



# LINESTACKER: a spectral line stacking tool for interferometric data

Jean-Baptiste Jolly<sup>1</sup>,<sup>\*</sup> Kirsten K. Knudsen<sup>1</sup> and Flora Stanley<sup>1</sup>

*Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, SE-439 92 Onsala, Sweden*

Accepted 2020 September 11. Received 2020 August 23; in original form 2019 October 26

## ABSTRACT

LINESTACKER is a new open access and open source tool for stacking of spectral lines in interferometric data. LINESTACKER is an ensemble of CASA tasks, and can stack both 3D cubes or already extracted spectra. The algorithm is tested on increasingly complex simulated data sets, mimicking Atacama Large Millimeter/submillimeter Array, and Karl G. Jansky Very Large Array observations of [C II] and CO(3–2) emission lines, from  $z \sim 7$  and  $z \sim 4$  galaxies, respectively. We find that the algorithm is very robust, successfully retrieving the input parameters of the stacked lines in all cases with an accuracy  $\gtrsim 90$  per cent. However, we distinguish some specific situations showcasing the intrinsic limitations of the method. Mainly that high uncertainties on the redshifts ( $\Delta z > 0.01$ ) can lead to poor signal-to-noise ratio improvement, due to lines being stacked on shifted central frequencies. Additionally, we give an extensive description of the embedded statistical tools included in LINESTACKER: mainly bootstrapping, rebinning, and subsampling. Velocity rebinning is applied on the data before stacking and proves necessary when studying line profiles, in order to avoid artificial spectral features in the stack. Subsampling is useful to sort the stacked sources, allowing to find a subsample maximizing the searched parameters, while bootstrapping allows to detect inhomogeneities in the stacked sample. LINESTACKER is a useful tool for extracting the most from spectral observations of various types.

**Key words:** methods: data analysis – techniques: interferometric – galaxies: high-redshift – galaxies: statistics – radio lines: galaxies – submillimetre: galaxies.

## 1 INTRODUCTION

One of the great challenges in astronomy, and more especially in the field of galaxy evolution, comes from the tendency to look primarily at the brightest sources of a given galaxy population. The further we look the fewer intrinsically faint objects are observable, posing a real difficulty in studying faint, low-mass, galaxy properties, or faint tracer of physical and chemical processes. In order to draw an accurate description of a galaxy population it is necessary to study representative samples, including the faint/undetected sources. One method that can be used for this purpose is stacking.

Stacking was first developed for optical data (Cady & Bates 1980) to determine the average properties of otherwise undetected sources and has, since then, been frequently used for studies at many different wavelength ranges (e.g. Knudsen et al. 2005; Hickox et al. 2007, 2009; Karim et al. 2011; Chen et al. 2013; Lindroos et al. 2015, 2016; Stanley et al. 2017). While radio and mm wavelengths observations are essential to study the gas content of high-redshift galaxies, arcsecond and sub-arcsecond angular resolution can only be obtained when doing interferometric observations. However, interferometry is not a direct imaging technique, but instead samples the Fourier transform of the brightness distribution of the sources being observed. Therefore, the produced images are a model representation of the actual data (see e.g. Thompson, Moran & Swenson 2001). While these models are well understood they often lead to the generation of artefacts, making stacking analysis of interferometric data less straightforward (Lindroos et al. 2015).

Stacking of radio and mm wavelength interferometric data has mainly been done for continuum data (e.g. Karim et al. 2011; Decarli et al. 2014; Ikarashi et al. 2015; Lindroos et al. 2016, 2018; Stanley et al. 2017), it has occasionally been done for spectral lines as well (e.g. Murray et al. 2014; Decarli et al. 2018; Bischetti et al. 2019; Fujimoto et al. 2019; Stanley et al. 2019). However, no general open access tool nor any thorough study of the spectral stacking method has yet been published.

In this paper, we describe the functionalities and performances of a new tool for stacking spectral lines: LINESTACKER. LINESTACKER is an ensemble of CASA tasks and it allows stacking of spectral cubes in the image-plane. Its main contribution is the stacking algorithm, but also includes embedded statistical tools for further analysis of the stacked data and optimization of the stacked signal. We also demonstrate the performances and capabilities of LINESTACKER, by testing it on increasingly complex simulated data sets, that mimic mm- and radio observation of emission lines from high-redshift galaxies. The tests are performed on both high and low signal-to-noise ratio (SNR) cases. The high SNR data sets test the reliability of the algorithm in a near-ideal case, while the low SNR data sets are used to verify the efficiency of noise reduction.

In Stanley et al. (2019), we used LINESTACKER to perform a spectral stacking analysis to search for faint outflow signatures in a sample of  $z \sim 6$  quasars. We used the main algorithm and accompanying tools presented in this paper, on a sample of 26 quasars with detected [C II] emission. Our work demonstrated the utility of LINESTACKER as a spectral stacking tool, when searching for faint emission at high redshift.

In Sections 2 and 3, we give a complete description of LINESTACKER, fully characterizing both the main algorithm and the

\* E-mail: [jean.jolly@chalmers.se](mailto:jean.jolly@chalmers.se)

embedded tools. In Section 4, we describe each simulated data set in detail. In Section 5, we give the results from our stacking analysis on the simulated data sets. We discuss the results of the analysis, and review possible outlooks in Section 6. Finally, Section 7 outlines the conclusions of this study.

## 2 LINESTACKER

LINESTACKER is an assembly of CASA tasks allowing stacking of data cubes, specifically cubes with two spatial dimensions and a frequency/velocity dimension. It has been developed specifically to stack spectral lines, with the capability to take into account varying redshift/central-velocities across the sample. In addition, embedded analysis tools are included within LINESTACKER, for further analysis of the stack results and sample. LINESTACKER is an extension of STACKER (Lindroos et al. 2015), a tool built for stacking continuum interferometric data. While STACKER allowed direct visibility stacking, the visibility stacking extension of LINESTACKER is still under construction.

### 2.1 Main algorithm

When stacking cubes, every source is stacked pixel to pixel, spectral bin to spectral bin. Spatial stacking positions as well as observed central frequency – or rest-frame frequency and redshift – of the sources are needed prior to stacking. The position of the source can typically be obtained through continuum observations, or, more generally, through prior observations at other wavelength. Subimages of  $N \times N$  pixels are stacked, centred on the stacking position. Similarly only a subset of the total number of spectral bins, centred on the estimated central frequency of the line in the observer frame, are stacked (note that all the spectral channels can be stacked if required by the user). A good prior knowledge of the observed central frequency of the line, i.e. of the redshift of the source, is needed in order to stack the spectra reliably and thus maximize the amplitude of the reconstructed line (see Section 5.2.1). Both median based (similar to Pannella et al. 2009) and weighted mean based (similar to Decarli et al. 2014) stacking have been developed. Both methods induce a theoretical rms noise reduction by a factor  $\sim \sqrt{N}$ , where  $N$  is the number of sources stacked (if measurements are independent and outliers are symmetric).<sup>1</sup>

Required user inputs to the algorithm are a list of data cubes, the spatial coordinates of the target sources, and the spectral coordinate of the associated line (i.e. either observed central frequency, or redshift of the source and the rest line frequency or central channel index). The spatial and frequency sizes of the subimages are specified by the user prior to stacking. For each target source, data is extracted from its associated cube, and filled into the associated empty subimage. All subimages are then buffered to facilitate access to data, this is especially relevant when using statistical tools implying numerous iterations of stacking. Finally, all stamps are stacked together, according to the user specified method (mean, median, or weighted-average). If weighted-average stacking is used, weights can be automatically calculated through a set of embedded methods

or input by the user. See Section 2.4 for a complete description of the automated weighting methods. The main steps taken by the main stacking algorithm can be seen in Fig. 1.

It should be noted that a difference between the results from median and mean stacking would imply a skewed distribution of the sources in the studied sample: while mean results would be driven by a few, brighter, outliers, they should have a lesser impact on the median results. Using multiple stacking methods can hence be a good diagnostic of a skewed distribution of the sample.

See Appendices A1 and A2 for examples using of LINESTACKER.

### 2.2 Edge treatment in spectra

When observing it is possible that the emission line of the target source is not centred within the observed spectral window, but falls near the edge. This results in only partial line coverage, and could inhibit the inclusion of such sources in stacking. In order to still include these sources, for each source, channels outside of the observed window are omitted from the stack. This will result in a certain range of spectral channels in the stack containing less sources. In such a situation, the noise will have higher values near the edges of the stacked spectrum. When stacking with LINESTACKER, the user gets, as an additional output, the number of sources used in every spectral bin, in order to take this effect into account when interpreting stacked data.

### 2.3 Estimating noise level

In order to calculate noise level in the data, two methods are available. The first computes the noise on the entire spectrum (through collapsing all frequency channels), while the second handles the noise channel by channel, allowing to account for noise variation with frequency (e.g. Bischetti et al. 2019). In addition, and in both cases, noise levels can be computed either through a user-defined region around the target sources, or across the entire cubes. Typically, noise levels are used as weights for sources in the stack. Calculating the noise across the entire cubes is therefore more relevant if there is only one source per cube, or at least if the dimensions of the cubes are comparable to the size of the sources. This is, however, left to the user to choose. Noise is calculated by computing the standard deviation of the data across the selected region. Computing noise levels across the entire cubes can be useful if, for example, cubes are obtained from different observations, as some may present much higher noise values (due to a lower integration time or varying observing conditions) and should therefore hold a lower weight in the stack. However, doing so leads to the inclusion of pixels far from the centre which will hold intrinsically higher noise level due to the reduction of the primary beam response. Computing the noise solely in more compact regions, centred around the target sources, allows to take the source position on the cube into account: sources closer to the phase centre should have lower noise and should hence have higher weights.

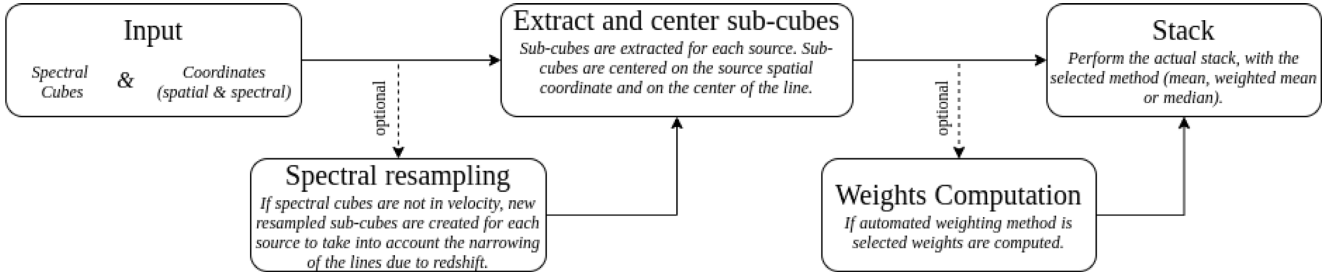
### 2.4 Automated weightings

If using weighted-average stacking the user can input customized weights for all sources individually or use the automated methods included in LINESTACKER. Automated methods include the following:

- (i) Weights inversely proportional to the noise of the cubes:

$$W_i = \frac{1}{\sigma_i^2},$$

<sup>1</sup> It should be noted that, while rms noise level goes down by a factor  $\sim \sqrt{N}$  in both mean and median stacking, the SNR may behave differently when using either. This would be the case if, for example, the studied sample is composed of many dim low SNR sources, and a few brighter high SNR sources. The brighter sources would not contribute to the median stack while they would be driving the SNR improvement in the mean stack.



**Figure 1.** Flow chart showing the main steps taken by the main stacking algorithm of LINESTACKER.

with  $W_i$  the weight of source  $i$ , and  $\sigma_i$  the standard deviation associated with source  $i$ . As stated in Section 2.3, noise can be computed across the entire cube. In which case  $W_i$  is the weight of cube  $i$ , and  $\sigma_i$  the standard deviation across cube  $i$ .

(ii) Alternatively different weights can be used for each frequency bin (depending on the individual noise values in that frequency bin), in this case,  $W_{i,j}$ , weight of source  $i$  at spectral channel  $j$  is defined as (Fruchter & Hook 2002):

$$W_{i,j} = \frac{1}{\sigma_{i,j}^2},$$

where  $\sigma_{i,j}^2$  is the standard deviation associated with source  $i$  at spectral channel  $j$ .

Following this, one can define  $W'_j$ , the total weight on stack channel  $j$  as

$$W'_j = \sum_{i=1}^n W_{i,j} = \sum_{i=1}^n \frac{1}{\sigma_{i,j}^2} = \frac{1}{\sigma_j'^2},$$

where  $n$  is the total number of sources and  $\sigma_j'$  the summed standard deviation at spectral channel  $j$ . Similarly to above, noise computation can be extended to the entire cube or restricted to regions around the target sources.

(iii) If the lines are individually distinguishable before stacking, weighting can be set proportionally to properties of the source, for example:

$$W_i = \frac{1}{A_i},$$

with  $W_i$  the weight and  $A_i$  the amplitude of line  $i$  (e.g. Stanley et al. 2019).

## 2.5 Linewidth change in frequency due to redshift: resampling

If the spectral cubes are not sampled in velocity space but in frequency (or wavelength) space instead, the width of the lines emitted by sources at different redshifts will change (getting narrower in frequency with increasing redshifts). In order to take this effect into account an option is included into LINESTACKER to resample sub-cubes according to their redshift, before stacking. Two options are available: to either use the line with the highest or lowest redshift as a reference – using the lowest redshift will lead to oversampling while using the highest implies undersampling, see below. The sub-cube associated with the line of reference is kept identical while the other are resampled such that

$$\Delta v_{\text{new}} = \frac{\Delta v_{\text{old}}}{z_{\text{ratio}}},$$

which is equivalent to

$$N_{\text{new}} = N_{\text{old}} * z_{\text{ratio}},$$

where  $\Delta v_{\text{new}}$ ,  $\Delta v_{\text{old}}$ ,  $N_{\text{new}}$ , and  $N_{\text{old}}$  are the channel size and number of channels after and before resampling, respectively. And  $z_{\text{ratio}} = \frac{1+z}{1+z_{\text{ref}}}$  where  $z$  is the redshift of the source being resampled, and  $z_{\text{ref}}$  is the redshift of the line chosen as reference. Once the new channel size is computed the resampling is performed through linear interpolation. Because such a treatment implies modification of the data – and may not be relevant if the sources stacked have similar redshifts – it is optional when using LINESTACKER. To avoid oversampling the data, we would advise against using the smallest redshift as reference.

