Publications of the Astronomical Society of Australia (2022), **39**, e050, 12 pages
doi:10.1017/pasa.2022.47

Research Article

Baseline correction for FAST radio recombination lines: A modified penalised least squares smoothing technique

Bin Liu[1], Lixin Wang[2], Junzhi Wang[3], Bo Peng[1] and Hongjun Wang[2]

[1]CAS Key Laboratory of FAST, National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Road, Beijing 100101, People's Republic of China, [2]Shaanxi University of Science and Technology, Weiyang University Park, Xi'an 710021, People's Republic of China and [3]Guangxi Key Laboratory for Relativistic Astrophysics, School of Physical Science and Technology, Guangxi University, Nanning 530004, People's Republic of China

Abstract

A pilot project has been proceeded to map 1 deg^2 on the Galactic plane for radio recombination lines (RRLs) using the Five-hundred-metre Aperture Spherical Telescope (FAST). The motivation is to verify the techniques and reliabilities for a large-scale Galactic plane RRL survey with FAST aiming to investigate the ionised environment in the Galaxy. The data shows that the bandpass of the FAST 19 beam L-band is severely affected by radio frequency interferences and standing wave ripples, which can hardly be corrected by traditional low order polynomials. In this paper, we investigate a series of penalised least square (PLS) based baseline correction methods for radio astronomical spectra that usually contain weak signals with high level of noise. Three promising penalised least squares based methods, AsLS, arPLS, and asPLS are evaluated. Adopting their advantages, a modified method named rrlPLS is developed to optimise the baseline fitting to our RRL spectra. To check their effectiveness, the four methods are tested by simulations and further verified using observed data sets. It turns out that the rrlPLS method, with optimised parameter $\lambda = 2 \times 10^8$, reveals the most sensitive and reliable emission features in the RRL map. By injecting artificial line profiles into the real data cube, a further evaluation of profile distortion is conducted for rrlPLS. Comparing to simulated signals, the processed lines with low signal-to-noise ratio are less affected, of which the uncertainties are mainly caused by the rms noise. The rrlPLS method will be applied for baseline correction in future data processing pipeline of FAST RRL survey. Configured with proper parameters, the rrlPLS technique verified in this work may also be used for other spectroscopy projects.

Keywords: radio lines: ISM – methods: data analysis – surveys – ISM: clouds

(Received 15 February 2022; revised 12 September 2022; accepted 18 September 2022)

1. Introduction

The current world largest single dish radio telescope, Five-hundred-metre Aperture Spherical Telescope (FAST, Qiu 1998; Nan et al. 2011; Qian et al. 2020), was built in late 2016, and started fully operating in early 2020. Large single dishes have been proven dominantly in observations of radio recombination lines (RRLs) tracing ionised interstellar medium in the Galaxy (Alves et al. 2015; Liu et al. 2019; Anderson et al. 2021, etc.). With its unprecedentedly high sensitivity, FAST has great potential to study diffuse ionised gases along the Galactic plane using RRLs. Using the FAST 19 beam L-band receiving system, a pilot observation has been made to image the Galactic plane with RRLs, from which the data will be used to verify the techniques and reliabilities for a large-scale Galactic RRL survey.

The modern developments of electronic devices and wireless communication technologies have made the microwave environment more and more lousy for radio telescopes. Therefore, radio spectroscopy observations are often contaminated by radio frequency interferences (RFIs) and baseline problems, especially in centimetre wavelength. Although some frequency ranges are protected for astronomical studies, such as the 21 cm neutral hydrogen (H I) line around 1420 MHz, there are hardly protections for studies which need a wide frequency coverage. A typical example is the observations of RRLs, whose line rest frequencies cover the entire radio frequency range from ∼100 GHz to ∼100 MHz.

Baseline removal is an essential preprocessing step for spectral data analysis. The purpose is to remove the artificial baseline structure caused by the electronics or broad RFI features, and to retain the astronomical signal unaffected. A common way of baseline estimation is to perform a low order ($\leq$3) polynomial least square fitting. Since our observations were made during the early operation of FAST, the frequency bandpass were not ideally clean and flat. Many RRL spectral segments were affected by RFIs and baseline ripples, to which the low order polynomial baseline fitting are mostly in vain. New method of baseline correction is then indispensable before the line profiles to be accurately fitted.

Based on penalised least squares (PLS) smoothing technique, baseline correction methods have been developed and applied in Raman and infrared spectroscopic analysis. The basic idea of PLS is to balance between fidelity to the original data and the roughness of the fitted baseline by combining least squares smoothing together with a penalty on roughness of an estimation. The PLS algorithm for baseline correction was first introduced by Eilers (2003, 2004) and named as asymmetric least squares (AsLS).

Corresponding authors: Bin Liu, email: bliu@nao.cas.cn; Junzhi Wang, email: junzhiwang@gxu.edu.cn; Bo Peng, email: pb@nao.cas.cn

Cite this article: Liu B, Wang L, Wang J, Peng B, and Wang H. (2022) Baseline correction for FAST radio recombination lines: A modified penalised least squares smoothing technique. *Publications of the Astronomical Society of Australia* **39**, e050, 1–12. https://doi.org/10.1017/pasa.2022.47

Table 1. The sky coverage for the 1 deg^2 RRL mapping.

Observing area	Values
Map centre (Gal)	$l = 34\overset{\circ}{.}5, b = 0\overset{\circ}{.}0$
Map centre (Equ)	$18^{\rm h}54^{\rm m}17\overset{\rm s}{.}95 + 01^{\circ}23'43\overset{''}{.}9$
RA range (J2000)	$18^{\rm h}52^{\rm m}42\overset{\rm s}{.}58 \sim 18^{\rm h}56^{\rm m}16\overset{\rm s}{.}26$
Dec range	$+00^{\circ}35'24\overset{''}{.}5 \sim +01^{\circ}56'10\overset{''}{.}0$
OFF position	$18^{\rm h}48^{\rm m}02\overset{\rm s}{.}57 + 01^{\circ}48'44\overset{''}{.}2$

Table 2. The backend configuration.

Parameters	Values
Targeted RRLs (Hnα)	H165α $-$ H184α
Frequency range	1050 $-$ 1450 MHz
Digital bandwidth	500 MHz
Number of channels	2^{20}(1 M)
Frequency resolution	0.478 kHz
Velocity resolution range	0.099 $-$ 0.137 km s^{-1}
Integration per sample	1 s

To improve the results of baseline correction, several modified methods inspired by AsLS have been developed subsequently, they are: 1) adaptive iteratively reweighted penalised least squares (airPLS, Zhang et al. 2010); 2) improved asymmetric least squares (IAsLS, He et al. 2014); 3) asymmetrically reweighted penalised least squares (arPLS, Baek et al. 2015); and 4) adaptive smoothness parameter penalised least squares (asPLS, Zhang et al. 2020b).

In this paper, we focus on the application and evaluation of the PLS-based baseline fitting algorithms applied to the RRL spectra obtained with FAST. Section 2 describes the RRL observations made by FAST and the data reduction pipeline for the spectral line imaging. Section 3 reviews the theory of the existing PLS-based algorithms and introduces our modified method for FAST RRL data, rrlPLS. Section 4 presents the simulation work, where the AsLS, arPLS, asPLS, and rrlPLS are evaluated using simulated data set and the optimised parameters are listed. Their applications to actual observed data are shown in Section 5. In Section 6, we verify the rrlPLS method using the real data cube with artificial line profiles injected. Conclusions are given in Section 7.

2. RRL observation and data reduction using FAST

2.1. *The RRL observation*

The pilot project covers a field of 1 deg^2 along the Galactic plane, which was observed using the FAST Multi-Beam On-The-Fly (MBOTF) mode. This field centring at $l = 34\overset{\circ}{.}5, b = 0\overset{\circ}{.}0$ is chosen for it contains active star-formation regions, thus intensive RRLs both from discrete H II regions and diffuse ionised gas are expected. A reference position off the Galactic plane is adopted for bandpass calibration. Table 1 gives the detailed information of the targeted region.

MBOTF observations are deployed in the Equatorial system. We scan the targeted region twice in each session, along RA and Dec axis respectively, with a scan speed of ~33″ s^{-1}. The offset position was observed for 5 min before and after each MBOTF mapping and a flux calibrator was observed at the beginning to confirm the power stability of the noise diode during different sessions. Flux calibration was done adopting the temperature of the noise diode provided by the FAST official website. To summarise, the observing procedure for each session is: 1) flux calibrator; 2) offset position; 3) MBOTF in RA; 4) offset position; 5) MBOTF in Dec; 6) offset position.

Table 2 lists the backend configuration. The frequency bandpass of the FAST L-band is from 1050 to 1450 MHz, which covers twenty hydrogen α−RRLs from H165α to H184α. The spectrometer records one spectrum per second which covers a digital bandwidth of 500 MHz with 2^{20} channels resulting the frequency resolution of ~0.478 kHz. The corresponding velocity resolutions of the twenty RRL segments are from 0.099 to 0.137 km s^{-1}.

2.2. *Data reduction*

A data reduction pipeline has been developed to process the FAST spectra from MBOTF observations. Three major steps are applied including radio frequency interference (RFI) excision, calibration, and baseline removal. After calibration, the full bandpass are cut into individual RRL segments, to which the baseline removal is deployed. The system properties adopted for calibration are given in Jiang et al. (2020).

The frequency channels affected by strong and broad RFIs, which may come from satellites, ground radar or communication stations, are firstly flagged out. To excise weak, narrow, and transitory RFIs, a median absolute deviation filter is applied (Liu et al. 2019), with a window width of 25 channels and intensity threshold above 3 times of the spectral rms.

