

# Archiving Visibility Data Using Lossy Baseline-Dependent SVD Techniques

---

Mukundi Ramanyimi

*Supervisor(s):*

Dr. Sphehile Makhathini

Prof. Marcellin Atemkeng (Rhodes University)



A dissertation submitted in partial fulfilment of the requirements for the degree of  
Master of Science in Physics

in the

School of Physics

University of the Witwatersrand, Johannesburg

8 June 2025

# Declaration

I, Mukundi Ramanyimi, declare that

- The research reported in this dissertation, except where otherwise indicated, is my original research.
- This dissertation has not been submitted for any degree or examination at any other university.
- This dissertation does not contain other persons' data, pictures, graphs or other information unless specifically acknowledged as being sourced from other persons.
- This dissertation does not contain other persons' writing unless specifically acknowledged as being sourced from other researchers.
- This dissertation does not contain text, graphics or tables taken from the Internet, unless specifically acknowledged, and the source is detailed in the dissertation and in the References sections.



Mukundi Ramanyimi

8 June 2025

# *Abstract*

Modern radio interferometer arrays, such as the MeerKAT [1], the Australian Square Kilometre Array Pathfinder (ASKAP) [2, 3], the Low Frequency Array (LOFAR) [4], the Murchison Widefield Array (MWA) [5, 6], and the upcoming Square Kilometre Array Observatory (SKAO) [7], generate large volumes of data due to their high temporal and spectral resolutions, large number of baseline configurations, and wide bandwidths. Managing these data volumes poses substantial challenges in terms of storage and processing. To address the growing costs, averaging techniques are widely used to reduce data sizes. However, averaging leads to signal loss in radio interferometric images, resulting in smeared or blurred source emissions and reduced source amplitudes.

Moreover, the extent of this smearing is baseline-dependent, as the signal phase depends on baseline length. Specifically, longer baselines are more affected than shorter ones. This is addressed by Baseline Dependent Averaging (BDA), which applies variable averaging intervals - longer for shorter baselines and shorter for longer baselines. BDA achieves high data volume reduction since radio interferometers generally have more shorter baselines, which can be aggressively averaged with minimal smearing effects. However, BDA changes the time-frequency grid structure of the data, making it incompatible with the standard storage format in the field, the Measurement Set (MS).

A promising approach to data compression was presented by Atemkeng et al. [8], who developed a compression technique based on Singular Value Decomposition (SVD). This approach exploits the inherent structure of raw visibility data, representing it as a low-rank matrix approximation where each component corresponds to a specific Fourier component of the sky distribution. By approximating the data with a reduced rank, the essential features of the original data can be captured using fewer components, effectively reducing data size.

In this work, we build on the methods introduced by Atemkeng et al. [8], specifically evaluating the effectiveness of SVD in compressing large volumes of data while preserving image quality and data fidelity for long-term archival. Although

our study focuses on the MeerKAT telescope, the approach can be adapted for use with any other radio telescope.

Our findings demonstrate that for a bright point source (1 Jy), whether located at the phase centre or away from it, the data features can effectively be captured using a single component, recovering over 99.90% of the source amplitude and achieving a data size reduction of over 97%. For fields with multiple sources, the features can be fully captured using 3-4 components out of 24, recovering over 99.90% of the source amplitude for a source at the edge of the Field of View (FoV), which is around 1.1 deg for the MeerKAT at a frequency of 1.4 GHz. This results in a data size reduction of over 91%.

Additionally, we found that the source or field direction does not impact SVD compression. On the other hand, the Signal to Noise Ratio (SNR) significantly affects SVD compression. For sources with low SNR or faint sources, all components are required to recover more than 97% of the source amplitude, making the compression ineffective. In such scenarios, it would be more advantageous to first denoise the data or to use BDA.



# Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Sphehile Makhathini, for his invaluable support, expert guidance, and above all, his patience. Each interaction with him has been a learning experience, and I am truly grateful for the knowledge he has shared with me. I also extend my sincere thanks to my co-supervisor, Prof. Marcellin Atemkeng, for his support and guidance. I have gained so much from his insights and look forward to continuing to learn from him.

Additionally, I would like to thank my family and friends for their unwavering support during this journey, with special thanks to my mother, who has been my greatest source of encouragement. Finally, I would like to acknowledge the National Research Foundation (NRF) and the South African Radio Astronomy Observatory (SARAO) for supporting my postgraduate studies. In particular, South African Radio Astronomy Observatory-Human Capital Development (SARAO-HCD) Group Grant that funded this research.