However, because both methods imply linear interpolation of the data, it is advised to work directly with cubes binned in velocity when working with a sample with a very large redshift range. Most tools for imaging interferometric data, such as CASA, can produce cubes binned in velocity. Such algorithms used for regridding the cubes take into account extreme cases like a very sharp line that may be missed with our method due to the use of linear interpolation.

## 2.6 1D-STACKER

In addition to the cube stacker, a module allowing 1D stacking is also included in LINESTACKER. The required user inputs are the spectra and the corresponding line centres. The line centres can also be identified automatically, through Gaussian fitting, if the lines are detectable before stacking. Similarly to cube stacking, individual weights for weighted-average stacking can be input by the user or automatically calculated (see Section 2.4 for a detailed description of the different weights). Median stacking is also available.

The 1D module of LINESTACKER can be used on spectra extracted from cubes beforehand, allowing individual custom spectra extraction for each source. This can be useful if, for example, sources are known to have different spatial extent.

Unlike cube stacking, 1D-STACKER does not require CASA functions, and can be run in a PYTHON session. See Appendix A3 for an application example of 1D-STACKER.

## 3 STATISTICAL ANALYSIS TOOLS

Here, we present the statistical tools included in LINESTACKER to assist with the analysis of the stack spectra. Some tools are meant to be applied after stacking (post-stacking), to determine the robustness of the stack. Other tools can be used before stacking (pre-stacking) to get a better insight of the distribution of the stacked population.

### 3.1 Estimating the significance of the stack result

In order to estimate the significance of the stack result one can stack source-free positions and compare the result to the initial stack. For every source, a random position on the map excluding the region

around the source is chosen to be stacked. The new set of source-free targets is then stacked, similarly to the original target sources (same weighting scheme, etc.). This process is performed a large number of times (user defined, typically of the order of  $10^4$ – $10^5$ ), as a Monte Carlo process, to reach a good statistical significance of the empty positions (homogeneously probing the field and thus avoiding peculiar or peak noise values). Comparing the distribution of the results from the source-free stacks to the result of the source stack, allows for a good estimate of the significance of the stack. This method shows to what extent the result obtained from stacking the original target sources could be reproduced by stacking only noise.

### 3.2 Bootstrapping

When coupled to stacking bootstrapping can be used to probe the distribution of the parameters of the lines (amplitude of the stacked line, width of the line or integrated flux) in the original sample.

In statistics, bootstrapping methods are methods of statistical inference that allow estimation of the distribution of the sample parameters, through randomly resampling the original sample with replacement. Each source added to the new sample is randomly chosen from the entire pool of sources, allowing for multiple selections of the same source. The total number of possible combinations of resampling  $N$  elements is  $\Gamma_N^N = \frac{(2N-1)!}{N!(N-1)!}$  which, for  $N = 30$ , is of the order of  $10^{16}$ . Therefore, bootstrapping methods are most commonly combined with stochastic methods such as Monte Carlo analysis. See Appendix A3 for an example use of bootstrapping.

### 3.3 Subsampling

Subsampling consist in choosing a new, smaller sample of sources from the original target sources. This method is performed by randomly choosing a new sample size (between 1 and  $N$ , where  $N$  is the total number of sources), and randomly filling it with any of the target sources (without replacement<sup>2</sup>). The stack is then performed again, using this new set of sources. A grade is assigned to sources present in the subsample, depending on how well their stack compares to the original/full stack (the grading system depends on what specific characteristics the user is trying to probe, and hence what kind of test is applied to the data set). Performing this procedure a high number of times allows to identify if some specific subset of sources exhibit an average higher grade. Similarly to bootstrapping, this aims at studying the sample's distribution, but subsampling could allow individual identification of outliers. A good example of the use of subsampling can be found in Stanley et al. (2019), where we used it to identify sources more likely to show an outflow component out of a sample of high-redshift quasars. See Appendix A3 for an example use of subsampling.

### 3.4 Spectral rebinning

Spectral rebinning consists in changing the size of the spectral bins of each cube/spectra depending on the width of the line. As shown in Section 5.1.5 stacking lines with different widths will impact the stacked line profile: even if all lines have Gaussian profiles initially, the resulting stacked line will not be Gaussian. This can be a source of bias if trying to give a diagnostic of lines profile (e.g. while looking for outflow signatures). If the linewidth

is identifiable pre-stacking, it is possible to change the bins size, individually for each source, so that all lines span the same number of spectral channels. The resulting stacked line will then retain a Gaussian shape. It should be noted that, after such treatment, the channel size of each spectrum can be defined as in Stanley et al. (2019):  $cw_{\text{rebin}} = cw_{\text{origin}} \times \text{FWHM}_{\text{origin}}/\text{FWHM}_{\text{min}}$ , where  $cw_{\text{rebin}}$  is the channel width after rebinning,  $cw_{\text{origin}}$  is the original channel width,  $\text{FWHM}_{\text{origin}}$  is the full width at half-maximum (FWHM) of the line before rebinning, and  $\text{FWHM}_{\text{min}}$  is the FWHM of the narrowest line (used as a reference to rebin all the spectra). After stacking the channel width of the stacked spectrum can be thought of as the mean channel width of all rebinned spectra. See Decarli et al. (2018) and Stanley et al. (2019) for example use of spectral rebinning. See Appendix A2 for an example use of spectral rebinning.

## 4 SIMULATIONS

To evaluate the performances of LINESTACKER on different observable cases, we simulate data sets mimicking interferometric data. We concentrate on two different data types: the full spectral cubes (3D data cubes), and extracted spectra (1D spectra). While stacking simulated 3D data sets allows us to characterize the general performances of LINESTACKER's main algorithm, stacking 1D spectra permits the study of the effect of complex line profiles on the stack. Every set of simulations and the associated analysis is performed 100 times in the case of the 3D data sets and 1000 times for the 1D data sets to increase statistical significance. While multiple weighting schemes are available in LINESTACKER all data sets are stacked using a weighting of  $w = 1$  for all sources. Characteristics of the simulation sets are given in Tables 1 and 2. Throughout we assume  $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ ,  $\Omega_M = 0.3$ , and  $\Omega_\Lambda = 0.7$ .

### 4.1 3D simulated data sets: general characteristics

All 3D simulated data sets but two are generated using the CASA task SIMALMA (a task performing simulations of ALMA observations<sup>3</sup>) simulating ALMA cycle 6, configuration C43–2 – a short baseline configuration: max baseline = 314 m, corresponding to an angular resolution  $\sim 1 \text{ arcsec}$  at 230 GHz<sup>4</sup> and a primary beam FWHM of  $\sim 25.2 \text{ arcsec}$ . This ALMA configuration has been chosen because most studied cases focus primarily on the frequency signature of sources and less on the spatial distribution. We hence favour faster computing time over better spatial resolution. If not specified otherwise, we simulate band 6 observations: with a central frequency of 230 GHz and a bandwidth of 4 GHz. Unless specified otherwise, all simulations have a velocity resolution of  $100 \text{ km s}^{-1}$  which corresponds to a frequency resolution of  $\sim 80 \text{ MHz}$ . Furthermore, we simulate VLA observations, using CASA task SIMOBSERVE (a more general simulation task in CASA), simulating configuration C, corresponding to an angular resolution of  $0.95 \text{ arcsec}$  and a FOV of  $120 \text{ arcsec}$ . The central frequency is set to 22 GHz, falling in the middle of the  $K$  band. The total bandwidth is 500 MHz with a resolution of  $100 \text{ km s}^{-1}$ , which corresponds to a frequency resolution of  $\sim 7.3 \text{ MHz}$ . We set the total integration time to 20 min for each pointing, corresponding to typical ALMA/VLA pointing time. When the channel width is set in frequency instead of velocity (data set7 and data set8b), we chose to divide the bandwidth in 60 channels,

<sup>2</sup>Once picked the source is removed from the pool, preventing sources to be placed twice in the same sample.

<sup>3</sup><https://casa.nrao.edu/docs/TaskRef/simalma-task.html>

<sup>4</sup><https://almascience.nrao.edu/tools/proposing/proposers-guide#section-53>

**Table 1.** Characteristics of 3D simulated data sets.

Data set <sup>a</sup>	Number of sources <sup>b</sup> per cube	Number of cubes <sup>c</sup>	Line peak <sup>d</sup> randomization	Linewidth <sup>e</sup> FWHM (km s <sup>-1</sup> )	Velocity <sup>f</sup> resolution (km s <sup>-1</sup> )	Position <sup>g</sup>	FOV <sup>h</sup> (arcsec)	Source size <sup>i</sup>	Simulation <sup>j</sup> type	Redshift <sup>k</sup> range	Foreground <sup>l</sup> source
1a	1	30	No	400	100	Centre	6	PS	ALMA	7.19 < z < 7.32	No
1b	1	30	No	400	10	Centre	6	PS	ALMA	7.19 < z < 7.32	No
2a	1	30	Yes	400	100	Centre	6	PS	ALMA	7.19 < z < 7.32	No
2b	3	10	Yes	400	100	Random	28	PS	ALMA	7.19 < z < 7.32	No
3	1	30	No	200–1000	100	Centre	6	PS	ALMA	7.19 < z < 7.32	No
4	3	10	Yes	200–1000	100	Random	28	PS	ALMA	7.19 < z < 7.32	No
5a, 5b, 5c	1	30	No	400	100	Offset	28	PS	ALMA	7.19 < z < 7.32	Yes
6a	1	30	No	400	100	Centre	6	0.1–1.5 arcsec	ALMA	7.19 < z < 7.32	No
6b	1	30	No	400	100	Offset	6	0.1–1.5 arcsec	ALMA	7.19 < z < 7.32	No
7	3	10	Yes	400	~80	Random	28	PS	ALMA	5.89 < z < 7.99	No
8a	3	10	Yes	200–1000	100	Random	120	PS	VLA	4.12 < z < 4.36	Yes
8b	3	10	Yes	200–1000	~120	Random	120	PS	VLA	3.34 < z < 5.40	Yes
9	1	30	No	400	100	Centre	6	PS	ALMA	7.19 < z < 7.32	No

<sup>a</sup>Data set number ID. <sup>b</sup>Number of stacking target sources on each cube. <sup>c</sup>Total number of cubes stacked for each stack iteration. <sup>d</sup>Is the [CII] peak value randomized ('Yes') or fixed ('No'). <sup>e</sup>FWHM of the simulated Gaussian emission lines. <sup>f</sup>Velocity resolution (i.e. size of the spectral channels) of the simulated cubes. <sup>g</sup>Spatial position of the sources on the cubes: on the centre, randomized, or offset. <sup>h</sup>Total spatial extent of the simulated cube. <sup>i</sup>Physical source size, 'PS' stands for point source. Corresponding sizes once convolved with ALMA beam are ~1 arcsec for the point sources (and ~0.95 arcsec in the VLA case), from ~1 to ~1.8 arcsec in the case of extended sources. <sup>j</sup>Interferometer simulated for the observations. <sup>k</sup>Redshift range of the target sources. <sup>l</sup>Presence of bright foreground sources.

**Table 2.** Characteristics of 1D simulated data sets.

Data set <sup>a</sup>	Number of sources <sup>b</sup>	Linewidth <sup>c</sup> (km s <sup>-1</sup> )	Velocity resolution <sup>d</sup> (km s <sup>-1</sup> )	Spectral signature <sup>e</sup>	$\Delta z$ ( <sup>f</sup> )
10	30	400	100	Gaussian	True
11a	30	100–700	100	Double peaked	False
11b	30	400	100	Double peaked	True
12	30	400	100	Two Gaussian components	False
13	30	400	100	Gaussian	False

<sup>a</sup>Data set number ID. <sup>b</sup>Total number of stacked spectra for each stacking iteration. <sup>c</sup>FWHM of the simulated emission lines. <sup>d</sup>Velocity resolution (i.e. size of the spectral channels) of the simulated spectra. <sup>e</sup>Spectral shape of the simulated emission lines. <sup>f</sup>Usage, or not, of uncertainties on the observed redshift, resulting in uncertainties in the central line position.

meaning that the average channel width in frequency may slightly differ from the constant 100 km s<sup>-1</sup> of the other sets (see Table 1).

Every source's image is generated through a component list (a list of functional representation of the sky brightness) which serves as a skymodel (model image of the observing field) whose observation is then simulated with the CASA task SIMALMA (or SIMOBSERVE in the case of VLA data). From this simulation we get the visibilities (i.e. the interferometric data), which are then imaged using the CASA task CLEAN,<sup>5</sup> the images being the final product that will be stacked.

In the absence of bright foreground sources the number of CLEAN iterations is set to 0 (i.e. solely imaging), otherwise CLEAN is performed down to a given threshold (~10 times fainter than the bright foreground source). For every source the spectrum consists of an emission line with noise, the noise being directly generated through the SIMALMA or SIMOBSERVE task depending on the data set.

Every pointing position is selected randomly within a circle of 10 arcmin radius, centred at J2000 3<sup>h</sup>31<sup>m</sup>00.00–27<sup>d</sup>40<sup>m</sup>00.00.<sup>6</sup>

Every sky model consists of 30 sources, distributed on either 30 or 10 images (either 1 source per image in the centre, or 3 sources at random positions, see Table 1). The sample size was chosen such that it is small enough to be representative of most stack cases and

large enough to present a relevant noise reduction.<sup>7</sup> When only one source is simulated in the middle of each image, the data cubes are produced with a size of 6 arcsec × 6 arcsec, which is a sufficient size as we are interested in the central source and not the edges. When multiple sources are generated on one image, sources are randomly distributed within the primary beam of the simulated data, which is 25.2 arcsec, and the cubes are imaged to a larger size of 28 arcsec × 28 arcsec. This allows to test for effects that may arise with sources far from the centre. In both cases the angular resolution is set to 0.25 arcsec pixel<sup>-1</sup>, the synthesized beam having a size of ~1 arcsec.