The bandpass of the FAST 19 beam L-band receiver is affected by standing wave ripples with a typical width of ~100 km s^{-1} (see Jiang et al. 2020). Figure 1 shows the averaged baseline of the twenty segments over 60 s. We show the averaged spectra in Figure 1 only for a better illustration of the baseline features. In the pipeline, the baseline removal was applied to the raw spectrum with 1 s integral time. Automatic polynomial or sinusoid fitting could not deal with such unstable baseline situations.

As a test, AsLS was applied in the pipeline, which was the first PLS-based methods originally developed for baseline correction in Chemistry and Raman spectroscopy (Eilers 2004; Peng et al. 2010; Zhang et al. 2020a). Differing from its original application, where both the baseline ripple and the spectral line intensity are strong while the noise are negligible, in our data the baseline ripples and the noise are significant but the spectral line signals are usually weak. The red lines in Figure 1 illustrate the result of the AsLS test. For our pipeline, a optimised PLS-based method, rrlPLS, was finally adopted, which is introduced in Section 3.4.

The spectra of individual RRL are spatially re-sampled and grided into data cube with 1’ pixel size (~1/3 beam size), with a Gaussian kernel following the instruction given by Mangum et al. (2007). In each observing session, one data cube is created for each RRL segment from the combined data sets of the two MBOTF scans. The cubes for the same RRL segment from different sessions are then averaged. Finally, we stack the data cubes of all segments in order to achieve a high signal-to-noise ratio. Since the beam size of a telescope varies with frequency, the spatial resolutions are different over those twenty RRLs. Before stacking, the cubes of different lines are convolved to an uniform beam size of $3\overset{'}{.}4$, which is

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

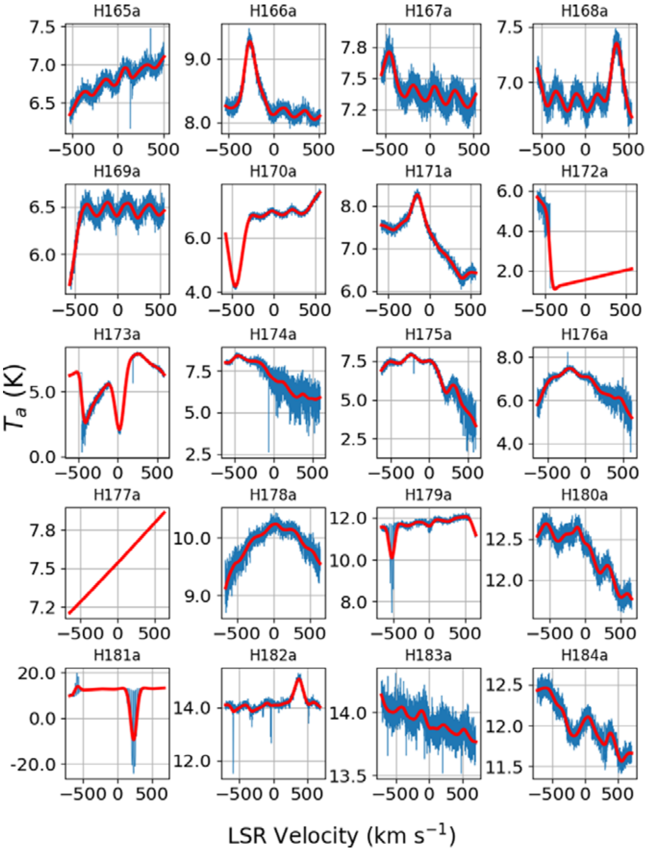

Figure 1. The averaged spectra of RRL segments over a 60 s OTF scan. The blue lines are the spectra and the red lines are the results of the asymmetric least squares smoothing (AsLS). In the pipeline, baseline removal was applied to the raw spectrum with 1 s dumping time. We show the averaged spectra only for the purpose of illustration since the baseline features are hard to be seen from the individual spectrum.

the FAST Half Power Beam Width (HPBW) at 1050 MHz (near the rest frequency of H184α).

3. The PLS-based methods for baseline correction

As the first PLS-based baseline fitting method, the AsLS was proposed by Eilers (2003, 2004) and has proved effective. Since then several improved PLS-based algorithms have been developed including airPLS (Zhang et al. 2010), IAsLS (He et al. 2014), arPLS (Baek et al. 2015), and asPLS (Zhang et al. 2020b). arPLS and asPLS were designed to deal with noisy spectrum, they are described and discussed bellow in details along with the AsLS and our modified method rrlPLS. No further analysis is applied to the airPLS and IAsLS methods since our test with these two did not present effective baseline fitting results to spectra with high noise level.

3.1. The AsLS method

To consider a power spectrum with length of m obtained by a radio telescope, its vector model $\mathbf{y} = [y_1, y_2, \cdots, y_i, \cdots, y_m]^{\mathrm{T}}$ is a composition of the profile of spectral line $\mathbf{s} = [s_1, s_2, \cdots, s_i, \cdots, s_m]^{\mathrm{T}}$, a baseline vector $\mathbf{b} = [b_1, b_2, \cdots, b_i, \cdots, b_m]^{\mathrm{T}}$, and random noise $\mathbf{n} = [n_1, n_2, \cdots, n_i, \cdots, n_m]^{\mathrm{T}}$, which gives

$$\mathbf{y} = \mathbf{s} + \mathbf{b} + \mathbf{n}. \tag{1}$$

Based on the Whittaker smoother (Eilers 2003), Eilers (2004) proposed the function to be minimised for a smoothing background,

$$Q = \sum_{i=1}^{m} w_i \left(y_i - b_i\right)^2 + \lambda \sum_{i=1}^{m} \left(\Delta^2 b_i\right)^2. \tag{2}$$

Δ is the first-order difference and Δ^2 stands for the second-order difference, which gives

$$\begin{aligned}\Delta^2 b_i &= \Delta(\Delta b_i) = (b_i - b_{i-1}) - (b_{i-1} - b_{i-2}) \\ &= b_i - 2b_{i-1} + b_{i-2}.\end{aligned} \tag{3}$$

The weight vector $\mathbf{w} = [w_1, w_2, \cdots, w_i, \cdots, w_m]^{\mathrm{T}}$ are chosen asymmetrically according to

$$w_i = \begin{cases} p, & y_i > b_i \\ 1-p, & y_i \le b_i \end{cases}, (0 < p < 1). \tag{4}$$

p and λ are smoothing parameters which should be optimised based on the data properties and preset by the user.

For convenience of implementation in programming and to simplify the equations, we adopt the form of linear algebra. Let $\mathbf{W}$ to be $m \times m$ diagonal matrix with $\mathbf{w}$ on its diagonal

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_m \end{bmatrix}, \tag{5}$$

and $\mathbf{D}$ as the $(m-2) \times m$ matrix such that $\mathbf{Db} = \Delta^2\mathbf{b}$. According to Equation (3),

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}. \tag{6}$$

Thus Equation (2) can be rewritten to

$$Q = (\mathbf{y} - \mathbf{b})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{b}) + \lambda \mathbf{b}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{b}, \tag{7}$$

By finding the vector of partial derivatives and equating it to zero

$$\frac{\partial Q}{\partial \mathbf{b}^{\mathrm{T}}} = -2\mathbf{W}(\mathbf{y} - \mathbf{b}) + 2\lambda \mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{b} = 0, \tag{8}$$

$$(\mathbf{W} + \lambda \mathbf{D}^{\mathrm{T}}\mathbf{D})\mathbf{b} = \mathbf{W}\mathbf{y}. \tag{9}$$

Solving Equation (9), we will obtain the optimal solution of baseline $\mathbf{b}$.

Difficulty lies in choosing values of p and λ objectively for the AsLS method. Experience has shown that this algorithm, using visual inspection to choose the parameters p and λ is effective and fast. For a baseline estimate, p near zero and rather large λ make $\mathbf{b}$ follow the valleys of $\mathbf{y}$, that is, $p = 0.001$ and $\lambda = 10^5$.

To start the calculation, the initial weights have to be assigned. Thus $w_i = 1$ is set to obtain an initial baseline $\mathbf{b_0}$, which is then adopted to derive new weights. Multiple literation are then followed to update the weight vector $\mathbf{w}$ and to estimate better baseline

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

b. The converging solution will be reached quickly and reliably in about 10 iterations.

3.2. *arPLS*

In order to perform baseline correction in noisy environment, Baek et al. (2015) proposed the arPLS algorithm. Given the optimising Equation 9 from AsLS, they assign the weight vector **w** according to the following equation:

$$w_i = \begin{cases} \text{logistic}\left(y_i - b_i, m_{\mathrm{d}^-}, \sigma_{\mathrm{d}^-}\right), & y_i > b_i \\ 1, & y_i \le b_i \end{cases} \tag{10}$$

where $m_{\mathbf{d}^-}$ and $\sigma_{\mathbf{d}^-}$ are the mean and standard deviation of $\mathbf{d}^-$. Defined as $\mathbf{d} = \mathbf{y} - \mathbf{b}$, $\mathbf{d}^-$ is the negative values of $\mathbf{d}$ when $y_i < b_i$. The logistic function is introduced as follows:

$$\text{logistic}(d, m, \sigma) = \frac{1}{1 + e^{k(d-(-m+s\sigma))/\sigma}}, \tag{11}$$

where k and s are asymmetric and shifting coefficients, which can be used to squeeze the transient region and to shift the weight curve along x-axis. The default values given by Baek et al. (2015) is $k = 2$ and $s = 2$.

3.3. *asPLS*

In order to attenuate the baseline boost at line peaks, Zhang et al. (2020b) proposed the asPLS method. With the increase of λ, the smoothed curve in the line peak region is closer to the actual baseline, while in line free regions the curve deviates further from baseline. Their idea is to adopt different smoothing parameter λ for different channels of the spectrum, meaning that to set large λ in line peak regions and small value in line free regions.