# Contents

|                                                                |             |
|----------------------------------------------------------------|-------------|
| <b>Declaration</b>                                             | <b>i</b>    |
| <b>Abstract</b>                                                | <b>ii</b>   |
| <b>Acknowledgements</b>                                        | <b>iv</b>   |
| <b>List of Figures</b>                                         | <b>viii</b> |
| <b>List of Tables</b>                                          | <b>xi</b>   |
| <b>1 Introduction</b>                                          | <b>1</b>    |
| 1.1 Overview of Radio Astronomy . . . . .                      | 1           |
| 1.2 Radio Interferometry Fundamentals . . . . .                | 3           |
| 1.2.1 Young's Double Slit Experiment . . . . .                 | 3           |
| 1.2.2 Visibility Function . . . . .                            | 5           |
| 1.2.3 The Radio Interferometric Measurement Equation . . . . . | 9           |
| 1.2.4 Smearing Effects . . . . .                               | 16          |
| 1.3 Compression of Visibility Data . . . . .                   | 18          |
| 1.3.1 Simple Averaging . . . . .                               | 20          |
| 1.3.2 Baseline Dependent Averaging . . . . .                   | 23          |
| 1.3.3 Singular Value Decomposition . . . . .                   | 25          |
| 1.4 Problem Statement . . . . .                                | 29          |
| 1.5 Research Aims . . . . .                                    | 29          |
| <b>2 Research Methodology</b>                                  | <b>31</b>   |
| 2.1 Simulations . . . . .                                      | 31          |
| 2.2 Data Compression . . . . .                                 | 32          |
| 2.3 Statistical Analysis of Residuals . . . . .                | 33          |
| 2.4 Imaging . . . . .                                          | 33          |

|          |                                                             |           |
|----------|-------------------------------------------------------------|-----------|
| 2.5      | Compression Factor . . . . .                                | 34        |
| <b>3</b> | <b>Results and Discussion</b>                               | <b>35</b> |
| 3.1      | Individual Singular Values . . . . .                        | 35        |
| 3.1.1    | Phase centre . . . . .                                      | 35        |
| 3.1.2    | Away from the Phase Centre . . . . .                        | 40        |
| 3.1.3    | Point Sources . . . . .                                     | 44        |
| 3.1.4    | Point and Extended Sources . . . . .                        | 47        |
| 3.1.5    | Signal to Noise Ratio (SNR) . . . . .                       | 50        |
| 3.1.6    | Point Sources with a total flux of 10 Jy . . . . .          | 53        |
| 3.1.7    | Source Direction . . . . .                                  | 54        |
| 3.2      | Cumulative Singular Values . . . . .                        | 58        |
| 3.2.1    | Phase Centre . . . . .                                      | 58        |
| 3.2.2    | Away from the Phase centre . . . . .                        | 59        |
| 3.2.3    | Point Sources . . . . .                                     | 60        |
| 3.2.4    | Point and Extended Sources . . . . .                        | 61        |
| 3.2.5    | SNR . . . . .                                               | 63        |
| 3.2.6    | Point Sources with a total flux of 10 Jy . . . . .          | 65        |
| 3.2.7    | Source Direction . . . . .                                  | 66        |
| <b>4</b> | <b>Software Package</b>                                     | <b>70</b> |
| 4.1      | Compressing and Archiving Visibility Data . . . . .         | 70        |
| 4.1.1    | Simple SVD . . . . .                                        | 70        |
| 4.1.2    | Baseline-Dependent Singular Value Decomposition (BDSVD)     | 71        |
| 4.2      | Data Decompression . . . . .                                | 72        |
| <b>5</b> | <b>Conclusions and Future Work</b>                          | <b>73</b> |
| 5.1      | Conclusions . . . . .                                       | 73        |
| 5.2      | Future Work . . . . .                                       | 74        |
| <b>6</b> | <b>Appendix A</b>                                           | <b>76</b> |
| 6.0.1    | Signal to Noise Ratio (SNR)-1000 . . . . .                  | 76        |
| 6.0.2    | Signal to Noise Ratio (SNR)-50 . . . . .                    | 79        |
| 6.0.3    | Signal to Noise Ratio (SNR)-5 . . . . .                     | 82        |
| 6.0.4    | Source Direction - Declination of $-60^\circ$ deg . . . . . | 85        |

|       |                                                   |    |
|-------|---------------------------------------------------|----|
| 6.0.5 | Source Direction - Declination of 0 deg . . . . . | 88 |
|-------|---------------------------------------------------|----|

|                     |  |           |
|---------------------|--|-----------|
| <b>Bibliography</b> |  | <b>92</b> |
|---------------------|--|-----------|