We create 15 × 2 different simulated data sets: 15 data sets, with specific characteristics, are tested in both a high and a low SNR. Each version serves a different purpose. The high SNR (~200) sets allow for a near-ideal test of the algorithm, where noise is almost negligible, and shows the best result one can expect. Low SNR (~1) sets test the reduction of noise through stacking, but also the limitations when applied to a case where noise levels are important. We mostly concentrate on point like sources but we also examine cases with extended sources. For flux conservation near the edges of the map every simulation has been primary beam corrected. We note that primary corrections should be done by the user as it is not a part of the image stacking routine of LINESTACKER.

<sup>5</sup>The CLEAN algorithm was originally described in Högbom (1974).

<sup>6</sup>Corresponding to the Extended Chandra Deep Field South field (ECDFS; Arnouts et al. 2001) it has been chosen arbitrarily and does not have any impact on the produced data and its analysis.

<sup>7</sup>When stacking, noise goes down as  $\sqrt{N}$  where  $N$  is the number of stack positions; 30 sources corresponds to an average noise reduction of  $\sqrt{30} \sim 5.5$  (see Section 6.1.2 and Fig. 8).

## 4.2 Spectra

For our simulated data sets we assume a sample of high-redshift galaxies. We chose two scenarios: [C II] emission lines, at  $z \sim 7$ , observed with ALMA in band 6 and CO(3–2) at  $z \sim 4$  observed with VLA in band 4. In both cases, the spectra consist solely of the emission line and noise. These lines have been chosen arbitrarily, as typically observed emission lines at high redshift. The lines and redshift choice should have no impact on the produced data and hence on the conclusions drawn in this article. The high SNR samples are generated simulating emission lines peaking at 100 mJy, versus 0.5 mJy for the low SNR samples; 100 mJy is typically brighter than what would be expected, this value is picked in order to have near-ideal, very high SNR values.

In some sets (see Table 1), the amplitude is not fixed, but randomized uniformly over a given range (i.e. top hat distribution). The ranges: from 50 to 150 mJy and from 0.1 to 0.9 mJy for the high and the low SNR sets, respectively, are chosen such that the mean amplitude reaches the same value as the fixed one, to allow easier comparison between sets. When fixed, the line FWHM is set to  $400 \text{ km s}^{-1}$  (which is typically for high-redshift galaxies, e.g. Bothwell et al. 2013; Gullberg et al. 2015, 2018; Decarli et al. 2018). When random, the FWHM is randomized uniformly from  $200 \text{ km s}^{-1}$  to  $1000 \text{ km s}^{-1}$ . Exact values do not have a big impact and are chosen to be representative of observed values where large variations in line widths are seen (e.g. Bothwell et al. 2013; Decarli et al. 2018). Throughout the text line width will refer to the FWHM of the line.

## 4.3 Simulated data sets: 3D

In Table 1, we give the characteristics of each 3D simulated data set, and here we provide further information on each data set

### 4.3.1 Most basic case – set1a

To test the performance of LINESTACKER in a near-ideal, fully controlled simulation, we start with the simplest [C II] line simulation, using a realistic spectral resolution. It consists of a point source in the centre of every image, with all the same spectrum which is simply a [C II] emission line with a width of  $400 \text{ km s}^{-1}$ . A total of 30 cubes are simulated, with a channel width of  $100 \text{ km s}^{-1}$ , which is typical of stacked data where the channels are collapsed together beforehand in order to maximize the amplitude of the signal to the detriment of frequency resolution. This set will be considered as a reference, to be compared to the following sets to see the impact of the tested parameters.

### 4.3.2 Most basic case, high spectral resolution – set1b

When the emitted line is observed, it is binned to a finite number of channels. When the number of channels across the line is small (of the order of  $\sim 5$  across the line FWHM) the line's amplitude will be systematically underevaluated. To test this effect, and show that the amplitude loss is due to the finite spectral resolution and not to some other intrinsic bias, we repeat data set1a with a much better spectral resolution. The number of channels is 10 times higher than in set1a for the same bandwidth. Even though such a spectral resolution is better than the one typically expected, the goal here is to build a near-perfect reference set, estimating the signal loss that can be expected due to the finite number of channels.

### 4.3.3 Random amplitude – set2a

Data set2a is designed to check the effect of a given distribution of the lines amplitude on the stack. Amplitudes of each line are randomized uniformly on a range of 50–150 and 0.1–0.9 mJy for the high and low SNR sets, respectively. The width of every line is kept at  $400 \text{ km s}^{-1}$ .

### 4.3.4 Random amplitude and position – set2b

Stacking can be expected to be used on individual cubes containing multiple sources. Because LINESTACKER performs stacking on sub-cubes the effect coming from such configuration should be limited. However, such a case of configuration is likely to be a common user case, and therefore is relevant to test. Here, unlike set2a the number of sources per image is set to 3, with positions uniformly randomized across the whole field. Amplitudes of each line are still randomized uniformly on a range of 50–150 and 0.1–0.9 mJy for the high and low SNR sets, respectively. The width of every line is kept at  $400 \text{ km s}^{-1}$ .

### 4.3.5 Random linewidth – set3

Data set3 aims at quantifying the effect on the stacked line when including lines with a range of line widths. This situation is expected in real data since galaxies exhibit a range of line width due to different masses, orientation and level of turbulence. In order to properly evaluate the consequence of such an inhomogeneous distribution this data set will be based on set1a but with a FWHM randomly picked from a uniform top-hat distribution from 200 to  $1000 \text{ km s}^{-1}$ .

### 4.3.6 Random amplitude, linewidth, and position – set4

Data set4 is a combination of data set2b and 3, with randomized positions on the field as well as randomized amplitude and width of the line.

### 4.3.7 Bright foreground source in the centre – sets 5a, 5b, and 5c

These data sets aim at quantifying the effect that could arise from stacking faint sources from a map containing a bright central point source. Such a case can be expected when stacking faint peripheral sources present in a field centred on a bright source. The presence of bright sources (continuum or spectral line) affects the quality of the interferometric image products, as imperfect modelling of the bright source can leave artefacts in the final data cubes; such artefacts increase the noise and could potentially affect the stacking result. In the simulations, the central bright source has a continuum flux density of 1 and 0.1 Jy for high and low SNR simulations, respectively. While 1 Jy is higher than typically expected values, the foreground source has to be brighter than the amplitude of the line of the target sources – which are already very bright in the high SNR sets. The target sources have the same properties as sources from data set1a. To properly diagnose the impact of the bright foreground source we created three type of data sets. In data set5a the target source is located at a distance of 2.5 arcsec from the bright source, 5 arcsec in data set5b and 10 arcsec in data set5c. The bright foreground source was removed using the CASA task CLEAN, and the stacking was performed on the residual image.

### 4.3.8 Extended sources – set6a

Data set6a is composed of extended sources, to test cases where sources are resolved. We investigated such a case by simulating

a data set similar to data set1a, but containing extended sources. Sources have Gaussian shape, with their orientation and size taken at random. The length of the major and minor axes are randomized uniformly between 0.1 and 1.5 arcsec (this is the physical extent of the sources and corresponds to a size range of  $\sim 1.00$ – $1.80$  arcsec after convolution with the ALMA beam). The orientation and size of the sources will have an important impact since different sources will not be spanning the same area. Each source is simulated with a spatially integrated peak flux value of 100 and 0.5 mJy for high and low SNR sets, respectively.

#### 4.3.9 Extended sources with an offset – set6b

Typically, the stacking positions of sources are defined based on observations at other wavelength. This can induce an offset between stacking position and source position at the observed wavelength. In the case of the observation of emission lines the region of interest may be offset from the position defined at a different wavelength (e.g. [C II] may be distributed differently on the studied galaxy than dust or the stellar population e.g. Rybak et al. 2019; Fujimoto et al. 2019). In order to reproduce and quantify such an effect, we produced another data set, similar to data set6a, but where the stacking position is offset compared to the source position. This offset is taken at random within a uniform distribution between 0 and 1 arcsec. This will impact the stack results since sources will not be properly aligned when stacked. It should be noted that point sources are of course also subject to potential offset. However, due to the nature of the observations, their observed size will be equal to size of the synthesized beam. The effect coming from the offset of point sources is hence comparable to the effect tested in this data set.

#### 4.3.10 Random central frequency, larger redshift range – set7

Data set7 is based on data set2b but studies sources spanning a larger redshift range and thus a larger observational central frequency range. The redshifts are drawn randomly from a uniform distribution in the range  $5.89 < z < 7.99$  (corresponding to a range of possible observed central frequencies covering the entire band 6). The bin width is set to a fixed frequency size of  $\sim 67$  MHz, 60 channels over the 4 GHz bandwidth, corresponding to  $\sim 80$  km s $^{-1}$  for the average frequency. Trying to stack observations done on a large frequency range will have different implications, one of the most direct is the shape and size of the beam, which changes with frequency and could hence be different from one observation to the other. In addition, because of the large redshift range, it is necessary to take into account the change of width of the lines when working in frequency. Each sub-cube is therefore resampled before stacking according to the method described in Section 2.5. The highest redshift is used as reference for resampling, leading to a reduced number of channels after resampling.

#### 4.3.11 VLA type simulation – set8a

We simulate CO(3–2) observations with VLA. To both showcase that the tool is not solely usable on ALMA type data and also to study cases with a larger field of view and more polluting bright sources. On each cube we simulate three target sources with characteristics similar to data set4. Additionally, we add: one very bright foreground point source with a line amplitude of 1 Jy in both the high and low SNR sets, as well as two ‘medium bright’ foreground point sources with amplitude 100 and 10 mJy in the high and low SNR sets,

respectively. All sources positions are uniformly randomized across the whole field. The frequency coverage for this set corresponds to a redshift range of  $4.12 < z < 4.36$ . It should be noted that since the line amplitudes of the target sources are kept the same as in the other sets, the SNR will be slightly reduced (due to the lower sensitivity of simulated VLA data):  $\text{SNR} \sim 1$  in our simulated ALMA data corresponds to  $\text{SNR} \sim 0.8$  in our simulated VLA data.

#### 4.3.12 VLA type simulation, larger redshift range – set8b

Set8b is an extension of set8a to the entire *K* band, the central frequency is chosen at random so that the entire bandwidth is contained between 18 and 26.5 GHz. Bandwidth is kept the same as in set8a but channel size is kept constant in frequency at  $\sim 8.3$  MHz: 60 channels over the 500 MHz bandwidth, corresponding to  $\sim 112$  km s $^{-1}$  for the average frequency. The goal is similar to set7: to study the impact of the large redshift range on the stack, but with a larger redshift range: from  $3.34 < z < 5.4$ . Similarly to set7 a resampling is applied to each sub cube to take into account the redshift of the sources.

#### 4.3.13 1D stack of spectra extracted from cube – set9

To allow easy comparison between the 3D and the 1D data sets, we built a data set where the spectra are extracted from individual cubes and then stacked using the 1D module of LINESTACKER. The sources’ properties are the same as set1a and the spectra are extracted from the central pixel.

### 4.4 Simulations data sets: 1D

The 1D simulated data sets allow us to test some specific spectral signatures more easily than when using full 3D simulations: we examine cases of double peaked line profiles, outflow signatures, the effect of redshift uncertainties on the stack, and the effect of stacking lines located on the edge of the observed spectral window. The data sets are generated with a bandwidth of 3000 km s $^{-1}$ , and a resolution of 100 km s $^{-1}$ . These spectral only simulations are not generated through CASA, allowing faster computing time. Individual spectra are generated by creating individual Gaussian components and adding randomly generated Gaussian noise on top of each channel. Similarly to the 3D sets two SNR configuration: one with high (pre-stacking) SNR ( $\sim 200$ ) and one with an SNR of order unity. Lines are generated with an amplitude of 200 mJy in the high SNR data sets and 1 mJy in the low SNR data sets. The noise follows a Gaussian distribution centred at 1 mJy. Linewidths differ from simulation to simulation, see Table 2 for a complete description of the characteristics of each simulated data set.

#### 4.4.1 Diagnostic of redshift uncertainties – set10

One of the biggest challenges when stacking lines of distant galaxies arises from redshift uncertainties. Every simulation performed previously has been computed expecting a perfectly good knowledge of the redshift. But, realistically, redshift is never known with 100 per cent accuracy, and redshift uncertainties can cause lines not to be stacked on the same central frequency.

Consequently, the amplitude and width of the stacked line will be washed out and potentially become indiscernible from the noise. In order to quantify this problem we construct spectra data sets, and test different levels of redshift uncertainties ( $\Delta z$ ). The linewidths are all set to a velocity of 400 km s $^{-1}$ . The redshift uncertainties,  $\Delta z$ , are

set to values of 0, 0.001, 0.005, 0.01, 0.05, and 0.1 (corresponding to velocity shifts of  $\sim 0, 36, 180, 360, 1800, 3600 \text{ km s}^{-1}$ ) and the observed redshifts of the sources in each simulated data set are chosen randomly between  $z_{\text{true}} \pm \Delta z$ , where  $z_{\text{true}}$  is the real redshift of the line. For every  $\Delta z$ , 1000 data sets are created.

#### 4.4.2 Double peaked spectrum – set11a

Rotational signatures from galaxies can be seen as double peaked line profile. In order to study such case we create a simulated 1D sample with such properties. The spectra were designed to have a distance,  $D$ , between the two peaks ranging uniformly from 200 to  $600 \text{ km s}^{-1}$ , and an amplitude of  $10 \text{ mJy}$ , while the width of the line ranges from 100 to  $700 \text{ km s}^{-1}$  (with the same linewidth for both components). The central frequency for stacking is taken as the centre of the two peaks.