To implement the asPLS algorithm, a coefficient vector α is introduced to tune the amplitude of λ. The minimising equation, Equation (2), can then be rewritten as

$$Q = \sum_{i=1}^{m} w_i \left(y_i - b_i\right)^2 + \sum_{i=1}^{m} (\alpha_i \lambda) \left(\Delta^2 b_i\right)^2, \tag{12}$$

where α_i follows

$$\alpha_i = \frac{\text{abs}\left(y_i - b_i\right)}{\max\left(\text{abs}\left(\mathbf{y} - \mathbf{b}\right)\right)}, \tag{13}$$

where abs() is to calculate the absolute value and max() is to find the maximum value. According to Equation (13), a large value of α_i is given in the line peak region where the difference between **y** and **b** is large. And small α_i are introduced in line free regions. Zhang et al. (2020b) introduced the weight function for asPLS following

$$w_i = \frac{1}{1 + e^{k\left(d_i - \sigma_{\mathrm{d}^-}\right)/\sigma_{\mathrm{d}^-}}}, \tag{14}$$

where k is asymmetric coefficient with a default value of 2.

3.4. *A modified method: rrlPLS*

The arPLS and asPLS methods both introduced pros and cons for baseline estimations compared to AsLS. In order to optimise the fitting results to the real RRL data observed with FAST, a modified method is introduced by combining the features of arPLS and asPLS, which is named as modified penalised least square for FAST radio recombination lines (rrlPLS).

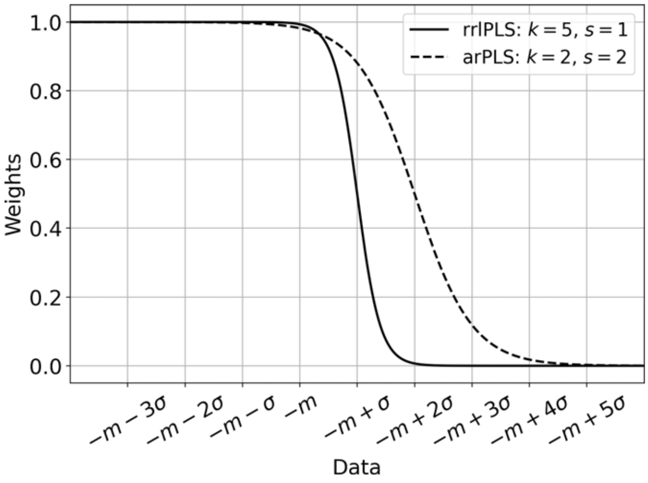

Figure 2. The weight curve for rrlPLS (solid line) with $k = 5$, $s = 1$ and the default weight curve of arPLS (dashed line) with $k = 2$, $s = 2$.

As is described in Section 2, the observation of RRL mapping with FAST uses MBOTF mode. The raw spectra, which are recorded with a changing pointing, have to be processed directly. Averaging is not an option until data cubes are being produced during re-gridding. Thus in the raw data to be processed, RRL signals are commonly weak and accompanied by relatively high noise.

In our modified method, a re-shaped weight function is derived from Equation (11), where the asymmetric coefficient is set to $k = 5$ and shifting coefficient $s = 1$. Comparing to the default values of arPLS, the new curve assigns smaller weights to positive differences and follows a more sharp trend on the negative side (see Figure 2).

Meanwhile, we adopt the idea of setting different λ with the α according to Equation (13). Flatter baseline is obtained to the line peak region with larger λ, whereas smaller λ produces more curvy baseline for the line free regions. Since the weight curve is fixed, it remains only one parameter, λ, to be optimised.

4. Investigations with simulated spectra

The function of PLS-based methods for baseline correction strongly depends on line intensities, noise level, and the amplitude of baseline ripple. Also the position of the line on the baseline ripple (or the 'phase' of the standing wave where the line is located) affects the fitting results significantly. For an unbiased comparison between all the methods, we perform a simulation by varying all the related conditions to obtain a statistically significant conclusion.

4.1. *The simulation configuration*

The spectra for simulation are generated following Equation 1. To match with the RRL spectra given by the FAST pipeline, the local standard of rest (LSR) velocity range is from –400 to +400 km s^{-1} with a resolution of 0.5 km s^{-1}. Accordingly, the length of the spectral vector is 1600. The line profile **s** is modelled with a Gaussian function, whose amplitude is 1 as the relative line peak intensity and FWHM is 20 km s^{-1} for the typical line width of Galactic RRLs. To imitate standing wave ripples in the frequency

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

bandpass of FAST, the baseline vector **b** is modelled by a sinusoidal function with a period of 200 km s^{-1} in velocity.

Considering that the RRL intensities vary from sources, different baseline and noise conditions are configured. We define the signal-to-baseline ratio R_b as the ratio of the line peak intensity (always equals to 1) to the amplitude of sine wave of baseline. Random noise with Normal distribution is added to the spectral model according to the preset signal-to-noise ratio R_n, which is the ratio of the line peak intensity to the standard deviation of the noise vector **n**. Spectra are then simulated with two different pairs of R_b and R_n for different case studies:

Case A ($R_b = 5, R_n = 5$) This is an ideal case consisting a clear detection of strong line with weak baseline ripples.

Case B ($R_b = 3, R_n = 3$) This is a difficult scenario, in which the line signal is relatively weak due to intensive noise level and strong baseline ripples.

Baseline fittings with the four methods introduced in Section 3 to the simulated spectra are then performed. For each case, multiple spectra are simulated with 200 different line peak velocities from −100 to +100 km s^{-1} and with random noise generated from 50 different seeds for each velocity. Thus, 10000 tests are conducted for each pair of parameters for each method.

Two factors are introduced to examine the fitting results. One is the relative loss of the line peak intensity, which is defined as

$$loss = \frac{F_{\rm fit} - F_{\rm sim}}{F_{\rm sim}} \times 100\%, \qquad (15)$$

where $F_{\rm fit}$ is the fitted line peak intensity of the corrected spectrum and $F_{\rm sim}$ is the original line peak intensity for simulation. The astronomical spectra are always noisy and the spectral line intensities are normally weak. The fitted baseline is usually overestimated in line peak regions when the noise level is high, thus the flux loss of line peak intensity is introduced. Similarly we also examine the relative deterioration of the spectral rms noise, which is defined as

$$deterioration = \frac{\sigma_{\rm res} - \sigma_{\rm noi}}{\sigma_{\rm noi}} \times 100\%, \qquad (16)$$

where $\sigma_{\rm res}$ is rms of residual of the corrected spectrum after removing the fitted line profile, and $\sigma_{\rm noi}$ is rms of the simulated noise. Better baseline removal causes smaller rms deterioration.

For stable and reliable baseline fitting, the standard deviation of the distribution of the two factors should be small and the mean should be close to zero. The distribution of those two factors are evaluators in the procedures of parameter optimisation. In order to obtain the optimised values of each method, we manually iterate over the parameter space with small steps to approach the values that yield the best results.

4.2. Results

Table 3 summarises the fitting results with optimised values of parameters for conditions of both Case A and B. The details of results of the four PLS-based methods are discussed as follows.

4.2.1. AsLS

The smoothing parameter λ and weighting parameter p of AsLS are configured within $10^2 < \lambda < 10^9$ and $0.001 < p < 0.5$ to suit for different conditions as suggested. When $p = 0.5$, the algorithm is actually the Hodrick-Prescott filtering algorithm (Hodrick & Prescott 1997) that is widely used for macroeconomic time series.

Table 3. The summary table of optimised simulation results for AsLS, arPLS, asPLS, and rrlPLS methods.

Simulation conditions			Methods	Optimised parameters				Fitting results		
Case	R_b	R_n		λ	p	k	s	Factor	μ(%)	σ(%)
A	5	5	AsLS	5×10^5	0.03	–	–	Loss	−9.6	3.3
								Deter.	1.5	0.6
			arPLS	1×10^6	–	–	–	Loss	−8.0	2.7
								Deter.	−0.3	0.2
			asPLS	5×10^5	–	–	–	Loss	−2.2	6.8
								Deter.	7.8	5.2
			rrlPLS	1×10^7	–	5	1	Loss	−3.3	3.8
								Deter.	−0.2	0.3
B	3	3	AsLS	5×10^5	0.03	–	–	Loss	−11.5	5.6
								Deter.	1.3	0.5
			arPLS	1×10^6	–	–	–	Loss	−16.8	4.4
								Deter.	−0.2	0.2
			asPLS	5×10^5	–	–	–	Loss	−4.7	10.6
								Deter.	7.5	4.9
			rrlPLS	1×10^7	–	5	1	Loss	−6.6	6.3
								Deter.	−0.1	0.3

Col. 1-3 show the conditions of simulated spectra following the description given in Section 4. Col. 4 lists the names of PLS-based methods. Col. 5-8 are the optimised values of parameters for each method. '-' is marked if not applied. Col. 9-11 give the simulation results for the two factors defined by Equations (15) and (16). μ and σ are the mean and standard deviation of the results of 10000 tests for one method under each condition.

In order to demonstrate the utility of AsLS, we present two fitting examples with two sets of parameters for both Case A and B. The two pairs of λ and p are $\lambda = 5 \times 10^4$, $p = 0.001$ and $\lambda = 5 \times 10^6$, $p = 0.45$. The fitting results are presented in Figures 3 and 4, both of which contain three panels. The top panel shows the original spectrum (grey) with the simulated line profile (red) and baseline (blue) overlaid. For comparison, the two fitted baselines are also plotted in this panel. The middle and bottom plots are the corrected spectra of the two fittings with the simulated and fitted line profiles. Affected by the noise level, the fitted baseline is likely to be apart from the 'real' baseline with a negative offset, especially when $p << 0.5$. So one more step to correct the spectrum, after removing the fitted baseline, is to further remove the median value of the subtraction.