# List of Figures

|      |                                                                                                                     |    |
|------|---------------------------------------------------------------------------------------------------------------------|----|
| 1.1  | Young's double slit experiment . . . . .                                                                            | 4  |
| 1.2  | Two antennas observing a radio source in the sky . . . . .                                                          | 5  |
| 1.3  | Visibility amplitude vs baseline length . . . . .                                                                   | 9  |
| 1.4  | Point and extended source images . . . . .                                                                          | 10 |
| 1.5  | A correlator . . . . .                                                                                              | 11 |
| 1.6  | Time and bandwidth smearing . . . . .                                                                               | 17 |
| 1.7  | Smearing effects as a function of source position . . . . .                                                         | 18 |
| 1.8  | Smearing effects in the image plane . . . . .                                                                       | 19 |
| 1.9  | Smearing (decorrelation) effects as a function of baseline length . . .                                             | 20 |
| 1.10 | MeerKAT $uv$ -coverage . . . . .                                                                                    | 21 |
| 1.11 | Simple averaging interval selection . . . . .                                                                       | 22 |
| 1.12 | Time and frequency averaging effects . . . . .                                                                      | 23 |
| 1.13 | MeerKAT baseline distribution . . . . .                                                                             | 24 |
| 1.14 | Baseline-Dependent Averaging . . . . .                                                                              | 24 |
| 3.1  | Singular values for the sky distribution with a point source at the<br>phase centre . . . . .                       | 36 |
| 3.2  | Images showing the sky distribution with a point source at the phase<br>centre. . . . .                             | 39 |
| 3.3  | Noise power profile in the visibility plane . . . . .                                                               | 41 |
| 3.4  | Singular values for the sky distribution with a point source off-centre. .                                          | 42 |
| 3.5  | Images showing the sky distribution with a point source off-centre. .                                               | 43 |
| 3.6  | Singular values for the sky distribution with point sources . . . . .                                               | 44 |
| 3.7  | Images showing the sky distribution with point sources . . . . .                                                    | 46 |
| 3.8  | Singular values for the sky distribution with point and extended sources                                            | 47 |
| 3.9  | Images showing the sky distribution with point and extended sources                                                 | 49 |
| 3.10 | Power profiles along the centre of the images for the sky distribution<br>with point and extended sources . . . . . | 50 |

|                                                                                                                         |    |
|-------------------------------------------------------------------------------------------------------------------------|----|
| 3.11 Singular values for the sky distribution with a point source off-centre with an SNR of 100 . . . . .               | 51 |
| 3.12 Images showing the sky distribution with a point source off-centre with an SNR of 100 . . . . .                    | 52 |
| 3.13 Singular values for the sky distribution with 100 point sources and a total flux of 10 Jy . . . . .                | 53 |
| 3.14 Images showing the sky distribution with 100 point sources with a total flux of 10 Jy . . . . .                    | 55 |
| 3.15 Singular values for the sky distribution with a point source at a declination of $-90^\circ$ . . . . .             | 56 |
| 3.16 Images showing the sky distribution with an off-centre point source at a declination of $-90^\circ$ . . . . .      | 57 |
| 3.17 Residual images for the sky distribution with a point source at the phase centre . . . . .                         | 59 |
| 3.18 Residual images for the sky distribution with an off-centre point source away from the phase centre . . . . .      | 60 |
| 3.19 Residual images for the sky distribution with point sources . . . . .                                              | 62 |
| 3.20 Residual images for the sky distribution with point and extended sources . . . . .                                 | 64 |
| 3.21 Residual images for the sky distribution with an off-centre point source with an SNR of 100 . . . . .              | 65 |
| 3.22 Residual images for the sky distribution with 100 point sources with a total flux of 10 Jy . . . . .               | 67 |
| 3.23 Residual images for the sky distribution with an off-centre point source at a declination of $-90^\circ$ . . . . . | 68 |
| 6.1 Singular values for the sky distribution with a point source off-centre with an SNR of 1000 . . . . .               | 77 |
| 6.2 Images showing the sky distribution with an off-centre point source with an SNR of 1000 . . . . .                   | 78 |
| 6.3 Residual images for the sky distribution with an off-centre point source with an SNR of 1000 . . . . .              | 79 |
| 6.4 Singular values for the sky distribution with a point source off-centre with an SNR of 50 . . . . .                 | 80 |