#### 4.4.3 Double peaked spectrum with redshift uncertainties – set11b

This data set, like the previous one, shows spectra with double peak line profile. This time all lines are simulated with the same characteristics, a width of  $400 \text{ km s}^{-1}$  and a distance between the peaks of  $400 \text{ km s}^{-1}$ . An uncertainty on the redshift is added. Like data set10 the  $\Delta z$  values of 0, 0.001, 0.005, 0.01, 0.05, and 0.1 are tested. Hence, trying to quantify the distortion of this specific spectral signature when confronted to redshift uncertainties, and the extent to which one can recover such spectral characteristics when stacking.

#### 4.4.4 Outflows – set12

In cases where outflows are present, the galaxy’s emission lines could include a second, fainter and broader, component (see Stanley et al. 2019). In many cases observation will not be deep enough to detect this signature, and therefore stacking is a useful tool. From a testing perspective, studying outflows allows us to use line stacking in a different fashion: to look for specific spectral signatures below the noise, while the main line is visible: using the main line not as a source but as a reference to stack signal below it. All the lines are simulated with a width of  $400 \text{ km s}^{-1}$ , and a broad component with a width of  $1000 \text{ km s}^{-1}$ . The amplitude of the broad component is one-tenth of the amplitude of the main line. The two SNR configurations (200 and 1) are based on the broad component amplitude. The amplitude of the main lines will be  $2 \text{ Jy}$  in the high SNR configuration and  $10 \text{ mJy}$  in the low SNR regime – corresponding to an amplitude of the broad component of 200 and  $1 \text{ mJy}$  in the high and low SNR configuration, respectively.

#### 4.4.5 Lines on the edge – set13

One of the strengths of LINESTACKER is its ability to stack lines located on the edge of the observed spectral window. To showcase this capability and test its performance we produced a data set where all lines are located on the edge. While it is unrealistic to have a data set where all lines are so far from the centre, we decided to test such an extreme case to demonstrate the expected result in the worst-case scenario. All simulation parameters are similar to data set10 with a redshift uncertainty  $\Delta z = 0$ , but lines are centred at a distance uniformly randomized between 0 and  $200 \text{ km s}^{-1}$  from either of the spectral edges. This means that a significant part of the line will be outside of the observed spectral window – in the worst case where the line is centred exactly on the edge, 50 per cent of the line will not

be observed, in the best case, when the line is centred at a distance of  $\frac{1}{2} \text{ FWHM} = 200 \text{ km s}^{-1}$ , roughly 30 per cent of the line is outside of the observed window.

## 5 RESULTS

We used LINESTACKER on the simulated data sets described in Section 4. From our stacking analysis we extract the amplitude and the width of the line as well as the integrated flux and compare them to the mean input values. Stacking is performed with subimages of size  $16 \times 16$  pixels and 32 spectral channels. To retrieve the amplitude and width we fit the resulting stacked line with a Gaussian and extract the fit’s parameters. The integrated flux is computed by summing the flux in each channel covering the detected line and multiplying by the channel width. The spectral area of integration has a size of two times the input line FWHM, centred on the line central frequency. When the amplitude and/or width are simulated at random we calculate their average and use this average for comparison with stack values. Presented reconstruction fractions are the ratio  $(1 - |1 - \frac{\text{measured}}{\text{expected}}|) \times 100$  for each parameter in both low and high SNR configurations (chosen such that the reconstruction rate is always  $< 100$  per cent).

The results, average of all the stacks, are presented in Table 3. Presented standard deviations are the standard deviations of the stack results across the studied simulation set. In the case of 1D simulations, different parameter reconstruction are tested for every simulation, results from stacking analysis of each one-dimensional data sets are shown in Tables 4–6.

### 5.1 Stacking results from 3D simulated data sets

We first present the results from stacking the 3D simulated data sets. For every data set we analyse separately the average results from the two SNR cases and compare them to the average input parameters. The presented results, for a given data set and a given SNR, are average of 100 stacks. Each studied parameters (amplitude, width and integrated flux) are discussed individually.

#### 5.1.1 Most basic case – set1a

After stacking, and fitting our result with a Gaussian, we find a reconstruction fraction of  $94.0 \pm 0.22$  per cent,  $92.7 \pm 0.23$  per cent, and  $99.6 \pm 0.16$  per cent for the amplitude, width and integrated flux of the line in the high SNR configuration. Similar results are found for the low SNR sets:  $95.4 \pm 5.2$  per cent,  $97.4 \pm 5.21$  per cent, and  $97.1 \pm 2.36$  per cent. As it will be shown in the next section the missing 6 per cent in the amplitude reconstruction are systematic errors due to the low velocity resolution. Even though line amplitude is at the same level as the noise in the faint sets, the reconstruction is extremely accurate. One should note that reconstruction of the integrated flux has a stronger dependence to SNR than the amplitude of the line does. This is due to the fact that, to have a proper integrated flux reconstruction, one will need proper reconstruction in every bin containing the line, meaning also the channels containing the outer part of the line profile, which, if the SNR is low, will be under the noise level. The reconstruction rates obtained from this data set will be used as references when rating success of following data sets, as data set1a has been designed to be the simplest data set, and will hence yield the best results. Example results from stacking data set1a can be seen in Fig. 2.

**Table 3.** Stacking results from all 3D simulations. Presented results are obtained by averaging 100 stacks. Amplitude and linewidth are obtained through Gaussian fitting, while integrated flux is obtained by integrating a given number of channels (see Section 5). Presented errors are computed standard deviation of the given parameter in the 100 stacks.

Data set <sup>a</sup>	Mean line amplitude <sup>b</sup> (mJy)			Mean linewidth <sup>c</sup> (km s <sup>-1</sup> )			Mean integrated flux <sup>d</sup> (Jy km s <sup>-1</sup> )		
	Stack		Simulated	Stack		Simulated	Stack		Simulated
	Value	Std dev		Value	Std dev		Value	Std dev	
1a Bright	93.98	0.22	100.0	429.1	0.91	400.0	42.38	0.07	42.57
1a Faint	0.477	0.026	0.5	410.5	20.82	400.0	0.206	0.005	0.212
1b Bright	99.92	0.025	100.0	403.8	0.25	400.0	42.58	0.021	42.57
1b Faint	0.496	0.019	0.5	408.3	20.93	400.0	0.213	0.010	0.212
2a Bright	93.38	4.630	99.33	429.2	0.971	400.0	42.12	2.081	42.29
2a Faint	0.478	0.046	0.498	411.0	22.95	400.0	0.205	0.017	0.212
2b Bright	93.92	5.71	99.85	429.5	4.59	400.0	42.39	2.60	42.51
2b Faint	0.472	0.042	0.500	429.1	32.35	400.0	0.211	0.018	0.213
3 Bright	90.20	1.59	100.0	637.0	47.87	592.1	60.33	5.18	63.03
3 Faint	0.462	0.020	0.5	599.9	51.05	601.6	0.293	0.023	0.320
4 Bright	90.49	5.34	100.3	645.1	41.73	596.4	61.31	5.85	63.71
4 Faint	0.452	0.038	0.500	636.6	68.10	593.8	0.301	0.033	0.316
5a Bright	92.84	0.308	100.0	430.4	0.861	400.0	42.52	0.082	42.57
5a Faint	0.464	0.026	0.5	429.6	25.16	400.0	0.210	0.014	0.212
5b Bright	92.43	0.266	100.0	430.6	1.156	400.0	42.33	0.065	42.57
5b Faint	0.467	0.024	0.5	433.7	30.83	400.0	0.213	0.012	0.212
5c Bright	92.04	0.273	100.0	430.4	0.932	400.0	42.14	0.098	42.57
5c Faint	0.468	0.025	0.5	421.0	29.11	400.0	0.206	0.016	0.212
6a Bright	95.51	2.66	100.0	428.6	1.006	400.0	41.33	1.14	42.57
6a Faint	0.446	0.054	0.5	356.4	32.79	400.0	0.164	0.015	0.212
6b Bright	92.04	1.63	100.0	428.6	1.095	400.0	39.83	1.05	42.57
6b Faint	0.405	0.052	0.5	332.2	35.11	400.0	0.138	0.015	0.212
7 Bright	95.40	4.776	99.44	415.6	1.698	400.0	42.13	2.087	42.34
7 Faint	0.496	0.052	0.507	410.7	32.31	400.0	0.214	0.025	0.216
8a Bright	90.88	4.353	99.81	633.5	46.20	596.3	63.05	5.564	63.35
8a Faint	0.438	0.087	0.507	629.7	87.33	611.0	0.289	0.040	0.330
8b Bright	90.77	4.788	100.8	639.0	48.93	599.8	63.58	6.389	64.39
8b Faint	0.411	0.061	0.497	651.6	111.1	608.0	0.286	0.044	0.322
9 Bright	94.00	0.21	100.0	429.3	0.92	400.0	42.86	0.07	42.57
9 Faint	0.475	0.021	0.5	410.5	17.98	400.0	0.205	0.006	0.212

<sup>a</sup>Data set number ID. <sup>b</sup>Average line amplitude from all the 100 simulations of the studied set, and from their corresponding stacks. <sup>c</sup>Average FWHM of the emission lines, from all the 100 simulations of the studied set, and from their corresponding stacks. <sup>d</sup>Average integrated flux from all the 100 simulations of the studied set, and from their corresponding stacks.**Table 4.** Stacking results from 1D simulated data set10. Presented results are obtained by averaging 1000 stacks. Parameters are obtained through fitting. Presented errors are computed standard deviation of the given parameter in the 1000 stacks.

Data set	$\Delta z^a$	Mean line amplitude <sup>b</sup> (mJy)		Mean linewidth <sup>c</sup> (km s <sup>-1</sup> )		Mean integrated flux <sup>d</sup> (mJy km s <sup>-1</sup> )	
		Value	Std dev	Value	Std dev	Value	Std dev
10 Bright	0	200.01	0.13	400	2.1	83289	53
10 Faint	0	1.02	0.13	400	64	416	52
10 Bright	0.001	195.66	1.11	408	2.3	83027	151
10 Faint	0.001	1.01	0.13	407	67	416	51
10 Bright	0.005	138.91	9.71	579	44	75592	2219
10 Faint	0.005	0.72	0.12	581	123	381	50
10 Bright	0.01	87.73	10.63	924	127	57866	4905
10 Faint	0.01	0.47	0.14	910	289	292	57
10 Bright	0.05	—	—	—	—	—	—
10 Faint	0.05	—	—	—	—	—	—
10 Bright	0.1	—	—	—	—	—	—
10 Faint	0.1	—	—	—	—	—	—
13 Bright	0	200.0	0.11	399.99	0.30	85155	158
13 Faint	0	0.99	0.42	398	74	421	168

<sup>a</sup>Average redshift uncertainty, leading to uncertainty of the line-centre position. <sup>b</sup>Average resulting stacked line amplitude.<sup>c</sup>Average resulting stacked line FWHM. <sup>d</sup>Average resulting stacked line integrated flux.

**Table 5.** Stacking results from 1D simulated data sets with double peak profiles (set11a and set11b). Presented results are obtained by averaging 1000 stacks. Parameters are obtained through fitting. Presented errors are computed standard deviation of the given parameter in the 1000 stacks.

Data set <sup>a</sup>	$\Delta z^b$	Mean line amplitude <sup>c</sup> (mJy)			Mean peak distance <sup>d</sup> (km s <sup>-1</sup> )			Mean linewidth <sup>e</sup> (km s <sup>-1</sup> )			Mean integrated flux <sup>f</sup> (mJy km s <sup>-1</sup> )			
		Stack		Simulated	Stack		Simulated	Stack		Simulated	Stack		Simulated	
		Value	Std dev		Value	Std dev		Value	Std dev		Value	Std dev	Value	Std dev
11a	0	182.4	7.2	200.0	413.7	23.9	400.0	428.2	36.7	400.9	168 778	9490	172 595	8841
11a	0	0.91	0.21	1.0	414.3	334.7	400.0	427.4	157.4	399.9	768	263	805	47
11b	0	199.9	0.13	200.0	400.0	0.15	400.0	400.0	0.28	400.0	169 689	64	170 301	–
11b	0	1.007	0.170	1.0	400.0	419.7	400.0	397.8	97.62	400.0	850	65	851	–
11b	0.001	195.6	2.33	200.0	400.0	0.16	400.0	408.9	4.77	400.0	169 566	71	170 301	–
11b	0.001	0.98	0.17	1.0	400.7	337.3	400.0	409.7	99.19	400.0	849	66	851	–
11b	0.005	136.9	29.21	200.0	399.9	3.99	400.0	584.2	90.00	400.0	164 482	1604	170 301	–
11b	0.005	0.68	0.216	1.0	399.6	336.0	400.0	586.4	154.8	400.0	822	66	851	–
11b	0.01	84.72	46.83	200.0	400.2	114.4	400.0	944.4	221.8	400.0	145 492	5947	170 301	–
11b	0.01	0.41	0.26	1.0	337.3	336.5	400.0	985.0	247.7	400.0	726	73	851	–
11b	0.05	–	–	200.0	–	–	400.0	–	–	400.0	–	–	170 301	–
11b	0.05	–	–	1.0	–	–	400.0	–	–	400.0	241	92	851	–
11b	0.1	–	–	200.0	–	–	400.0	–	–	400.0	–	–	170 301	–
11b	0.1	–	–	1.0	–	–	400.0	–	–	400.0	124	88	851	–

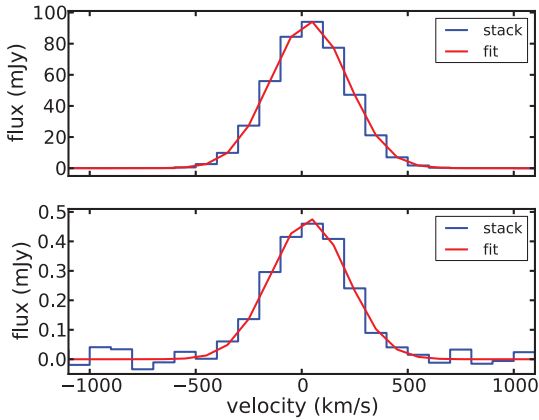
<sup>a</sup>Data set number ID. <sup>b</sup>Average redshift uncertainty, leading to uncertainty of the line-centre position. <sup>c</sup>Average line amplitude, from the stack and from simulations. <sup>d</sup>Average distance between the two peaks, from the stack and from simulations. <sup>e</sup>Average single line FWHM (both lines have the same width), from the stack and from simulations. <sup>f</sup>Average line integrated flux, from the stack and from simulations.