The optimised parameters of AsLS is found to be $p = 0.03$ and $\lambda = 5 \times 10^5$ for both cases. We plot histograms of the loss and deterioration factors with the optimised parameters (see Figure 5 for Case A and Figure 6 for Case B). The mean flux loss is −9.6% with a standard deviation of 3.8% for Case A and −11.5% with $\sigma = 5.6\%$ for Case B. The rms deterioration distribution has the mean of 1.5% with $\sigma = 0.6$ for Case A and 1.3% with $\sigma = 0.5\%$ for Case B. This experiment suggests that AsLS can fit the FAST baseline ripples effectively. However, due to the high noise feature of our data, this method may cause an average line peak intensity loss of ∼10%.

4.2.2. arPLS

Adopting the default weight function given by Equation (10), we obtained the optimised $\lambda = 5 \times 10^6$ for both cases (see Table 3). To

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

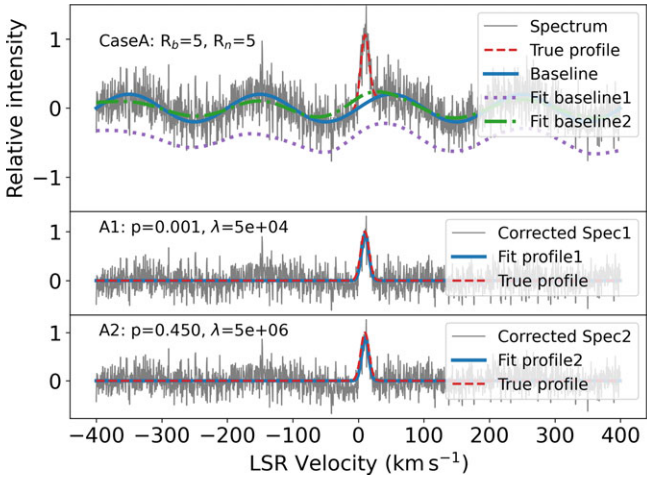

Figure 3. The simulated spectrum and AsLS fitting results under Case A condition. The top panel shows the simulated spectrum (solid grey), which is the combination of a Gaussian peak (dashed red) as the line profile, a sine wave (solid blue) as the baseline ripple, and white noise. AsLS baseline fitting results from two different parameter configurations are also plotted (dotted blue and dash-dotted green). The middle and bottom panel give the baseline corrected spectra (solid grey) from two different parameter configurations which are overlaid by their fitted Gaussian line profiles (solid blue). The simulated Gaussian peaks (dashed red) are also shown for comparison.

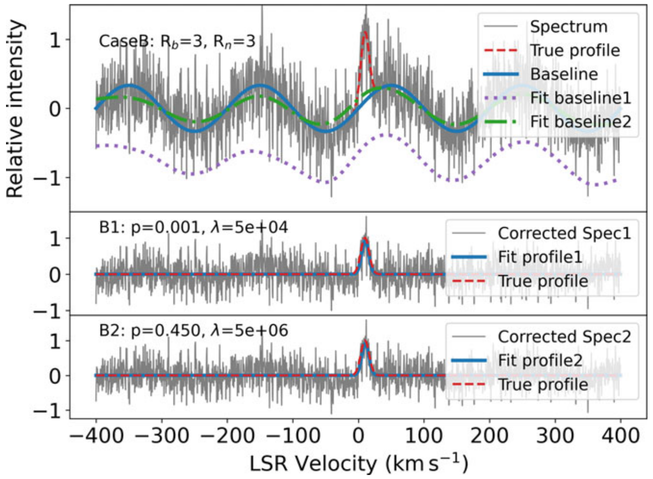

Figure 4. The simulated spectrum and AsLS fitting results under Case B condition. The plots of the three panels are following the same instruction given in Figure 3.

examine the distribution of the results, we also plot the histograms in Figure 7 for Case A and Figure 8 for Case B.

The distribution of flux loss has a mean of -8.0% with σ of 2.7% for Case A and -16.8% with 4.4% for Case B. The mean of rms deterioration is -0.3% with $\sigma = 0.2\%$ for Case A and -0.2% with $\sigma = 0.2\%$ for Case B. In comparison with the AsLS results, the smaller value of standard deviation of the flux loss distribution implies that the arPLS method is more stable than AsLS for different conditions. Although it works better to strong signals as in Case A, it causes more flux loss on average than AsLS for weak signals in Case B. The negative amplitude of noise deterioration means that the baseline is slightly overfitted.

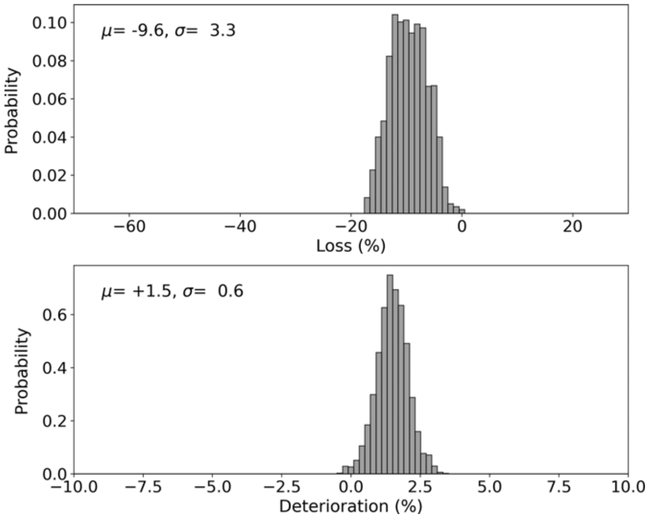

Figure 5. The distribution of simulation results for Case A using AsLS method with optimised parameters. The optimised parameters of AsLS method are $\lambda = 1 \times 10^5$ and $p = 0.03$. The upper panel is histogram of the flux loss and the lower panel shows the histogram of noise deterioration. The μ and σ values labelled in the figures are the means and standard deviations of their distributions.

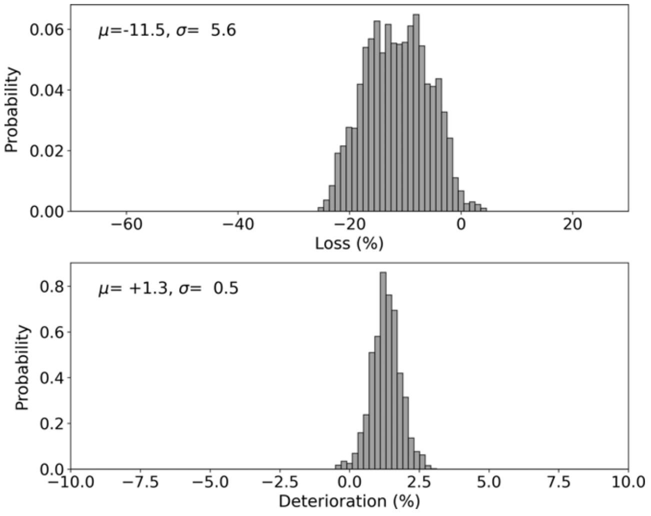

Figure 6. The distribution of simulation results for Case B using AsLS method with the same optimised parameters for Case A ($\lambda = 1 \times 10^5, p = 0.03$). The figure instruction follows that is given in Figure 5.

4.2.3. asPLS

The smoothing parameter of asPLS is optimised to be $\lambda = 5 \times 10^8$ for both Case A and B (see Table 3 for details). The resulted distributions are also plotted in Figure 9 for Case A and Figure 10 Case B.

The mean of the flux loss distribution is -2.2% with σ of 6.8% for Case A and -4.7% with 10.6% for Case B. The mean of rms deterioration is -3.3% with $\sigma = 3.8\%$ for Case A and 7.5% with $\sigma = 4.9\%$ for Case B. Comparing with AsLS and arPLS, the flux loss distributes closer to zero although its standard deviation becomes larger. It seems that the asPLS method could improve the line intensity attenuation problem as expected, but its baseline fitting results may not be very stable for different situations.

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

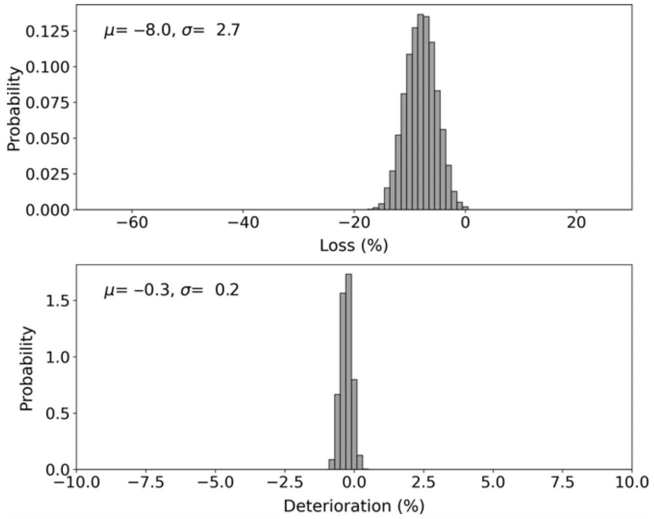

Figure 7. The distribution of simulation results for Case A using arPLS method with optimised parameter. The optimised value of parameter λ is 1×10^6. The figure instruction follows that is given in Figure 5.