|      |                                                                                                                                                  |    |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 6.5  | Images showing the sky distribution with an off-centre point source with an SNR of 50 . . . . .                                                  | 81 |
| 6.6  | Residual images for the sky distribution with an off-centre point source with an SNR of 50 . . . . .                                             | 82 |
| 6.7  | Singular values for the sky distribution with an off-centre point source with an SNR of 5 . . . . .                                              | 83 |
| 6.8  | Images showing the sky distribution with an off-centre point source with an SNR of 5 . . . . .                                                   | 84 |
| 6.9  | Residual images for the sky distribution with an off-centre point source with an SNR of 5 . . . . .                                              | 85 |
| 6.10 | Singular values and their corresponding weights for the sky distribution with an off-centre point source at a declination of $-60$ deg . . . . . | 86 |
| 6.11 | Images showing the sky distribution with a point source at a distance of $0.4$ deg and declination of $-60$ deg . . . . .                        | 87 |
| 6.12 | Residual images for the sky distribution with a point source at a distance of $0.4$ deg and a declination $-60$ deg . . . . .                    | 88 |
| 6.13 | Singular values and their corresponding weights for the sky distribution with an off-centre point source at a declination of $0$ deg . . . . .   | 89 |
| 6.14 | Images showing the sky distribution with a point source at a distance of $0.4$ deg and declination of $0$ deg . . . . .                          | 90 |
| 6.15 | Residual images for the sky distribution with a point source at a distance of $0.4$ deg and a declination $0$ deg . . . . .                      | 91 |

# List of Tables

|     |                                                                                                            |    |
|-----|------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Compression factors. The data sizes are given in units of GigaByte (GB) . . . . .                          | 40 |
| 3.2 | Recovered peak brightness for a sky distribution with SNR of 100. . .                                      | 64 |
| 3.3 | Recovered peak brightness for a sky distribution with multiple sources with a total flux of 10 Jy. . . . . | 66 |
| 3.4 | Summary of BDSVD compression performance across different scenarios . . . . .                              | 69 |
| 6.1 | Recovered peak brightness for a sky distribution with SNR of 1000. .                                       | 77 |
| 6.2 | Recovered peak brightness for a sky distribution with SNR of 50. . .                                       | 80 |
| 6.3 | Recovered peak brightness for a sky distribution with SNR of 5. . . .                                      | 83 |



# Listings

|     |                                         |    |
|-----|-----------------------------------------|----|
| 4.1 | The simple SVD command. . . . .         | 70 |
| 4.2 | The BDSVD command. . . . .              | 72 |
| 4.3 | The data decompression command. . . . . | 72 |