**Table 6.** Stacking results from 1D simulated containing an outflow component (data set12). Presented results are obtained by averaging 1000 stacks. The characteristics of both components are obtained through Gaussian fitting with two Gaussian. Presented errors are computed standard deviation of the given parameter in the 1000 stacks.

Mean main line amplitude <sup>a</sup> (mJy)			Mean outflow amplitude <sup>b</sup> (mJy)			Mean main line width <sup>c</sup> (km s <sup>-1</sup> )			Mean outflow width <sup>d</sup> (km s <sup>-1</sup> )		
Stack		Simulated	Stack		Simulated	Stack		Simulated	Stack		Simulated
Value	Std dev		Value	Std dev		Value	Std dev		Value	Std dev	
1999.	0.46	2000.0	200.0	0.47	200	399.9	0.07	400.0	999.9	1.086	1000.0
10.00	–	10.0	0.99	0.13	1.0	401.5	10.9	400.0	984.5	131.1	1000.0

<sup>a</sup>Average main component amplitude, from the stack and from simulations. <sup>b</sup>Average second component amplitude, from the stack and from simulations.

<sup>c</sup>Average main component line FWHM, from the stack and from simulations. <sup>d</sup>Average second component line FWHM, from the stack and from simulations.



**Figure 2.** Two example stack of 30 sources from data set1a and the corresponding Gaussian fit. Top: High SNR configuration. Bottom: Low SNR configuration.

### 5.1.2 Most basic case, high-frequency resolution – set1b

This data set aimed at simulating a near-perfect configuration, differing from set1a with a velocity resolution of 10 km s<sup>-1</sup>, 10 times higher than previously. As expected we reach reconstruction fractions of 99.9 ± 0.02 per cent, 99.0 ± 0.06 per cent, and 100.0 ± 0.05 per cent;

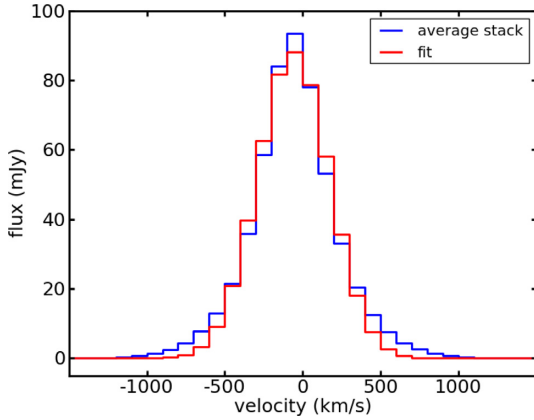
and 99.4 ± 3.8 per cent, 97.7 ± 5.23 per cent, and 96.5 ± 4.72 per cent for the amplitude, width, and integrated flux of the line compared to the input of the high and low SNR sets, respectively. Showing, as mentioned in the previous section, that the missing 6 per cent, when reconstructing the amplitude of the line in the set1a, were systematic errors due to the channels width. The almost perfect reconstruction in the low SNR case shows that random errors should be negligible in our stacking setup.

### 5.1.3 Random amplitude – set2a

This set has a uniform distribution of amplitude between 50–150 and 0.1–0.9 mJy in the high and low SNR sets, respectively. Here again we find a good reconstruction fraction of 94.0 ± 4.66 per cent, 92.7 ± 0.24 per cent, and 99.6 ± 4.92 per cent of the amplitude, width, and integrated flux in the high SNR case, and 95.9 ± 9.24 per cent, 97.2 ± 5.74 per cent, and 96.7 ± 8.02 per cent in the low SNR case.

### 5.1.4 Random amplitude and position – set2b

This set has a uniform distribution of amplitude between 50–150 and 0.1–0.9 mJy in the high and low SNR sets, respectively, with three sources per image at random positions. Here again we find a good reconstruction fraction of 94.1 ± 5.72 per cent, 92.6 ± 1.15 per cent,



**Figure 3.** Average stacked line from the 100 realizations of set3, high SNR. Here it is clear – by comparing with the fit – that, even if all input lines are Gaussian, the output stacked line is not anymore. One direct consequence of such behaviour is the underevaluation of the amplitude of the line, if done through fitting.

and  $99.7 \pm 6.12$  per cent of the amplitude, width, and integrated flux in the high SNR case, and  $94.4 \pm 8.4$  per cent,  $92.7 \pm 8.09$  per cent, and  $99.1 \pm 8.45$  per cent in the low SNR case, respectively. While flux loss could have been expected with sources far from the pointing centre, it is not observed here due to the use of primary beam correction.

#### 5.1.5 Random width – set3

Here, amplitudes and positions are fixed. The linewidth however is randomized, drawn from a flat distribution, between 200 and 1000  $\text{km s}^{-1}$ . When stacking Gaussian lines with different linewidth the resulting stacked line will not conserve a Gaussian shape, and can introduce what could be interpreted as a second component to the line. Fitting the stacked line with a Gaussian to extract parameters will hence give slightly biased results. As shown on Table 3, the amplitude is still well reconstructed – at  $90.2 \pm 1.59$  per cent and  $92.4 \pm 4.0$  per cent for the high and low SNR simulation sets, respectively. Such a difference with set1a is due to the use of Gaussian fitting to extract parameters. As a result the fitted amplitude is lower than expected (see Fig. 3). It is, however, interesting to note that, if, instead of fitting, we look at the maximum bin value, it shows  $\sim 96$  per cent reconstruction in both SNR cases (see Section 6.3). The width and mean integrated flux are well reconstructed, with a  $92.4 \pm 8.08$  per cent and  $95.7 \pm 8.22$  per cent reconstruction for the high SNR case, and  $99.7 \pm 8.49$  per cent and  $91.7 \pm 7.19$  per cent for the low SNR one, respectively.

#### 5.1.6 Random amplitude, width, and position – set4

This set includes random amplitudes, random positions, and random width, it is a combination of data sets 2 and 3 and has similar results. Similarly to set3 the reconstructed amplitude is only reconstructed at  $90.2 \pm 5.32$  per cent and  $90.4 \pm 7.6$  per cent in the high and low SNR setup. This is again due to the intrinsic bad fit, due to a non-Gaussian shaped stack line. Width and integrated flux are reconstructed at  $91.8 \pm 7.0$  per cent and  $96.2 \pm 9.18$  per cent in the high SNR case, and  $92.8 \pm 11.47$  per cent and  $95.2 \pm 10.44$  per cent in the low SNR one.

#### 5.1.7 Bright foreground source in the centre – sets 5a, 5b, and 5c

These sets aimed at studying cases where target sources are faint and at a fixed distance from a central bright foreground continuum source. Sources’ characteristics are similar to set1a (but target sources are not in the cube centre). The amplitude, width and integrated flux are reconstructed at  $92.8 \pm 0.31$  per cent,  $92.4 \pm 0.22$  per cent, and  $99.9 \pm 0.19$  per cent, respectively, in the high SNR case, and  $92.8 \pm 6.2$  per cent,  $94.2 \pm 5.12$  per cent, and  $97.6 \pm 8.02$  per cent, respectively, in the low SNR one for set5a. Set 5b shows reconstruction rates of  $92.4 \pm 0.27$  per cent,  $92.4 \pm 0.29$  per cent, and  $99.4 \pm 0.15$  per cent for the amplitude, width, and integrated flux, respectively, in the high SNR case and  $93.4 \pm 4.8$  per cent,  $91.6 \pm 7.71$  per cent, and  $99.5 \pm 5.66$  per cent, respectively, in the low SNR one. Finally, set 5c shows reconstruction rates of  $92.0 \pm 0.27$  per cent,  $92.4 \pm 0.23$  per cent, and  $99.0 \pm 0.23$  per cent for the amplitude, width, and integrated flux, respectively, in the high SNR case and  $93.6 \pm 5.0$  per cent,  $94.8 \pm 7.28$  per cent, and  $97.2 \pm 7.55$  per cent, respectively, in the low SNR one. The presence of the bright source is well handled through CLEANING and seems to have close almost no impact on the stack.

#### 5.1.8 Extended sources – set6a

For the cases where sources are extended (set 6a and 6b), the stacked sources are fitted with a two-dimensional Gaussian and the stacked spectra are extracted from an ellipsoidal region centred on the stacked source, where the size of each axis is set to two times the corresponding FWHM. Fluxes values are converted from  $\text{Jy beam}^{-1}$  to  $\text{Jy pixel}^{-1}$ . This set shows good reconstruction fractions:  $95.51 \pm 2.66$  per cent,  $92.85 \pm 0.25$  per cent, and  $97.08 \pm 2.67$  per cent in the high SNR case and  $89.2 \pm 10.8$  per cent,  $89.10 \pm 8.20$  per cent, and  $77.36 \pm 7.1$  per cent in the low SNR case, for amplitude, width, and integrated flux, respectively. Indicating that, if the stacking position is well known, stacking extended sources should yield similar results as stacking point sources. It should be noted however, in the low SNR case, the line width is underestimated, leading to an even worse estimate of the integrated flux. This issue arises in extended sources because the outer pixels of the source have a lower line amplitude (by construction), leading to a worse reconstruction of the associated spectra. In addition, the region from which the spectra are extracted can be more easily underestimated in the low SNR case. Consequently, some of the extended emission will not be successfully retrieved.

#### 5.1.9 Extended sources with an offset – set6b

Set 6b is build similarly to set6a, but stacking positions are off by a random factor, drawn uniformly between 0 and 1 arcsec. Similarly to set 6a the stacked sources are fitted with a two-dimensional Gaussian and the stacked spectra are extracted from a circular region of radius one FWHM, centred on the stacked source. As expected, amplitude reconstruction, as well as integrated flux, are not as good as in set 6a, and this effect is even more pronounced in the low SNR configuration. The amplitude, width, and flux being recovered at  $92.04 \pm 1.63$  per cent,  $92.85 \pm 0.28$  per cent, and  $93.56 \pm 2.46$  per cent in the high SNR case and  $81.0 \pm 10.4$  per cent,  $83.05 \pm 8.78$  per cent, and  $65.09 \pm 7.08$  per cent in the low SNR case. If the stacking positions were off by a too important factor then the line reconstruction would eventually be impossible. It should however be noted that, while the reconstruction fraction is inversely proportional to the uncertainties on the stacking positions, it is

also proportional to the source size. Hence, the effect coming from position uncertainty will be mitigated by the extent of the sources.

#### 5.1.10 Random central frequency, larger redshift range – set7

This set is similar to set2b but lines are simulated over a larger redshift range. Bins are set to a fixed frequency size, and rebinned to take into account the line width change due to redshift. Amplitude, linewidth, and integrated flux are properly reconstructed in both the high and low SNR simulations:  $95.9 \pm 4.8$  percent,  $96.1 \pm 0.42$  percent, and  $99.5 \pm 4.93$  percent for the amplitude, width, and integrated flux, respectively, in the high SNR case and  $97.8 \pm 10.26$  percent,  $97.3 \pm 8.08$  percent, and  $99.1 \pm 11.57$  percent, respectively, in the low SNR one. The reconstruction rate for the amplitude and flux is slightly better than in set 2b due to the overall smaller channel width (see Table 1). The standard deviation is, however, higher due to the channel width being fixed in frequency, and hence varying in velocity space. The near perfect reconstruction shows that we properly correct the effect coming from large redshift range.

#### 5.1.11 VLA type simulation – set8a

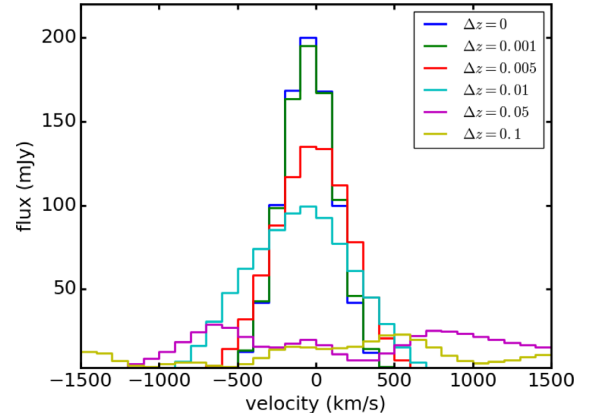
In this data set we simulate VLA observations. The reconstruction fractions are  $91.1 \pm 4.36$  percent,  $93.8 \pm 7.75$  percent, and  $99.5 \pm 8.78$  percent for the amplitude, width, and integrated flux, respectively, in the high SNR case, and  $86.4 \pm 17.16$  percent,  $96.9 \pm 14.29$  percent, and  $87.6 \pm 12.12$  percent, respectively, in the low SNR one. While parameters are well retrieved in the high SNR case, a lower reconstruction fraction is observed in the low SNR case compared to ALMA simulations. This is due to a worse sensitivity of our VLA simulations, as mentioned in Section 4.3.11.

#### 5.1.12 VLA type simulation, larger redshift range – set8b

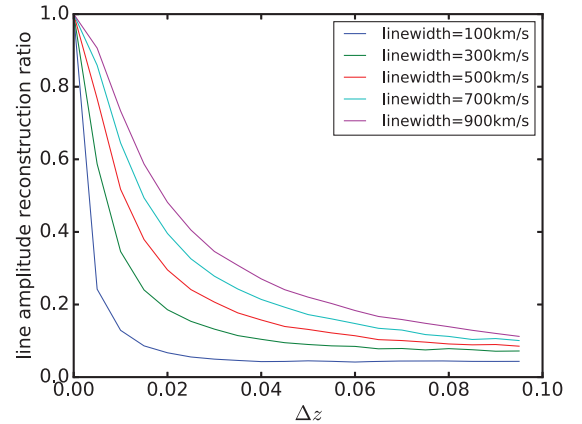
Here, we are looking for potential effects that could come from the large redshift range such as size difference in the primary beams as well as side effects from our resampling method, needed to stack sources with large redshift difference and not binned in velocity. Similarly to set7, a larger redshift range does not seem to have any substantial effect on the stacked data. While the reconstruction fractions are slightly lower than the ones in set8a, this is simply due to the velocity resolution which is  $\sim 20$  per cent higher (see Table 1). Reconstruction fractions are  $90.0 \pm 4.75$  percent,  $93.5 \pm 8.16$  percent, and  $98.7 \pm 9.92$  percent for the amplitude, width, and integrated in the high SNR case and  $82.7 \pm 12.27$  percent,  $92.8 \pm 18.27$  percent, and  $88.8 \pm 13.66$  percent in the low SNR one.