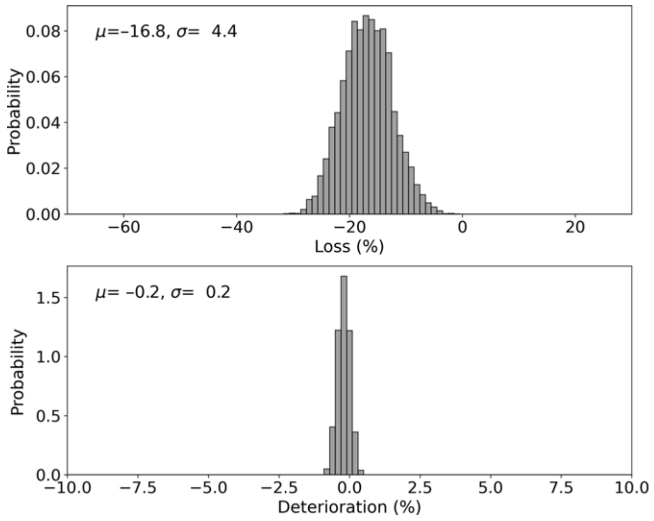

Figure 8. The distribution of simulation results for Case B using arPLS method with the same optimised parameter for Case A ($\lambda = 1 \times 10^6$). The figure instruction follows that is given in Figure 5.

Moreover, the spectral rms deteriorates significantly, thus asPLS is not an ideal method for our RRL data reduction.

4.2.4. rrlPLS

Similar as other methods, simulations with rrlPLS are conducted. When $\lambda = 1 \times 10^7$, we obtain the best baseline fitting results. Figures 11 and 12 present the histograms of the results for Case A and B.

The mean of the flux loss distribution is −3.3% with σ of 3.8% for Case A and −6.6% with 6.3% for Case B. The mean of rms deterioration is −0.2% with $\sigma = 0.3\%$ for Case A and −0.1% with $\sigma = 0.3\%$ for Case B. Comparing to the other three methods, the ~5% flux loss introduced with nearly ~0% noise deteriorations make rrlPLS the most promising baseline correction method to our RRL spectra.

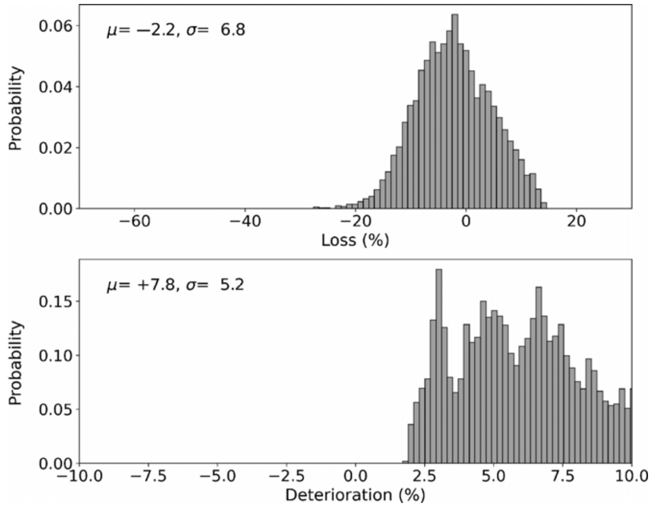

Figure 9. The distribution of simulation results for Case A using asPLS method with optimised parameters. The optimised value of parameter λ is 5×10^5. The figure instruction follows that is given in Figure 5.

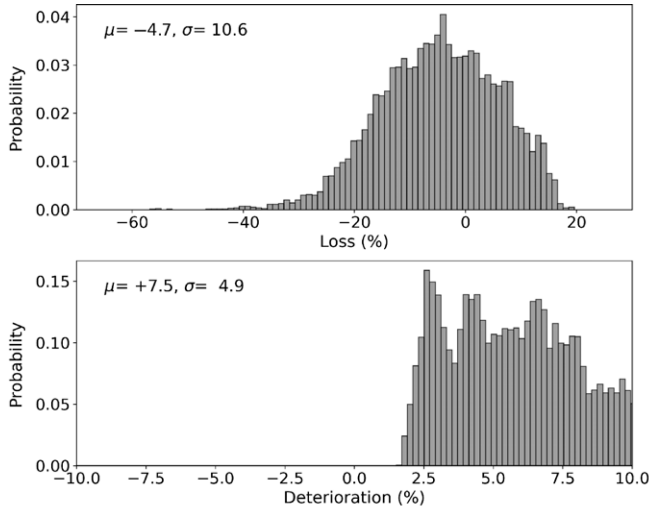

Figure 10. The distribution of simulation results for Case B using asPLS method with the same optimised parameters for Case A ($\lambda = 5 \times 10^5$). The figure instruction follows that is given in Figure 5.

5. Apply to real RRL data

The four methods discussed in Section 3 and 4 are employed to fit the baselines of observed raw spectra. We extract a spectral segment of H169α from the full bandpass of calibrated data. Then the spectra of the RRL segment are baseline removed using one of these methods. Finally the processed spectra are re-grided into data cubes.

For each method, the optimised parameters listed in Table 3 are firstly adopted. Considering the difference between the simulated and observed data sets, we further tuned the parameters by small steps. No clear improvements have been seen except for the rrlPLS, in which $\lambda = 2 \times 10^8$ is configured instead of 1×10^7.

To compare the FAST RRL results with previous studies, we provide the RRL 0th moment map (Figure 12) given by the Survey of Ionized Gas in the Galaxy Made with Arecibo (SIGGMA, Liu et al. 2013). The sensitivity of stacked RRLs from SIGGMA

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

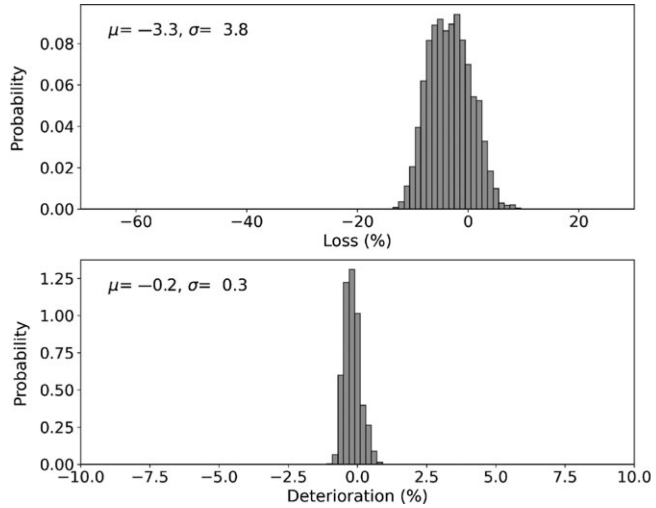

Figure 11. The distribution of simulation results for Case A using rrlPLS method with optimised parameters. The optimised values of parameters are $\lambda = 1 \times 10^7$, $k = 5$, and $s = 1$. The figure instruction follows that is given in Figure 5.

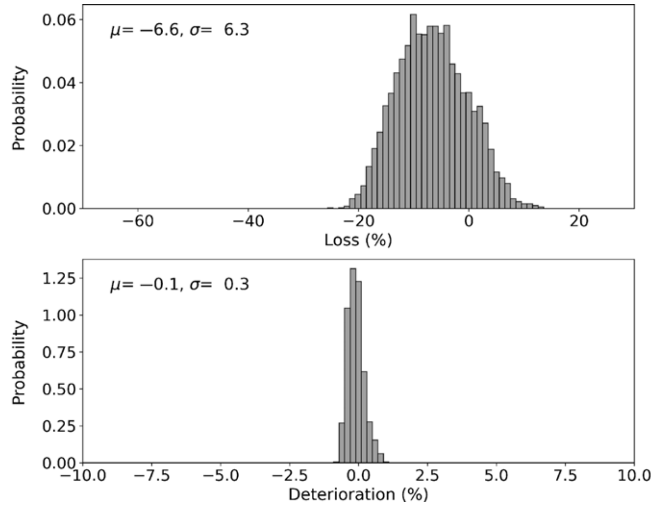

Figure 12. The distribution of simulation results for Case B using rrlPLS method with the same optimised parameters for Case A ($\lambda = 1 \times 10^7$, $k = 5$, and $s = 1$). The figure instruction follows that is given in Figure 5.

is remarkable, however, there may be some unreliable spatial-extended features in the map due to its survey strategy and data quality (Liu et al. 2019). Therefore, we also present the 1.4 GHz continuum map (Figure 13b) from the VLA Galactic Plane Survey (VGPS, Stil et al. 2006). For a better comparison, the VGPS map was convolved to the FAST HPBW and re-projected to the grid of FAST RRL maps. The data processing results using AsLS, arPLS, asPLS, and rrlPLS methods are presented in Figures 14, 15, 16, and 17. In each figure, the left panel presents the 0th moment map integrated over the velocity range between 20 and 100 km s^{-1} from the cube. The right panel gives two spectra at the locations marked as (A) and (B) in the moment map. (A) is a known bright H II region, where intensive RRL emission exists. (B) is a relatively 'empty' spot in the field, where no strong RRL signals are expected. The spectra of (A) and (B) are corresponding to the Case A and B in the simulation.

The AsLS method was first adopted for our project. In Figure 14, one can see smoothed gas structures and clean spectral baselines. However, it introduces notable flux loss to strong emitting sources. The arPLS method gives the best baseline fitting but also causes the most severe flux losses. Most of the RRL emissions are eliminated as demonstrated in Figure 15. The asPLS method was designed to retain line signals from noisy spectra. Although the corresponding flux loss is small, the resulting baseline quality is the worst comparing to the other three methods (Figure 16). Furthermore, because of the bad baseline, the line peak intensities of weak sources are inaccurate.