# Abbreviations

**ASKAP** Australian Square Kilometre Array Pathfinder.

**AVN** African VLBI Network.

**BDA** Baseline Dependent Averaging.

**BDSVD** Baseline-Dependent Singular Value Decomposition.

**DDE** Direction Dependent Effects.

**DIE** Direction-Independent Effects.

**EVN** European VLBI Network.

**FAST** Five-hundred-meter Aperture Spherical Telescope.

**FoV** Field of View.

**GB** GigaByte.

**GBT** Green Bank Telescope.

**LOFAR** Low Frequency Array.

**MIGHTEE** MeerKAT International GHz Tiered Extragalactic Exploration.

**MS** Measurement Set.

**MWA** Murchison Widefield Array.

**NRF** National Research Foundation.

**NVSS** NRAO VLA Sky Survey.

**PSF** Point Spread Function.

**RFI** Radio Frequency Interference.

**RIME** Radio Interferometric Measurement Equation.

**RMS** Root Mean Square.

**SARAO** South African Radio Astronomy Observatory.

**SARAO-HCD** South African Radio Astronomy Observatory-Human Capital Development.

**SEFD** Sytem Equivalent Flux Density.

**SKAO** Square Kilometre Array Observatory.

**SNR** Signal to Noise Ratio.

**SVD** Singular Value Decomposition.

**VLA** Very Large Array.

**VLA** VLA Sky Survey.

**VLBI** Very Long Baseline Interferometry.

**VSOP** VLBI Space Observatory Programme.

# Chapter 1

## Introduction

### 1.1 Overview of Radio Astronomy

Radio astronomy is a field that has gained prominence in recent years. Its origins trace back to the 1930s when Karl G. Jansky [9, 10], a radio engineer, discovered a radio signal coming from the Milky Way. Inspired by Jansky's work, Grote Reber [11, 12] built the first radio telescope in his backyard, a 9-meter parabolic dish, to further explore cosmic radio signals. The cosmic signals studied in radio astronomy are at the low-frequency, long-wavelength end of the electromagnetic spectrum, within the radio window ranging from 3 MHz to 300 GHz [13, 14], although these boundaries are flexible. This field has led to major discoveries and opened new avenues for astronomers. For instance, radio astronomy has revealed previously unobservable sources, such as quasars, which are extremely luminous active galactic nuclei powered by supermassive black holes [15, 16]. Another significant discovery is that of pulsars, which are highly magnetized, rotating neutron stars emitting pulses of radiation at regular intervals [17]. Additionally, radio astronomy enabled the first detection of the cosmic microwave background, the radiation emitted shortly after the Big Bang [18].

Since Grote Reber's pioneering efforts, the field of radio astronomy has advanced significantly with the construction of several larger telescopes. Among the most notable are the 100-meter Green Bank Telescope (GBT) in West Virginia, USA [19], and the now-decommissioned 305-meter Arecibo Telescope in Puerto Rico [20]. The Five-hundred-meter Aperture Spherical Telescope (FAST) [21] in China holds the title of the world's largest filled-aperture telescope and is second only to Russia's RATAN-600 [22] in single-dish size, with its 576-meter diameter circle of

rectangular radio reflectors. Single-dish telescopes face limitations in angular resolution, which is approximated as  $\theta \approx \lambda/D$  where  $\lambda$  is the observing wavelength and  $D$  is the telescope's dish diameter [14]. To improve angular resolution, increasing the dish diameter is essential, but structural and monetary constraints limit how large a telescope can be built. To overcome this challenge, interferometers use multiple dishes or antennas to observe a source simultaneously, effectively functioning as a larger telescope.

The concept of radio interferometry is rooted in optical interferometry techniques developed by Albert Michelson and Francis Pease in the 1920s for measuring the diameter of the star Betelgeuse [23, 24]. This technique was adapted by Martin Ryle and Derek Vonberg in 1946 [25] for observing the Sun and radio sources. The first radio interferometer, however, was used by McCready et al. in 1945 [26] for solar observations on a sea cliff [27]. In radio interferometry, the angular resolution is given by  $\theta \approx \lambda/B$  where  $B$  is the largest separation between a pair of dishes or antennas, known as the baseline [14]. High angular resolution is achieved by increasing the separation between antennas. For instance, the Very Large Array (VLA) [28], located in New Mexico, USA, consists of 27 antennas with a maximum baseline of 36 km and can achieve a maximum angular resolution of up to 0.04 arcseconds [29]. The SKAO [7, 30], being developed in Africa and Australia, is set to become one of the world's largest radio interferometer arrays. Its precursor, the MeerKAT telescope in Northern Cape, South Africa, has a maximum baseline of 8 km and achieves a maximum angular resolution of 1 arcsecond [1]. Once completed, the SKA is expected to provide an angular resolution of up to 20 milliarcseconds (mas) [31].

The Very Long Baseline Interferometry (VLBI) network [32] enables extremely high resolution by using single-dish telescopes located around the globe to observe the same sky position and correlate the signals, effectively creating a large-sized synthesized telescope. Several VLBI networks exist, and more are being developed, including the European VLBI Network (EVN) [33], which incorporates telescopes from Europe, Africa, Asia, and North America, and the African VLBI Network (AVN) [34], which is currently being developed and will use telescopes across Africa. Additionally, longer baselines can be achieved by correlating signals from space-based telescopes with those on the ground, as demonstrated by projects like the VLBI Space Observatory Programme (VSOP) and the RadioAstron project [35–

37].

## 1.2 Radio Interferometry Fundamentals

### 1.2.1 Young's Double Slit Experiment

Radio interferometry is analogous to Thomas Young's double slit experiment, conducted in 1801 [38, 39]. This seminal experiment provided critical insights into the wave nature of light and laid the groundwork for many modern optical and radio interferometry techniques. In Young's experiment, monochromatic light, a light of a single wavelength, passes through two closely spaced, parallel slits on an opaque plate. Let us name the holes  $p$  and  $q$ . The electric fields through these slits are described by:

$$E_p = I_p e^{i2\pi\nu t_p} \quad \text{and} \quad E_q = I_q e^{i2\pi\nu t_q} \quad (1.1)$$

where  $I_p$  and  $I_q$  are the maximum intensities of the electric fields,  $\nu$  is the frequency of the monochromatic light, and  $t_p$  and  $t_q$  are the respective arrival times at a given point.