#### 5.1.13 Spectra extracted from cube – set9

This data set is similar to set 1a but the spectra are individually extracted from each cube and stacked using the 1D module of LINESTACKER. The reconstruction fractions are very similar to the one in set1a:  $94.0 \pm 0.21$  percent,  $92.7 \pm 0.23$  percent, and  $99.3 \pm 0.16$  percent for the amplitude, width, and integrated in the high SNR case and  $95.0 \pm 4.2$  percent,  $97.4 \pm 4.5$  percent, and  $96.7 \pm 2.83$  percent in the low SNR one. Which shows a good agreement between our two stacking methods, and justifies our usage of 1D data sets as an easier way to test specific spectral effects in stacking.



**Figure 4.** Stack spectrum of 30 sources of width  $400 \text{ km s}^{-1}$  for different redshift uncertainties.  $\Delta z = 0.01$  already shows a  $\sim 50$  per cent reconstruction, and rapidly dropping. We chose here a high SNR configuration (SNR before stacking  $\sim 200$ ), to showcase the pure effect of redshift uncertainties on the stack. The corresponding velocity shifts are  $\sim 0, 36, 180, 360, 1800$ , and  $3600 \text{ km s}^{-1}$ .



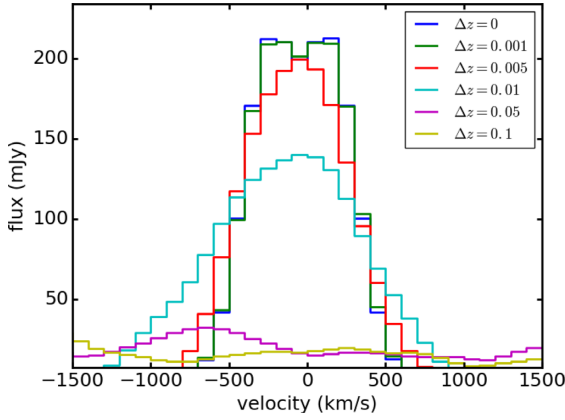
**Figure 5.** Amplitude reconstruction ratio, for a stack of 30 noise-free sources, as a function of redshift uncertainties for different linewidth (all stacked sources are simulated with the same linewidth). Stacking very narrow lines (i.e.  $\sim 100 \text{ km s}^{-1}$ ) requires a very precise redshift. Results are averaged from 1000 realizations.

## 5.2 Stacking results from 1D simulations

In the coming sections, we will be analysing results from stacking the spectra data sets presented in Section 4.4. We studied the reconstruction of the amplitude of the line, the linewidth and the integrated flux for all five data sets as well as some additional parameters specific to each set.

### 5.2.1 Diagnostic of redshift uncertainties – set10

In this set, we simulated lines with offset redshifts in order to study and estimate the effect of redshift uncertainties on the stacked line. It is important to note that the results of such a study depend on the average linewidth. Fig. 4 shows the average reconstructed line for different  $\Delta z$  at a given linewidth of  $400 \text{ km s}^{-1}$ , it shows that as soon as the redshift uncertainty becomes larger than 0.01 the stacked line cannot be recovered. Showcasing the importance of redshift accuracy. Fig. 5 shows the clear relation between linewidth and goodness of the reconstruction, when confronted to redshift uncertainties: if stacking high-velocity lines, the effect of redshift



**Figure 6.** Stack spectrum of 30 sources with double peak profile for different redshift uncertainties (both peaks are Gaussian, with a width of  $400 \text{ km s}^{-1}$ , and a distance of  $400 \text{ km s}^{-1}$  between the two peaks). An uncertainty higher than 0.001 already dilutes the two peaks, showing that similar profiles will be extremely hard to reconstruct through stacking. Similarly to Fig. 4 we chose a high SNR ( $\sim 200$  before stacking) to showcase the pure effect of redshift uncertainties. Corresponding velocity shifts are  $\sim 0, 36, 180, 360, 1800$ , and  $3600 \text{ km s}^{-1}$ .

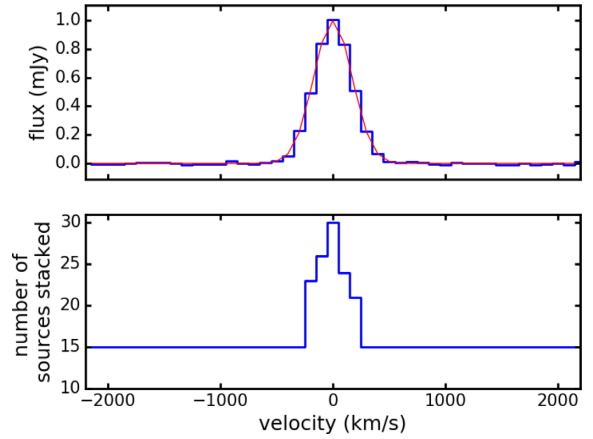
uncertainty on the reconstruction will be lowered. Alternatively reducing the spectral resolution would also mitigate the effect of redshift uncertainty on the reconstructed flux, at the expense of accuracy on the line profile measurement. Results values are summarized in Table 4.

### 5.2.2 Double peaked spectrum – set11a

This set studies a different spectral signature, of a double peak spectrum. The double peaked line is described by two Gaussians separated by a velocity difference,  $D$ , with  $200 \text{ km s}^{-1} < D < 600 \text{ km s}^{-1}$ . Both lines have the same width which is drawn at random between 100 and  $700 \text{ km s}^{-1}$ . Each stack consists of 30 spectra and is repeated for 1000 realizations. As shown in Table 5 stacking allows for a good reconstruction of all three parameters: 91 per cent, 97 per cent, and 94 per cent for the line amplitude, distance between the peaks and linewidth in the high SNR case and low SNR cases. The integrated flux is also well reconstructed at  $\sim 95$  per cent. The standard deviation of the studied parameters is low ( $\sim 5$  per cent) hence one can expect a good degree of confidence when studying similar cases with no redshift uncertainties.

### 5.2.3 Double peaked spectrum with a $\Delta z$ – set11b

Data set 11b is similar to set 11a but focuses on the effect of redshift uncertainties on the stack. Linewidth and distance between the peaks are kept constant, both at  $400 \text{ km s}^{-1}$ . Results presented are the average and standard deviation of 1000 realizations. Fig. 6 shows that, as soon as redshift uncertainty becomes worse than 0.001 the double peak feature is no longer distinguishable. This implies that such spectral signatures will be very hard to observe using stacking, and will require a very good redshift accuracy. Furthermore, the reconstruction of such a feature will also depend on the distance between the peaks. If the separation between the peaks is higher they will be easier to discern. Table 5 shows the best fit of the stack, using a double peak fit. While the reconstruction of all three parameters is near perfect at  $\Delta z = 0$ , from  $\Delta z \sim 0.05$  the line cannot be recovered (see Fig. 6). One should note that the integrated flux is not impacted



**Figure 7.** Average results from 1000 realizations of stacking 30 sources from set13 (in the low SNR configuration). Top: Resulting average stack and the corresponding Gaussian fit (in red). Bottom: Number of sources stacked at each velocity bin. Due to the nature of the data stacked the number of sources quickly drops to half of the sources when moving away from the central channels.

as much by the redshift uncertainties as the other parameters, thus in such cases it is advised to focus on the integrated flux rather than other line parameters.

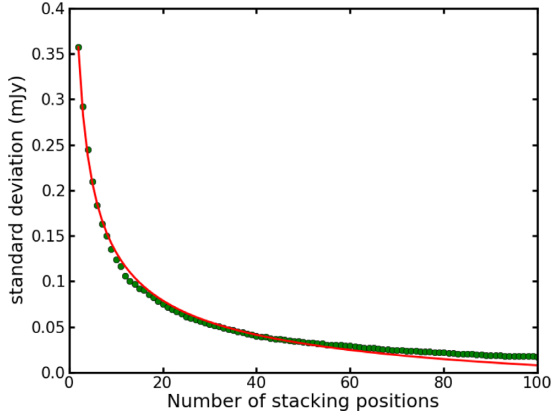
### 5.2.4 Outflows – set12

Data set12 was built to study cases with spectral signature of outflows. Spectra consist of two components, one main line with an amplitude of  $2000 \text{ mJy}$  and a width of  $400 \text{ km s}^{-1}$ , and an outflow component with an amplitude of  $200 \text{ mJy}$  and a width of  $1000 \text{ km s}^{-1}$  in the high SNR case. Amplitudes are set to 10 and  $1 \text{ mJy}$  for the main and outflow component, respectively, in the low SNR case (linewidth are the same as in the high SNR case). Results are obtained through fitting the stacks with two Gaussian components (results are averaged from the 1000 realizations). In the low SNR sets, the amplitude of the main component is fixed when fitting. This is done to avoid cases where good fitting is achieved with a brighter broad component and a fainter main line, leading to a much higher uncertainty on the broad component amplitude reconstruction (of the order of  $\sim 80$  per cent).

Once again LINSTACKER allows for a good reconstruction, with about  $\sim 99$  per cent retrieval of the parameters for both components in both SNR configuration (assuming the fitting method described above in the low SNR configuration). It should be noted however, that, when stacking lines of different linewidths, the variation in linewidths needs to be accounted for, the method would otherwise be biased in finding outflows. To do so a spectral rebinning method can be applied to the data pre-stacking (see Section 3.4 and Stanley et al. 2019).

### 5.2.5 Lines on the edge – set13

Data set13 was built to diagnose cases where all lines would lie on the edge of the spectral window. Lines have the same properties as data set10 when  $\Delta z = 0$ . Fig. 7 shows an example from such a stack product. One can see that the numbers of sources stacked rapidly drops on both sides of the central channel. However, reconstruction rates stay close to perfect for both noise configurations for all parameters, see Table 4. It should none the less be noted that, while average results are consistent with data set10 at  $\Delta z = 0$ , the standard



**Figure 8.** Standard deviation in the stack of empty cubes (simulated through SIMALMA) as a function of number of cubes stacked. Red line indicates the theoretical  $1/\sqrt{N}$  drop off. Presented results are averaged from 100 realizations.

deviation is significantly higher than set10 in the low SNR regime (of the order of  $\sim 3$  times higher for both the amplitude and the integrated flux), indicating a larger spread in the expected results.

## 6 DISCUSSION

### 6.1 Choice of parameters

In the following subsections, we justify the choice of parameters for our simulated data sets.

#### 6.1.1 Array configuration

Even if a specific array configuration has been chosen for most simulations (ALMA cycle 6 configuration 2), the method has been extensively tested in other array configurations (not presented here), as well as with other choice of interferometer – as showed through data sets 8a and 8b, simulating VLA observations – and neither should have any impact on the performance of the algorithm.

#### 6.1.2 Number of sources

For every simulation, we chose to stack 30 sources. The number 30 has been chosen as an intermediate number, allowing for both a good noise reduction ( $\sqrt{30} \sim 5.5$ ) and a satisfying computing time. This number is typical for small, but statistically significant, samples. A lower number would imply too important statistical fluctuations and, on the other hand, a higher number of sources (samples as big as a few thousands e.g. Dole et al. 2006, can be expected) would not change the conclusion from our analysis, simply allowing identification of fainter signal.

To verify the good behaviour of noise reduction as a function of number of sources stacked, we stacked an increasing number of empty fields (i.e. without sources, containing just noise). The fields have similar noise properties as in the rest of our simulations. Computing for each new number of stack positions the standard deviation across the map, see Fig. 8. The noise reduction is in good accordance with theoretical predictions (showing an 80 per cent noise reduction for 30 sources, similar to the 82 per cent theoretically expected) confirming the relevance of the number of sources in our stacks.

#### 6.1.3 Number of simulations

Each data set has been simulated, stacked, and analysed 100 times in the case of 3D data sets, and 1000 in the case of 1D data sets. The number of realizations has been chosen for practical reasons, limiting computation time, it should however be noted that the total number of resulting sources is high enough to provide statistically relevant numbers. This can be seen from the uncertainties on the reconstructed parameters – coming directly from the spread of results in the stacking iterations – that are small compared the found values.

#### 6.1.4 High and low SNR sets

Every 3D simulated data set was run both in a high and a low SNR configuration. While high SNR data sets reveal the ideal behaviour of our algorithm and show that it performs well, stacking will be mostly used on low SNR data. It was therefore essential to show that our tool was successful in such cases as well. Flux values have been chosen arbitrarily, mimicking typical values. The amplitude of the line in low SNR sets has been chosen such that individual lines would be of order of the noise, thus impossible to detect individually but strong enough to be studied once 30 of such sources are stacked.

#### 6.1.5 Source size

Most simulated data sets have been performed simulating point sources. It is not uncommon that high-redshift galaxies are seen as point sources because of the high resolution needed to resolve such objects, and working with point sources allows to have a more controlled data set. We have however studied extended sources as well, in data sets 6a and 6b, showing that no substantial effect would come from stacking extended sources instead of point sources. However, only homogeneous cases have been tested (where sources were either all point like or all extended). Furthermore, when studying extended sources all sources had similar sizes. It may happen that stacked sources have very different sizes, resulting in an inhomogeneous spatial distribution in the stack product. In this case, one should be prudent when extracting quantitative conclusions from such an analysis (such as sizes, density, spatial distribution, etc.). Using statistical methods like bootstrapping or subsampling should prove useful to diagnose such cases. Furthermore, if the sources extent is identifiable pre-stacking, extracting and stacking directly the spectra – using the 1D stacking module of LINESTACKER – could allow to get past these biases (see Stanley et al. 2019, for an example use of combination of 1D and 3D stacking).