Finally, the rrlPLS method presents the most promising results in Figure 17. Comparing the spectrum of (A) in Figure 17 with that in Figure 16, the line peak intensities are identical. The flux loss introduced by rrlPLS is as small as asPLS. Whereas the baseline of strong continuum source given by rrlPLS is not as good as that from AsLS or arPLS, and those of weak emissions are similar. Furthermore, the 0th moment map resulted from rrlPLS reveals the most intensive gas structures than that from other methods, implying that rrlPLS with $\lambda = 2 \times 10^8$ produces the best RRL data sets.

The results of RRL maps processed by the PLS methods agree with the simulation results in Section 4. Therefore, discussions on emission line searching, gas kinematics from LSR velocities, and gas morphology from relative line peak intensities are reliable. One should still be cautious about the uncertainties for further astrophysical analysis using the line profiles. Before line profile fitting and further calculations for individual strong continuum sources, a high order polynomial baseline removal is suggested by masking the velocity range of the detected RRL.

6. Evaluation with fake source injection

Finally, it is necessary to deploy a quantitative analysis on the line distortions through the whole process using rrlPLS. Since the true intensities of observed RRLs is unknown, simulated Gaussian profiles are added to the raw spectra centring at LSR velocity of −300 km s^{-1} so that fake signals will not overlap with real RRLs. To imitate a point source, all spectra were injected with weights according to FAST beam pattern and the angular distances from positions where spectra were obtained to the fake source location. To cover different baseline scenarios, three sources, located at strong, medium, and weak continuum background, were simulated and added to raw data sets. Data cube was then produced using our pipeline, from which the spectra of the fake sources were extracted and fitted.

Figure 18 shows the 0th moment map and the spectra of the three fake sources, labelled as (f1), (f2), and (f3). (f1) is located at a known bright H II region (strong continuum background), (f2) is within a extended gas structure (medium continuum background), and (f3) is at a weak emission spot (weak continuum background). The simulated (solid blue) and fitted (dashed red) line profiles are overlaid. For relatively weak (f2) and (f3), the simulated and processed line intensities are identical. But for the stronger (f1), the processed profile is notably weaker than the simulated.

To compare with the traditional method, baseline fitting with a 3rd order polynomial (Poly-3), with a velocity mask from −320 to −280 km s^{-1}, were also carried out besides rrlPLS. Table 4 presents the comparison of simulated and processed line profiles using both methods. Although they follows same trends, the results of rrlPLS

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

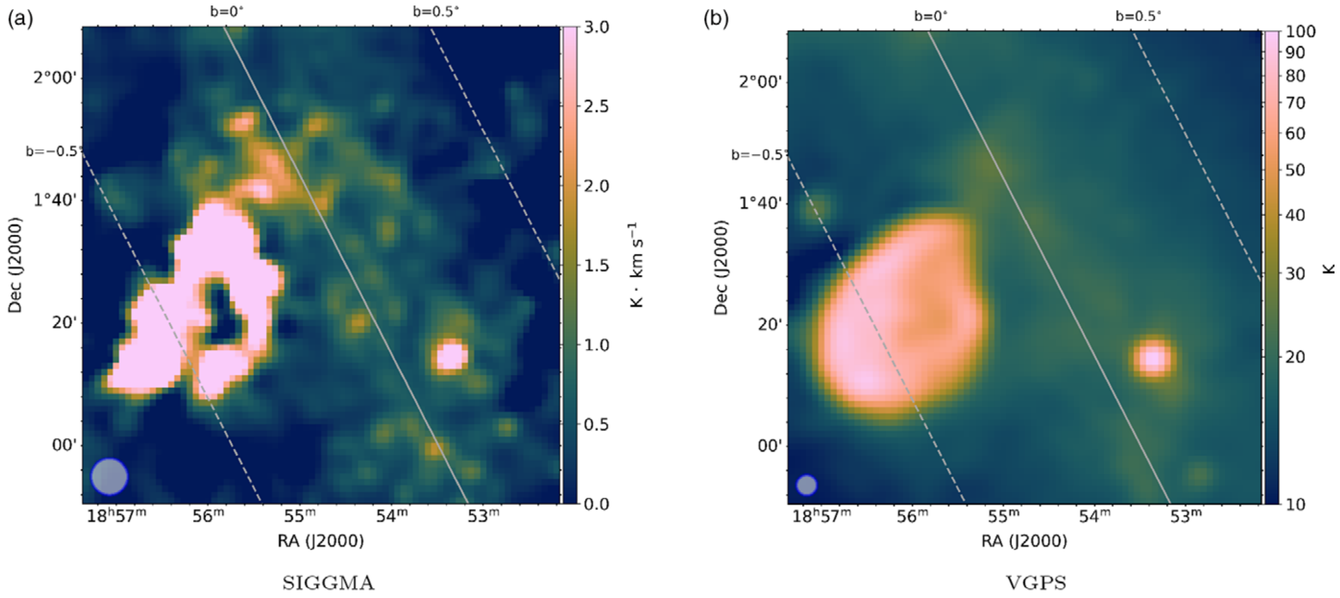

Figure 13. (a) The SIGGMA RRL 0th moment map integrated over the velocity range from 20 to 100 km s^{-1} (Liu et al. 2019). The blue circle at the bottom left corner shows the SIGGMA resolution of 6′. (b) The VGPS continuum map at 1.4 GHz (Stil et al. 2006). The VGPS data is convolved to FAST HPBW of 3′ at 1350 MHz (blue circle at the lower left corner). Both images are reprojected to match with the FAST image grid. The bright extended source located at the middle east in the field is the supernova remnant W44, who shows strong non-thermal continuum emission.

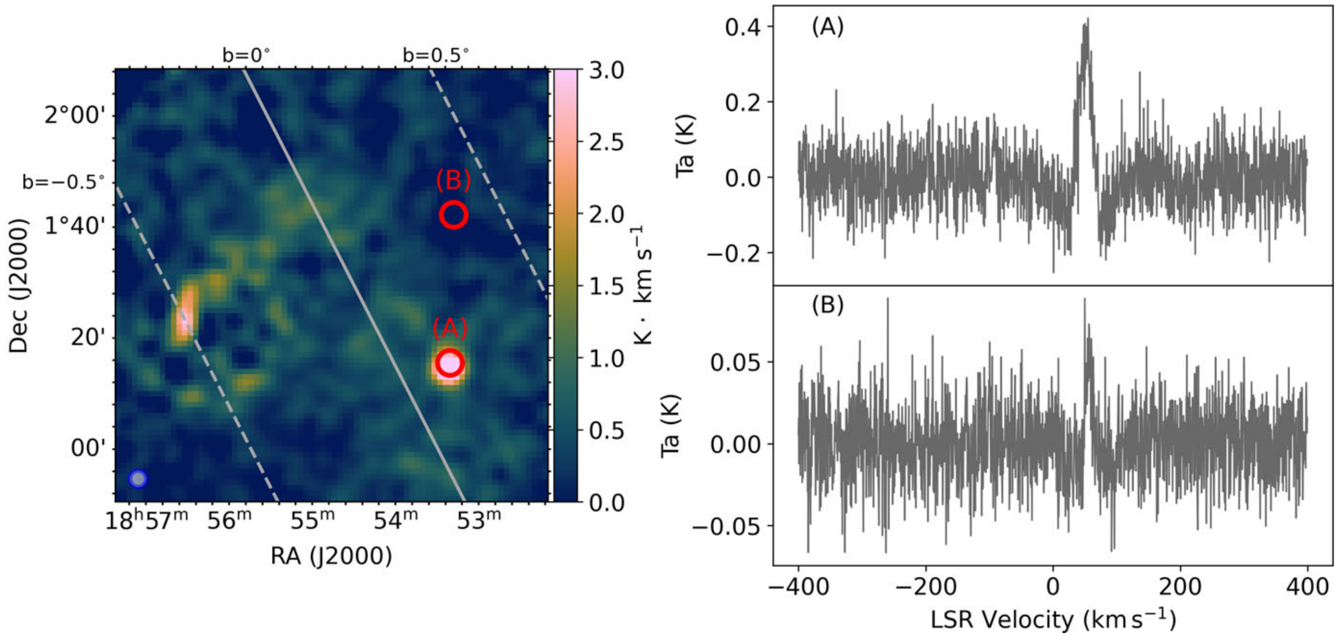

Figure 14. The results of H169α processed using AsLS method. The image on the left is the 0th moment map integrated over the velocity range from 20 to 100 km s^{-1} from the cube. The red circles marked as (A) and (B) in the map are locations with strong and weak RRLs. The blue circle at the bottom left corner shows the FAST beam size of 3′. The right panel plot two spectra at the locations marked as (A) and (B) in the left-hand moment map. (A) is apart from strong continuum source, where RRL signal is weak. (B) is a known bright H II region, who shows intensive RRL emission. The spectra of (A) and (B) are corresponding to the Case A and B in the simulation.

are more consistent than that of Poly-3. The latter does not reduce the standing wave ripples in the spectra, thus will not generate reliable emission structures in the map. The flux losses of (f2) and (f3) of rrlPLS are possibly introduced by the rms noise of spectra. After a 5th order polynomial baseline fitting to the (f1) spectrum from rrlPLS, the fitted line intensity is well recovered.

