叠加方法提供了理论依据.

当然, 随机噪声并不是随着谱线数目的叠加呈线性减小关系, 随着叠加到了一定数目以后, 由于观测设备精度不够所产生的其他非随机噪声的影响会越来越明显, 导致rms减小的速度会变慢, 具体的影响因素不是本文所要研究的内容, 这里不再赘述.

图5 (网络版彩图)rms值与叠加谱线条数的关系

Figure 5 (Color online) The relation between the rms of the stacked spectra and the number of co-added objects.

表 2 是所有样本星系群的叠加结果, 其中第 1 列是每个星系群的 ID 编号, 后面 8 列分为两组, 分别是该星系群中所有成员的叠加结果和 ALFALFA 观测数据中未探测到的成员的叠加结果, 两组分别显示了群成员数量 n , rms, S/N 及 flux 作为对照, 其中 flux 是指信号的峰值流量, S/N 由 flux/rms 而来. 需要说明的是所有这些数据都是叠加之后对谱线平滑 5 个通道得到的, 但是对每个星系群做相同的平滑不一定能得到信噪比最高的谱线, 而且对于未探测到谱线的标定也是人为的, 因此表 2 中右侧的结果不一定能说明随着叠加数量的增加得到的谱线 HI 信噪比一定增加的规律.

6 结论

从叠加样本的选取、叠加过程的实现及叠加过程的自动化 3 个方面详细介绍了 HI 谱线数据叠加技

术. 叠加结果显示谱线叠加的效果显著, 是可以用于分析星系群整体 HI 含量的. 谱线叠加技术作为一种射电天文数据处理手段的延伸, 随着射电天文观测设备精度的提高、数据量的增加以及计算处理能力的提升, 将有助于更精确地计算星系群(团)的 HI 含量, 从而更清晰地认识 HI 在宇宙中的分布规律, 为相关理论研究和预测提供帮助. 同时, 谱线叠加技术弥补了 ALFALFA 释放的 HI 源表对光学星系探测率低的问题. 通过 FAST 巡天我们有望得到一个全天区、大样本的星系群的准确的 HI 质量, 结合演化状态的相关性研究, 将可以得到星系群 HI 气体的演化规律. 通过分析 HI 质量大的星系群的特征, 可以为将来 FAST 探测高红移星系团尺度 HI 信号做准备.

未来我们也会把谱线叠加技术融入到 FAST 谱线数据处理流程之中, 并且要实现谱线叠加的实时化和批量化, 从而扩展 FAST 谱线数据处理能力, 进一步提高对 HI 巡天数据的处理能力.

致谢 感谢中国科学院 FAST 重点实验室开放课题项目的支持, 同时也感谢 Martha Haynes 教授提供 SDSS 星系群样本的 ALFALFA 观测数据.

参考文献

- 1 Mok A, Wilson C D, Golding J, et al. The JCMT nearby galaxies legacy survey—X. Environmental effects on the molecular gas and star formation properties of spiral galaxies. *Mon Not R Astron Soc*, 2016, 456: 4384–4406
- 2 Dressler A. A catalog of morphological types in 55 rich clusters of galaxies. *Astrophys J Suppl S*, 1980, 42: 565–609
- 3 Balogh M L, Morris S L, Yee H K C, et al. Star formation in cluster galaxies at $0.2 < z < 0.55$. *Astrophys J*, 1997, 488: L75–L78
- 4 Zabludoff A I, Mulchaey J S. The properties of poor groups of galaxies. I. Spectroscopic survey and results. *Astrophys J*, 1998, 496: 39–72
- 5 Mahajan S. Evidence for galaxies being pre-processed before accreted into clusters. *Mon Not R Astron Soc-Lett*, 2013, 431: L117–L121
- 6 Giovanelli R, Haynes M P, Kent B R, et al. The Arecibo Legacy Fast ALFA survey. I. Science goals, survey design, and strategy. *Astron J*, 2005, 130: 2598–2612
- 7 Haynes M P, Giovanelli R, Martin A M, et al. The Arecibo Legacy Fast ALFA survey: The α .40 HI source catalog, its characteristics and their impact on the derivation of the HI mass function. *Astron J*, 2011, 142: 170
- 8 York D G, Adelman J, Anderson Jr. J E, et al. The Sloan Digital Sky Survey: Technical summary. *Astron J*, 2000, 120: 1579–1587
- 9 Nan R, Li D, Jin C, et al. The Five-Hundred Aperture Spherical Radio Telescope (FAST) project. *Int J Mod Phys D*, 2011, 20: 989–1024
- 10 Ai M, Zhu M, Fu J. Integrated HI emission in galaxy groups and clusters. *Res Astron Astrophys*, 2017, 17: 101
- 11 Berlind A A, Frieman J, Weinberg D H, et al. Percolation Galaxy Groups and Clusters in the SDSS Redshift Survey: Identification, Catalogs, and the Multiplicity Function. *Astrophys J Suppl S*, 2006, 167: 1–25
- 12 Mei A, Ming Z. Evolution of galaxy types and H I gas contents in galaxy groups. *Astrophys J*, 2018, 862: 48
- 13 Haynes M P, ALFALFA Team. The Arecibo Legacy Fast ALFA (ALFALFA) extragalactic HI survey. *AIP Conf Proc*, 2008, 1035: 238–241
- 14 Shostak G S, Allen R J. Accurate HI profiles of several nearby galaxies. *Astron Astrophys*, 1980, 81: 167–171
- 15 Condon J J, Ransom S M. *Essential Radio Astronomy*. Princeton: Princeton University Press, 2016. 23–126
- 16 Fabello S, Catinella B, Giovanelli R, et al. ALFALFA H I data stacking - I. Does the bulge quench ongoing star formation in early-type galaxies? *Mon Not R Astron Soc*, 2011, 411: 993–1012

Extragalactic HI spectra data processing method based on Stacking

GUO YuanQi^{1,2,3}, ZHU Ming^{1,2*} & AI Mei^{1,2}

¹*National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China;*

²*CAS Key Laboratory of FAST, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China;*

³*University of Chinese Academy of Sciences, Beijing 100049, China*

Blind HI survey is a powerful tool for exploring the HI universe. During the extragalactic HI survey, due to the weakness of HI signal and the limitation of telescope sensibility, most HI emissions from galaxies are overwhelmed by noise signals and cannot be directly detected. However, using the technique of spectral stacking, we can increase the the signal-to-noise ratio (SNR) by co-adding a set of spectra if the optical redshifts and position are known. This paper presents a method to derive the integrated HI emission in a galaxy group by stacking all the HI spectra of the member galaxies in the group. This method can generate a stacked spectrum representing the total HI emission of the group. The stacked spectrum have a lower rms noise level and thus higher S/N ratio than that of the individual member galaxies. In some cases, the integrated HI emission shows clear detection ($S/N > 4.5$) in the stacked spectrum even though no detection is visible in the spectrum of individual member galaxy in the group. We have also developed a software code to automatically handle the stacking of large amount spectra in galaxy groups. As an example, we use this software to derive the stacked HI spectra of 80 galaxy groups selected from the SDSS galaxy group catalog, of which the HI data are from the ALFALFA HI survey. Such software will be a valuable tool to process the stacking of large mount of HI spectra from the future extragalactic HI survey with the Five hundred meter Aperture Spherical radio Telescope (FAST).

HI, spectra line, stacking, extragalactic systems

PACS: 95.75.-z, 95.75.Pq, 95.80.+p, 98.38.Gt, 98.65.Cw

doi: [10.1360/SSPMA2018-00181](https://doi.org/10.1360/SSPMA2018-00181)