### 6.2 Impact of redshift uncertainty on stack

When stacking spectral line data, the two most important properties to know are the position and redshift of the source. Uncertainties in the astrometry will result in the stacked source appearing more extended than in reality (see e.g. Lindroos et al. 2016, 2018, for further discussion on astrometric uncertainties). In Sections 5.2.1 and 5.2.3, we have shown that uncertainties on the redshift have a significant impact on the stacked line result as well as on the potential reconstruction of the average line profile. The three most common cases of high-redshift galaxies, where the uncertainties could be too large are (i) when using photometric redshifts for a large number of galaxies, (ii) if relying on the ultraviolet, broad emission lines of quasars, as these can often be shifted relative to the velocity of the quasar host (e.g. Coatman et al. 2017), and (iii) when studying Lyman

$\alpha$  emitters because their peak emission is known to be offset from the systemic redshift (e.g. Shapley et al. 2003; Rakic et al. 2011).

An uncertainty of  $\Delta z = 0.05$  results in the amplitude of a line of  $400 \text{ km s}^{-1}$ , with a spectral accuracy of  $100 \text{ km s}^{-1}$ , to be recovered at only  $\sim 20$  per cent. However, wider lines will show a better reconstruction (see Fig. 5) at similar redshift uncertainties. Bellagamba et al. (2012) have shown that one can typically expect  $\langle \frac{\sigma_z}{1+z} \rangle < 0.05$  from photometric redshift of weak lensing surveys, which at  $z \sim 6$  corresponds to  $\Delta z < 0.35$ . Combined with our analysis this shows that using photometric redshift to stack lines from high-redshift objects should not be viable.

Other examples of significant offsets can come from quasars and powerful AGNs, where the redshift is determined by lines that are several thousands  $\text{km s}^{-1}$  wide and can be offset due to outflows. This can result in shifts that are so large that the line might be shifted out of the interferometric bands. An example of this is W0410–0913 shown in Fan et al. (2018) where even the optical redshift was a few thousands  $\text{km s}^{-1}$  offset.

### 6.3 Gaussian fitting to recover stacked line parameters

When stacking lines with a wide range of line widths the resulting stacked line ends up not being Gaussian. Therefore, using Gaussian fitting to recover the line parameters will yield biased results. This effect is especially relevant if one is interested in retrieving the line shape. It should be noted however that this effect can be mitigated depending on the spread in line widths. For example, in simple cases such as the one given in Fig. 3, fitting a single Gaussian line profile is a good approximation for the used data sets. However, in more complicated or realistic cases, as for example the case given in Appendix A2 and Fig. A2, one can see that deriving parameters using single Gaussian fitting would yield extremely biased results. In addition, it is important to keep in mind that, even if not interested in line shape, using Gaussian fitting to retrieve the stacked line amplitude will lead to its systematic underestimation (see Fig. 3).

### 6.4 Stacking in the $uv$ -plane

In this paper, we focus on stacking in image-plane. Lindroos et al. (2015), conducted a systematic comparison between stacking interferometric data in the image-plane and in the  $uv$ -plane. The two approaches yield similar results, but the  $uv$ -stacked results are more robust for a range of cases; in some cases the difference was a few per cent, while in other cases the improvement was up to 15 per cent. The cases where  $uv$ -stacking has the potential to make a difference are (i) when the  $uv$ -coverage differs between data sets: this can affect the stacked result. For example, if stacking in the image-plane the data should be imaged to the same resolution, however, this might not always be easy, depending on the array configuration. Furthermore, if the stacked sources are faint, i.e. below the few sigma rms of the data, the sources will not have been cleaned in the imaging and therefore the stacked image will include a stacked version of the dirty beams. The potential complications of these effects are avoided when stacking directly in the  $uv$ -plane as the imaging is done afterwards. (ii) When bright sources are present in the data, these have to be removed from the data before stacking, both for  $uv$ - and image-plane. This is normally done using the CLEAN algorithm, and the residuals from this can leave noise and cleaning artefacts in the data. Often this affects only the short baselines, and can hence be more easily identified and dealt with in  $uv$ -stacked data than in the image-plane stacked result. (iii) When stacking extended source, it is possible to analyse the stacked  $uv$ -data for possible cleaning artefacts

or problematic baselines that could affect size measurements, and it is possible to estimate the extend of the emission on the visibilities.

In comparison between stacking in the  $uv$ -plane and image-plane, the image-plane stacking is computationally faster, in particular if also including statistical tests using Monte Carlo/resampling methods. In addition, image stacking allows for easier masking and is generally more intuitive to work with. It has been shown in Lindroos et al. (2015) to yield satisfying results when the imaging can be done reliably. Finally, future interferometers such as the SKA, will likely not archive all visibility data after processing and imaging because of the enormous amounts of data produced.

In this paper, we have focused on studying the stacking of spectral line data, and this has a number of complications that are independent of whether the stacking is done in the image- or the  $uv$ -plane. Spectral line stacking can be carried out in both image- and  $uv$ -plane, and the differences in performance is expected to be the same as was found for continuum stacking in Lindroos et al. (2015).

Finally, we note that if the line width of the stacked line is known a priori, an easy approach allowing a coarse version of line stacking in the  $uv$ -plane could be to treat the line channels as a continuum channel (see e.g. Fujimoto et al. 2019; Méndez-Hernández et al. 2020). Through this approach, it is the line intensity that is stacked, while all potential information about line profile is not included.

### 6.5 LINESTACKER as a spectral analysis tool

While we used LINESTACKER solely to study emission lines from high-redshift galaxies, the tool is flexible and would perform equally well on other type of data in the range GHz to THz. It is probable that the tool performs equally well on lower frequencies, but has not been tested. Additionally, while all the tests focused on stacking interferometric data, it is possible to use it to stack non-interferometric data (this could possibly include optical integral field spectroscopic data, though we note that performance has not been tested). The tool could for example be used on stars, clusters, gas clouds, local galaxies, galaxy clusters or any other object. In principle any line data can be studied, and it would also be possible to stack different lines from the same object together (from different transitions for example), provided that the sub-cubes spectral size is chosen as compact enough to avoid overlap between the different lines. Besides, stacking absorption lines is theoretically identical and should yield similar results. It would also be possible in principle, using Monte Carlo methods, to find better individual redshifts of the studied objects, by trying to optimize stacked line reconstruction. Indeed, stacked line amplitude should be maximized when all the target sources are stacked in phase, i.e. when all lines are stacked perfectly all in the velocity centre. One could then look for the set of redshifts maximizing the stacked line amplitude, recovering, in fine, better individual redshifts while also optimizing line reconstruction.

### 6.6 Spectroscopic stacking in the literature

Spectroscopic stacking as a method has been used more and more in the past decade. When presented in the literature authors typically have not shown tests of their algorithm nor made it publicly available. Among other, stacking has been used to study H I, both in emission and absorption (Murray et al. 2014, 2018). H I line profiles are usually complex, and, while the redshift precision is usually good, such complexity will typically be diluted through stacking. Aside from H I, other molecules have been studied, such as CO (e.g. Decarli et al. 2016, in multiple transitions) or [C II] (e.g. Decarli et al. 2018; Bischetti et al. 2019; Fujimoto et al. 2019; Stanley

et al. 2019). In Decarli et al. (2016), stacking was performed for a sample of high- $z$  galaxies, where the optical redshifts were uncertain by 200–1000 km s<sup>-1</sup>, yielding a low-significance detection. This highlights the challenge of stacking spectroscopic data without accurate redshifts or velocities. Bischetti et al. (2019) studied line profiles through stacking without rebinning their data. None the less, they build physically selected sub samples (depending for example on the width of the lines) to reduce the effect of stacking sources with different linewidth. In Stanley et al. (2019), we used LINESTACKER to search for outflows in high-redshift quasars. In addition to the main algorithm, we used some of tools included in LINESTACKER to improve our analysis: spectral rebinning to properly recover line profiles, as well as subsampling to identify the sources exhibiting the best outflow signature. Our analysis demonstrated the efficiency and flexibility of LINESTACKER to study faint emission at high redshift.

Murray et al. (2014) used a different technique for edge treatment than the one presented in Section 2.2. In their algorithm they simply added zeros to fill spectral channels, when sources' lines were too close to the edge of the observation window. This has a bigger, unwanted, impact on the stacked result than the method used in LINESTACKER, because zeros added to the stack will, once averaged with the rest of the sources, artificially drive the edges of the stacked spectral window to values closer to zero.

Using LINESTACKER would allow a fast, uniform and controlled way to do spectral stacking. It would also allow the use of statistical tools as well as data handling treatment in a more systematic and coherent way.

## 7 SUMMARY

We have carried out an extensive analysis of stacking of interferometric spectral line data using our new algorithm and tool LINESTACKER. We have used simulations of near-ideal and realistic cases of simulated data from two different interferometers, ALMA and VLA. All high SNR simulations emphasized the controlled behaviour of our algorithm while low SNR simulations focused on noise reduction and proved the efficiency and usefulness of stacking.

We showed and justified the need for statistical tools, both pre-stacking (e.g. rebinning) and post-stacking (e.g. bootstrapping), to better understand stacking results as well as the distribution of parameters of the stacked population.

We find that knowing the redshift of the sources with a good precision is a necessary condition for a good line reconstruction. And that, for an average linewidth of 400 km s<sup>-1</sup>, a redshift uncertainty below 0.01 implies a line reconstruction around 60 per cent, dropping to roughly 10 per cent at  $\Delta z = 0.1$ . Furthermore, it has been shown that more complex spectral signatures will tend to be smeared out by uncertainties on the redshift, and that hence, in most cases, more complex spectral signatures will disappear when stacked (either through averaging with other, different, spectral shapes, or because of the homogenization due to redshift error).

In addition, we showed that stacking Gaussian lines with different linewidth results in a non-Gaussian shaped stack, leading to a possible misinterpretation of the fitted result. This can be fixed by rebinning the spectra, if individual linewidths are identifiable before stacking. Such spectral configuration are especially problematic when trying to identify line profiles.

With the significantly improved capabilities of modern radio and mm interferometers in combination with deep optical and near-infrared surveys, the use of stacking is becoming increasingly relevant. As seen in the literature, there is a growing interest to exploring the faint, often individually undetected sources through

stacking. However, the number of public tools available are still limited. LINESTACKER provides the community with increased opportunity for optimal synergy between modern telescopes and the large astronomical surveys. Enabling multiwavelength studies is necessary also for faint sources in order to establish a complete understanding of the chosen population.

LINESTACKER is open source and open access. It can be downloaded at <https://www.oso.nordic-alma.se/software-tools.php>. The tool is provided with examples and documentation.

## ACKNOWLEDGEMENTS

We thank the anonymous referee for their instructive comments for the improvement of this paper. We thank John Conway for extensive discussion on the project. We thank the staff of the Nordic ALMA Regional Center node for their support. KKK acknowledges support from the Swedish Research Council (2015-05580). JJB thanks Martin Zwaan for helpful discussions.

## 8 DATA AVAILABILITY

No new data were generated or analysed in support of this research.

## REFERENCES

- Arnouts S. et al., 2001, *A&A*, 379, 740  
 Bellagamba F., Meneghetti M., Moscardini L., Bolzonella M., 2012, *MNRAS*, 422, 553  
 Bischetti M., Maiolino S., Carniani S., Fiore F., Piconcelli E., Fluetsch A., 2019, *A&A*, 630, A59  
 Bothwell M. S. et al., 2013, *MNRAS*, 429, 3047  
 Cady F. M., Bates R. H. T., 1980, *Opt. Lett.*, 5, 438  
 Chen C.-T. J. et al., 2013, *ApJ*, 773, 3  
 Coatman L., Hewett P. C., Banerji M., Richards G. T., Hennawi J. F., Prochaska J. X., 2017, *MNRAS*, 465, 2120  
 Decarli R. et al., 2014, *ApJ*, 780, 115  
 Decarli R. et al., 2016, *ApJ*, 833, 69  
 Decarli R. et al., 2018, *ApJ*, 854, 97  
 Dole H. et al., 2006, *A&A*, 451, 417  
 Fan L., Knudsen K. K., Fogasy J., Drouart G., 2018, *ApJ*, 856, L5  
 Fruchter A. S., Hook R. N., 2002, *PASP*, 114, 144  
 Fujimoto S. et al., 2019, *ApJ*, 887, 107  
 Gullberg B. et al., 2015, *MNRAS*, 449, 2883  
 Gullberg B. et al., 2018, *ApJ*, 859, 12  
 Hickox R. C. et al., 2007, *ApJ*, 671, 1365  
 Hickox R. C. et al., 2009, *ApJ*, 696, 891  
 Högbom J. A., 1974, *A&AS*, 15, 417  
 Ikarashi S. et al., 2015, *ApJ*, 810, 133  
 Karim A. et al., 2011, *ApJ*, 730, 61  
 Knudsen K. K. et al., 2005, *ApJ*, 632, L9  
 Lindroos L., Knudsen K. K., Stanley F., Muxlow T. W. B., Beswick R. J., Conway J., Radcliffe J. F., Wrigley N., 2018, *MNRAS*, 476, 3544  
 Lindroos L., Knudsen K. K., Vlemmings W., Conway J., Martí-Vidal I., 2015, *MNRAS*, 446, 3502  
 Lindroos L. et al., 2016, *MNRAS*, 462, 1192  
 Murray C. E., Stanimirović S., Goss W. M., Heiles C., Dickey J. M., Babler B., Kim C.-G., 2018, *ApJS*, 238, 14  
 Murray C. E. et al., 2014, *ApJ*, 781, L41  
 Méndez-Hernández H., et al., 2020, *MNRAS*, 497, 2771  
 Pannella M. et al., 2009, *ApJ*, 698, L116  
 Rakic O., Schaye J., Steidel C. C., Rudie G. C., 2011, *MNRAS*, 414, 3265  
 Rybak M. et al., 2019, *ApJ*, 876, 112  
 Shapley A. E., Steidel C. C., Pettini M., Adelberger K. L., 2003, *ApJ*, 588, 65  
 Stanley F., Jolly J. B., König S., Knudsen K. K., 2019, *A&A*, 631, A78