To summarise, line profiles with low signal-to-noise ratio were less affected by rrlPLS, whose uncertainty were mainly caused by the rms noise. Although the majority of the detected RRLs are weak, to which the rrlPLS method is acceptable, still the strong line peaks may affect the baseline fitting towards a few intensive positions in the field. Therefore, after the identification of strong RRLs,

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

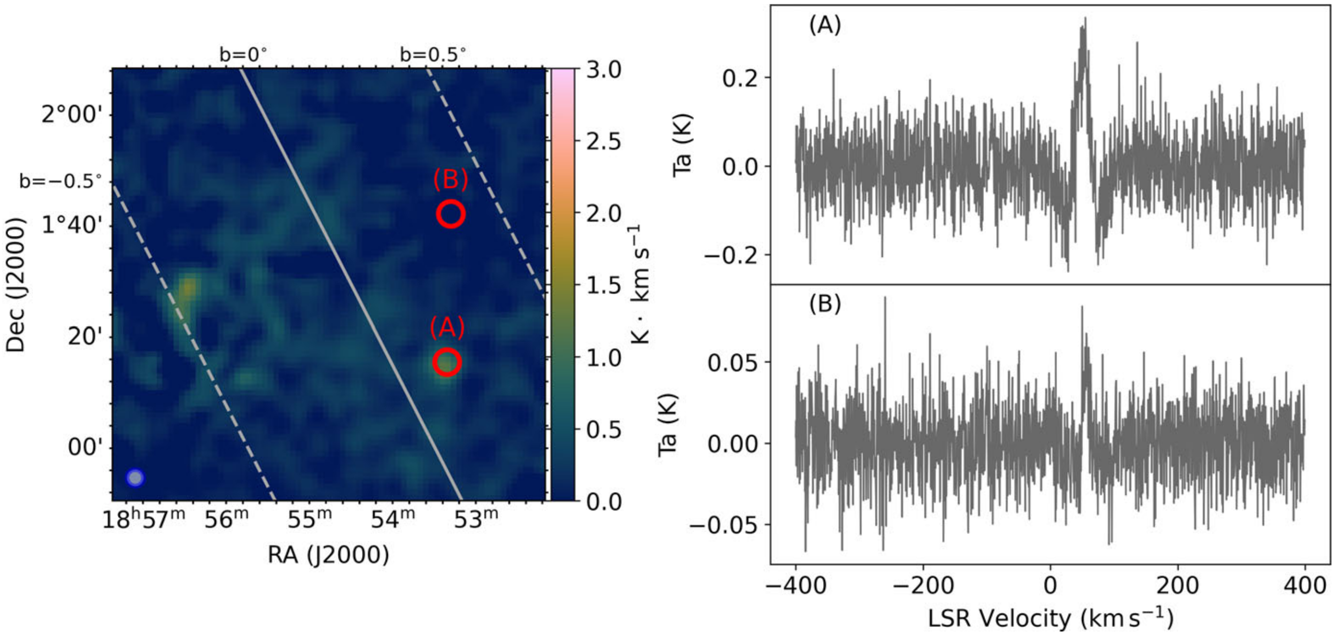

Figure 15. The results of H169α processed using arPLS method. The figure instruction follows that is given in Figure 14.

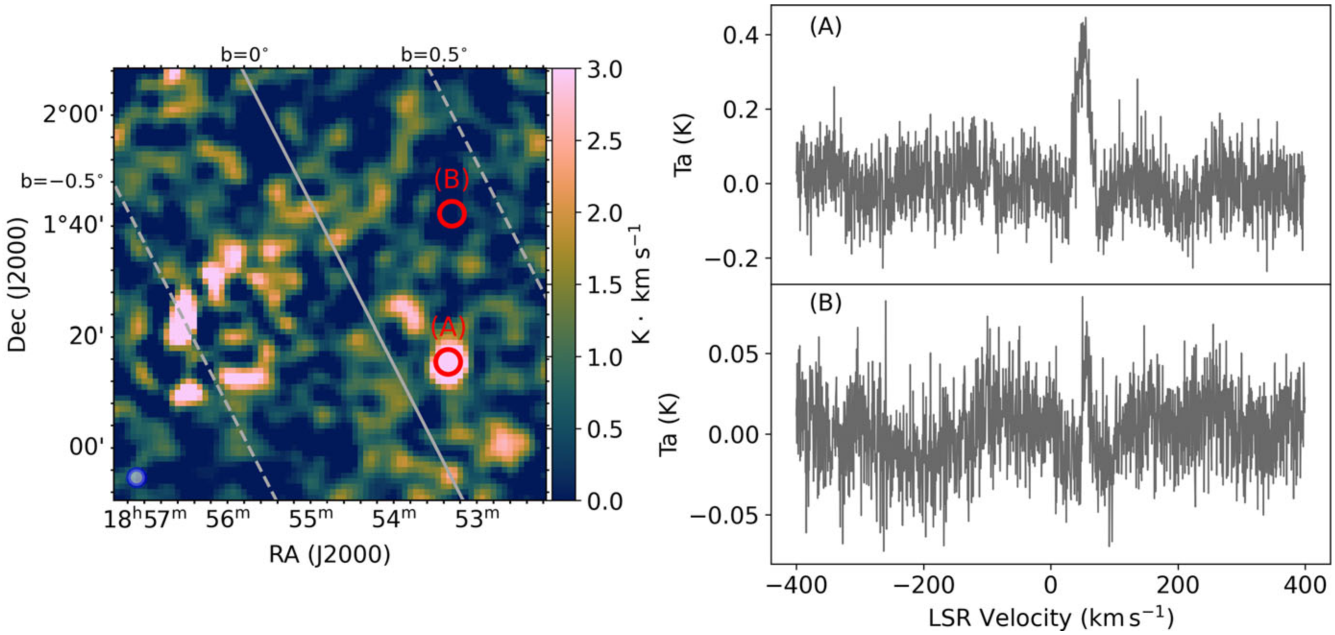

Figure 16. The results of H169α processed using asPLS method. The figure instruction follows that is given in Figure 14.

high order polynomial baseline removals are suggested for accurate line profile fittings. In addition, we note that the line widths are reduced after baseline processing from both Poly-3 and rrlPLS. Since the line narrowing is not distinct from the methods applied, it may be caused by the remained baseline ripples. For scientific discussion with the current data sets, one need to be careful about calculations using line widths, which may be under estimated.

7. Conclusions

To investigate the ionised environment in the Galaxy using FAST, RRL map of 1 deg^2 on the Galactic plane has been processed, which serves as a pilot study for a further large-scale Galactic plane RRL survey with FAST. In this paper, We introduced the observing details, survey configurations, and data processing pipeline

Table 4. The comparison of rrlPLS fitting results with simulated spectra injected into real RRL data.

Source	True profile			Method	Fitted profile			Flux loss	rms Noise
	Peak	V_{LSR}	FWHM		Peak	V_{LSR}	FWHM		
	(K)	$km\,s^{-1}$	$km\,s^{-1}$		(K)	$km\,s^{-1}$	$km\,s^{-1}$	(%)	(K)
f1	0.233	–300	25	rrlPLS	0.152 ± 0.019	-300.8 ± 1.0	15.6 ± 1.0	–34.8	0.050
				Poly-3	0.139 ± 0.032	-301.9 ± 1.6	14.5 ± 1.6	–40.3	0.061
f2	0.044	–300	20	rrlPLS	0.039 ± 0.007	-301.6 ± 1.3	15.9 ± 1.3	–11.4	0.019
				Poly-3	0.040 ± 0.008	-302.9 ± 1.7	16.9 ± 1.7	-9.1	0.021
f3	0.026	–300	15	rrlPLS	0.031 ± 0.006	-300.1 ± 1.0	10.5 ± 1.0	$+19.2$	0.018
				Poly-3	0.033 ± 0.006	-299.9 ± 1.1	10.8 ± 1.1	$+26.9$	0.018
f1[a]	0.233	–300	25	rrlPLS	0.216 ± 0.010	-300.4 ± 0.5	21.6 ± 0.5	-7.1	0.050

Col.1 are the name of injected fake sources. Cols.2-4 list the true profile parameters of the simulated spectra. Col.5 gives the baseline removal methods applied in the pipeline. 'Poly-3' stands for the 3rd order polynominal fitting. Col.6-8 list the fitting parameters of the simulated spectra after the baseline correction processed. Col.9 and 10 are the relative flux loss and the rms values of spectra, which are calculated from velocity range between $+300$ and $+400\ km\,s^{-1}$.

The line profile is fitted after an extra 5th order polynomial with velocity mask covering –320 to $-280\ km\,s^{-1}$.

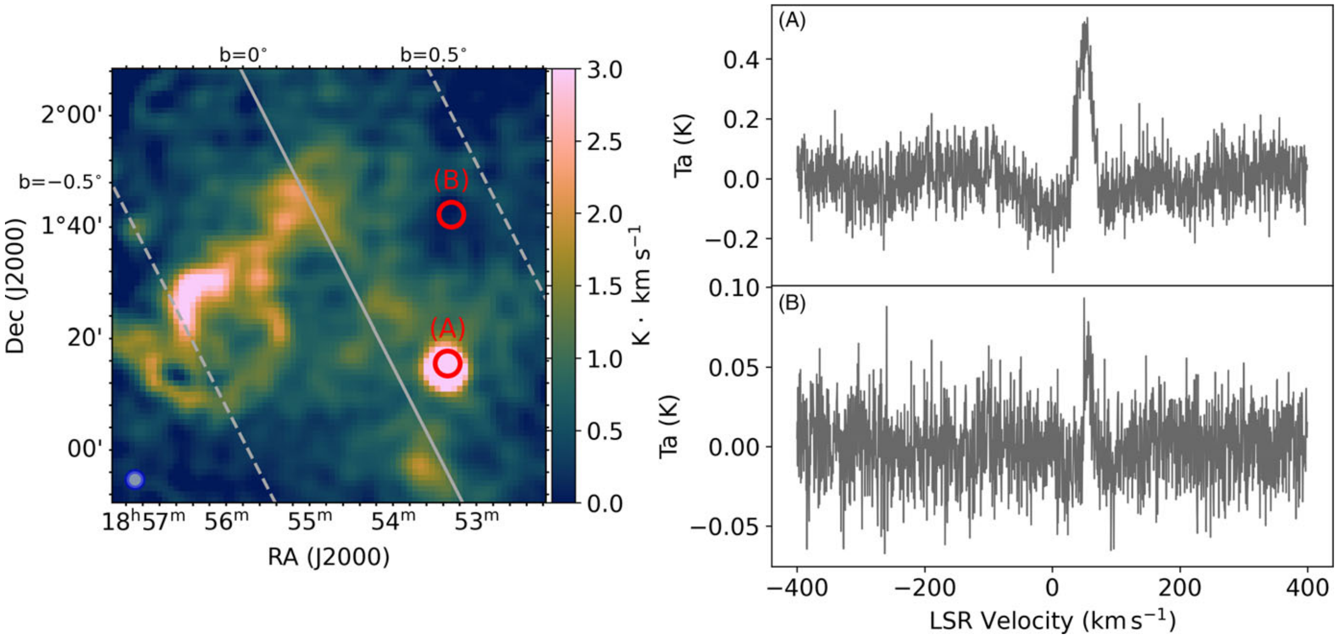

Figure 17. The results of H169α processed using rrlPLS method. The figure instruction follows that is given in Figure 14.

developments. The data shows that the frequency bandpass given by the FAST 19 beam L-band receiver is severely affected by RFIs and standing wave ripples, which brings a major challenge lying in the baseline fitting step. Low order polynomial baseline removal method, which is widely used for spectroscopy studies in radio astronomy, is not suitable for this complex scenarios.

To solve the baseline problem, we investigate a series of PLS-based baseline correction methods in this paper. The AsLS, arPLS, and asPLS methods were evaluated using simulated spectra according to the actual features of FAST bandpass. To further improve the results of baseline correction, we developed a modified method, rrlPLS, by adopting the advantages of arPLS and asPLS. Optimised parameters were obtained from our simulations. The four methods were then applied and compared by reducing the real FAST data. The rrlPLS with the optimised parameter $\lambda = 2 \times 10^8$ revealed the most sensitive and reliable RRL emission features in the 0th moment map, and thus was well-suited for our project.

We finally verified the line distortion, which the rrlPLS method may cause, using fake RRL sources injected into the raw data sets. Small distortions were identified by comparing the processed profiled to the simulated ones. It is concluded that the weaker the line intensity the less it may be affected by this baseline removal method. While for strong emission regions, an extra high order polynomial is suggested before fitting the line profile accurately, so that the uncertainty is mainly introduced by the rms noise. Note that the fitted line width may be smaller than the actual signal due to the baseline ripples and the current limitation of the method. Cautious considerations should be made for line width related science cases.

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press

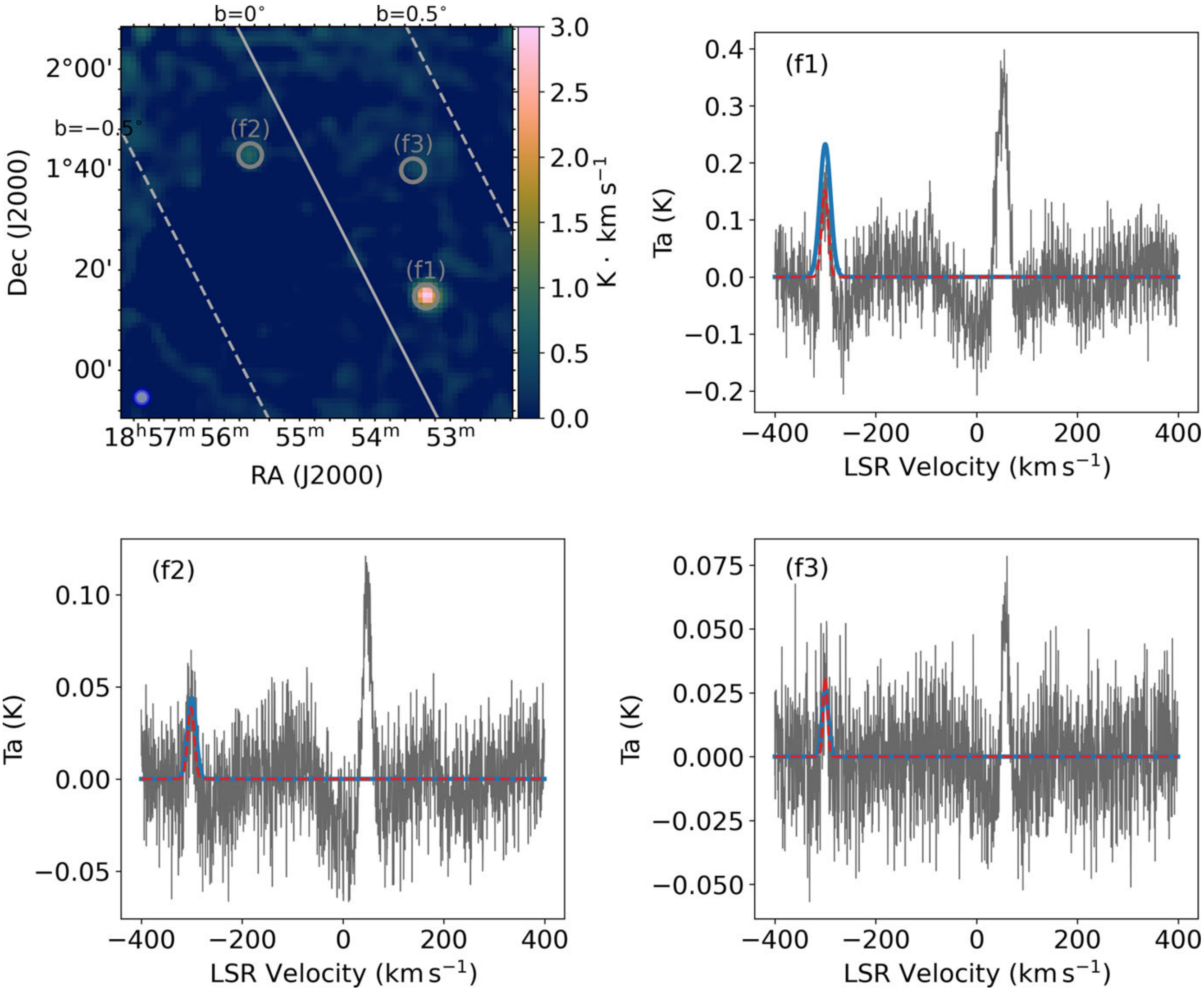

Figure 18. The comparison of rrlPLS fitting results with simulated spectra injected into H169α data. The top left is the 0th moment map integrated over the velocity range from −320 to −280 km s^{-1}, within which the fake line profiles are injected. The blue circle at the bottom left corner of the map shows the FAST beam size of 3′. The top right and bottom plots are the spectra extracted from data cube towards the fake sources. The solid grey lines are the processed spectra, solid blue lines are the injected Gaussian profiles, and dashed red lines are the fitted line profiles to the spectra.

A Galactic plane RRL mapping project with FAST is now in progress. The baseline correction technique verified in this paper will be applied in our future data processing pipeline.

Acknowledgements. This work is supported by the National Key Basic Research and Development Program of China(grant No. 2017YFA0402604 and 2018YFA0404703), and the National Natural Science Foundation of China (No.11703048) and the Open Project Program of the Key Laboratory of FAST, NAOC, Chinese Academy of Sciences. This work made use of the data from FAST (Five-hundred-metre Aperture Spherical radio Telescope). FAST is a Chinese national mega-science facility, operated by the National Astronomical Observatories, Chinese Academy of Sciences.

References

Alves, M. I. R., Calabretta, M., Davies, R. D., Dickinson, C., Staveley-Smith, L., Davis, R. J., Chen, T., & Barr, A. 2015, MNRAS, 450, 2025
Anderson, L. D., et al. 2021, ApJS, 254, 28
Baek, S.-J., Park, A., Ahn, Y.-J., & Choo, J. 2015, Analyst, 140, 250
Eilers, P. H. 2003, AC, 75, 3631
Eilers, P. H. 2004, AC, 76, 404
He, S., Zhang, W., Liu, L., Huang, Y., He, J., Xie, W., Wu, P., & Du, C. 2014, AM, 6, 4402
Hodrick, R., & Prescott, E. 1997, JMCB, 29, 1
Jiang, P., et al. 2020, RAA, 20, 064
Liu, B., Anderson, L. D., McIntyre, T., Anish Roshi, D., Churchwell, E., Minchin, R., & Terzian, Y. 2019, ApJSS, 240, 14
Liu, B., McIntyre, T., Terzian, Y., Minchin, R., Anderson, L., Churchwell, E., Lebron, M., & Roshi, D. A. 2013, AJ, 146, 80
Mangum, J. G., Emerson, D. T., & Greisen, E. W. 2007, A&A, 474, 679
Nan, R., et al. 2011, IJMPhD, 20, 989
Peng, J., Peng, S., Jiang, A., Wei, J., Li, C., & Tan, J. 2010, ACA, 683, 63
Qian, L., Yao, R., Sun, J., Xu, J., Pan, Z., & Jiang, P. 2020, Innovation(China), 1
Qiu, Y. H. 1998, MNRAS, 301, 827
Stil, J. M., et al. 2006, AJ, 132, 1158
Zhang, F., Tang, X., Tong, A., Wang, B., & Wang, J. 2020a, Sensors, 20, 2015
Zhang, F., Tang, X., Tong, A., Wang, B., Wang, J., Lv, Y., Tang, C., & Wang, J. 2020b, SL, 53, 222
Zhang, Z.-M., Chen, S., & Liang, Y.-Z. 2010, The Analyst, 135, 1138

https://doi.org/10.1017/pasa.2022.47 Published online by Cambridge University Press