Stanley F. et al., 2017, *MNRAS*, 472, 2221

Thompson A. R., Moran J. M., Swenson George W. J., 2001, *Interferometry and Synthesis in Radio Astronomy*, 2nd edn., Springer, Berlin

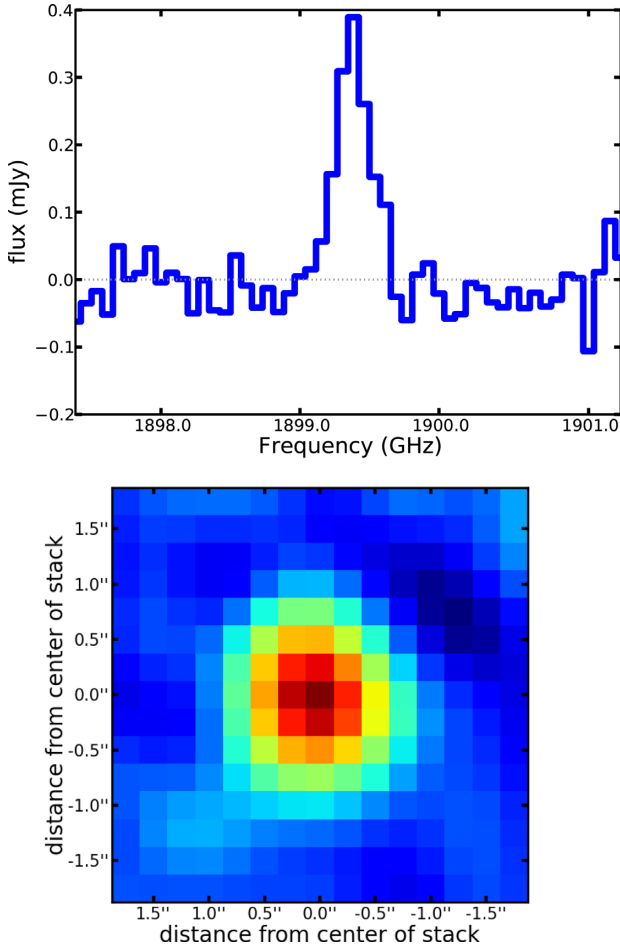
## APPENDIX A: EXAMPLE USE OF LINESTACKER

All the following examples are presented in more detail in the LINESTACKER library, including the lines of code, and the associated data sets, required to run them.

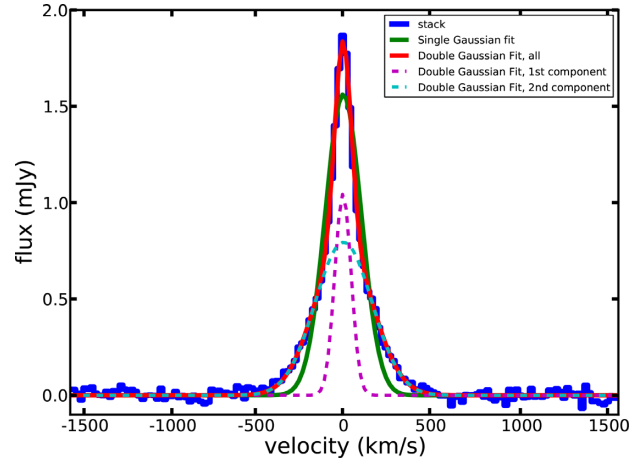
### A1 Cube basic example

The first example shows a basic use LINESTACKER. In many cases stacking is done for sources where the target lines are not individually detected and where the redshift or systemic velocity has been determined through other means (e.g. in another wavelength range). For example, this could be measuring the stacked CO line of optically detected galaxies, where the redshift is determined from optical spectroscopy.

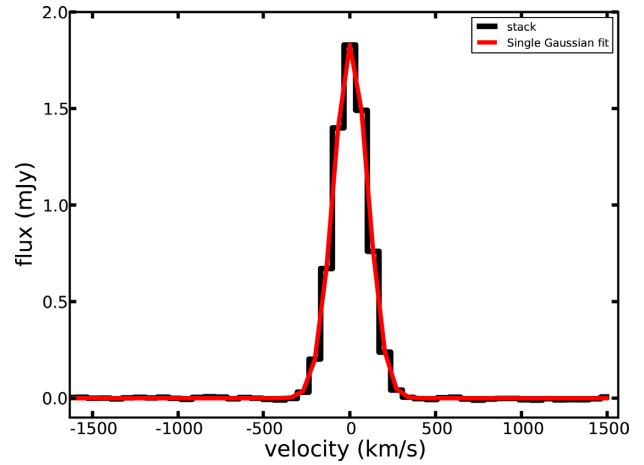
The input required together with the data cubes are source coordinates and line centre, which should be given in as a file. The coordinates are given in J2000 RA and Dec., while the line centre is tabulated with the source redshift and rest-frame frequency of the line. We note that several methods for determination of the line centre are available, for example central bin index and central



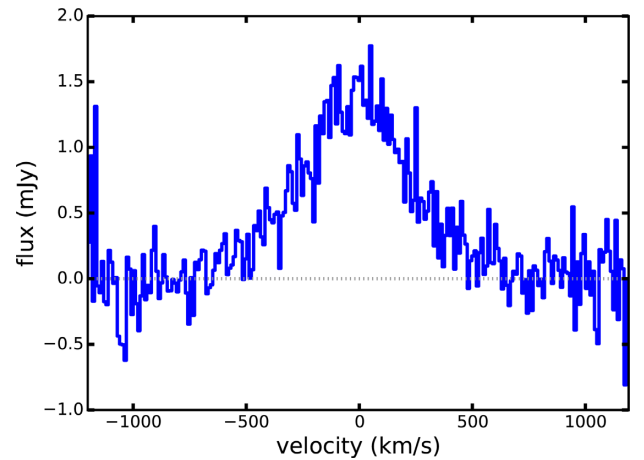
**Figure A1.** Results from the example of a basic stacking. Top: Spectrum extracted from central pixel of the stack. Note that the frequency values are arbitrary. Bottom: Moment-0 map of the mean stack.



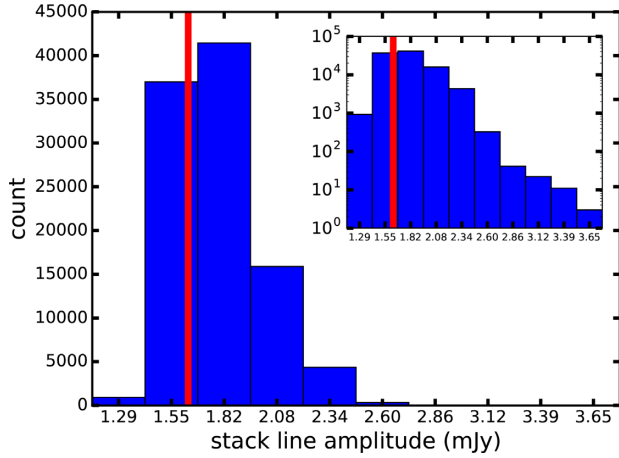
**Figure A2.** Original stack, before rebinning. Fitted with one (green line) and two (red line) Gaussians.



**Figure A3.** Stack result after rebinning. Shows very good agreement with single component Gaussian fitting.



**Figure A4.** Median stack of the 50 spectra with non-homogeneous line amplitude.



**Figure A5.** Distribution of the results from the median bootstrap analysis. The straight red line indicates the result from the stack of the entire sample.

velocity – full description of the line centre identification methods are available in the LINESTACKER documentation. Here, we show an example with 30 cubes, each containing a point source in its centre. The sources spectra consist solely of an emission line and some noise. The amplitude of the lines has, on average, the same amplitude as the noise (similarly to set1a in the low SNR regime). Mean stacking is performed on the cubes, and the result can be seen in Fig. A1.

### A2 Cube extended example

Most high- $z$  galaxies will display varying line profiles and line widths depending on properties such as mass and orientation. Here follows an example expanded from the previous subsection, where we use simulated data of sources with random line widths. As discussed in the main body of the paper, stacking sources of varying line widths can affect the final result. The present example details the usage of the rebinning method coupled to stacking. We note that using rebinning requires some prior realistic knowledge of the line width, this could for example be stacking of fainter isotopologue lines under the assumption that the line width is similar to individually detected main isotopologue lines (e.g.  $^{12}\text{CO}$  compared to  $^{13}\text{CO}$ ; e.g. Méndez-Hernández et al. 2020) or the search for high-velocity outflows (e.g. Decarli et al. 2018; Stanley et al. 2019).

The data used here consists of 50 cubes, with one point source each in their centre. The sources spectra consist of a bright ( $\text{SNR} \sim 100$ ) Gaussian line with random width. The randomization of the width is operated through Gaussian random centred around  $200 \text{ km s}^{-1}$ , with a minimum value of  $50 \text{ km s}^{-1}$ .

All cubes are stacked using mean stacking. The resulting spectrum can be seen in Fig. A2. Here, the spectrum is extracted from the stacked cube by summing all  $8 \times 8$  pixels of the stacked cube. One can see that the resulting stack is better fitted by two Gaussian line profiles, even if the lines were originally simulated with single Gaussian profiles.

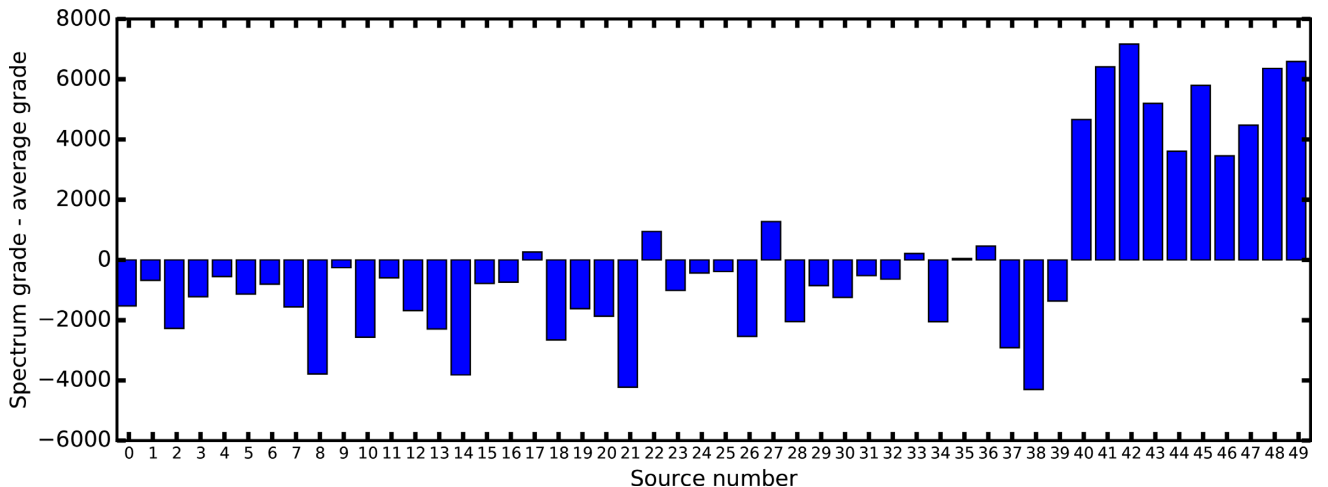
To avoid this effect *spectral rebinning* is performed on the cubes, using the method embedded in LINESTACKER. The method allows for automated fitting to estimate the line widths, which are used for the rebinning. The rebinned cubes are stacked again, using mean stacking. The stacking result is presented in Fig. A3, and one can see that, after rebinning, the resulting stack spectrum is purely Gaussian, as expected. The width of the stacked line is  $\sim 242 \text{ km s}^{-1}$  (to be compared to the average line width before stacking:  $\sim 231 \text{ km s}^{-1}$ ).

### A3 1D extended example

Because averaged properties of the studied population are retrieved from stacking, the underlying assumption in stacking data is that all sources have similar properties. In some cases, it is known that not all sources will show similar line properties, but it is not known a priori which sources. This can be well illustrated when searching for outflows, where orientation could affect the projected kinematics properties and thereby the line width (e.g. Stanley et al. 2019).

In this example, we present the use of bootstrapping and subsampling. Here, we will be using the 1D module of LINESTACKER, and hence stack spectra directly, and not whole cubes. The data set consists of 50 spectra. The spectra are composed of a Gaussian line, randomly centred, and some noise. The amplitude of the Gaussian line is typically of order of the noise, however, 10 of the sources have an average line amplitude 10 times higher. Such an inhomogeneous distribution has been chosen to showcase the performance and usage of both bootstrapping and subsampling. All spectra are stacked using median stacking, the stacking result is shown in Fig. A4.

The second step is to show the usage of bootstrapping methods. Here, we perform bootstrapping using median stacking, and iterating 100 000 times. Results can be seen in Fig. A5.



**Figure A6.** Results from subsampling analysis: the average grade is subtracted from each sources grade.

The bootstrapping results show a clear skewed distribution towards results of higher amplitude. This typically implies an inhomogeneous distribution of the sources amplitude. While a similar conclusion could have been drawn from looking at bootstrapping paired with mean stacking, it would not have been as easy to display, which is why we chose to show median stacking in this example. When using bootstrapping paired with mean stacking, the in-homogeneity of the results can be typically deduced from the width of the bootstrap distribution – showing a much larger spread of the results than it would have if the sample was homogeneous.

Since bootstrapping allowed to suspect the presence of outliers in the data, one can now use subsampling to try to identify them.

Because the bootstrapping analysis indicated a skewed distribution of the lines amplitude, the amplitude will be used as a criteria to separate the subsamples (see Section 3.3 for a description of the usage of grades in our subsampling method). The subsampling method is performed 100 000 times and its results can be seen in Fig. A6.

The distribution of source grade shows a clear in-homogeneity, and it is easy here to identify the last 10 sources as having higher amplitude.

This paper has been typeset from a  $\text{\TeX/L\AA\TeX}$  file prepared by the author.