As the light waves pass through the slits, they spread out and overlap, producing an interference pattern on the screen as illustrated in Figure 1.1. The resulting electric field is the sum of the individual fields:

$$E_{tot} = E_p + E_q. \quad (1.2)$$

The time-averaged electric field intensity  $I$  on the screen is then given by:

$$\begin{aligned} \langle I \rangle &\propto \langle E_{tot} E_{tot}^* \rangle \\ &= (E_p + E_q)(E_p + E_q)^* \\ &= I_p^2 + I_q^2 + I_p I_q e^{i2\pi\nu(t_p - t_q)} + I_p I_q e^{-i2\pi\nu(t_p - t_q)} \\ &= I_p^2 + I_q^2 + 2I_p I_q \cos(2\pi\nu\tau), \end{aligned} \quad (1.3)$$

where  $*$  denotes the complex conjugate, and  $\tau = t_p - t_q$  is the difference in arrival times. The terms  $I_p^2$  and  $I_q^2$  are constants, while the  $\cos(2\pi\nu\tau)$  term represents the

interference component, varying with the phase difference. This interference produces a pattern of alternating bright and dark fringes observed on the screen (see Figure 1.1).

Young's double-slit experiment was pivotal in demonstrating the principle of interference, a concept later adopted by Albert Michelson in the 1920s. Michelson utilized this principle to develop an optical interferometer to measure the size of the star Betelgeuse [23, 24]. In radio interferometry, this principle is adapted for use with radio waves instead of visible light. The core idea remains the same: pairs of antennas or dishes act analogously to the slits, capturing radio waves from astronomical sources. By analyzing the interference patterns of these radio waves, valuable information such as the intensity, shape, and position of the sources can be extracted.

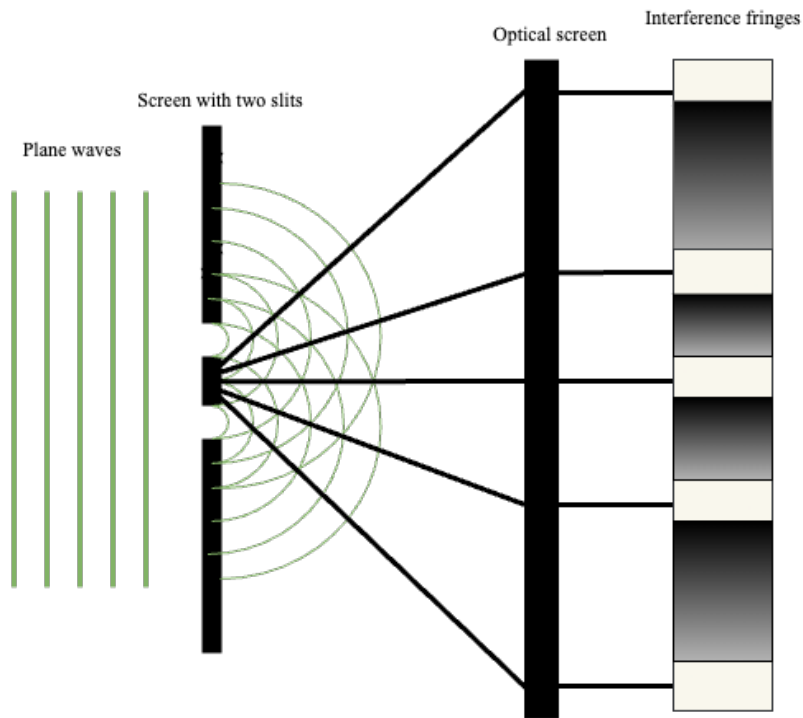


FIGURE 1.1: An illustration of Young's double slit experiment. A monochromatic plane wave passes through two narrow, closely spaced slits, producing an interference pattern on the optical screen. This pattern consists of alternating bright and dark bands caused by the constructive and destructive interference of the light waves.

Having established the foundational principles of interferometry using Thomas Young's double-slit experiment, we now turn our attention to their application in

radio interferometry. Here, these principles are used to derive the visibility function, which encodes information about the source's properties.

### 1.2.2 Visibility Function

Any pair of antennas  $p$  and  $q$  in a radio interferometer, separated by baseline vector  $\mathbf{b}$ , as illustrated in Figure 1.2, detects the electromagnetic (EM) radiation wave,  $\mathbf{e}$ , from a radio source in the sky. The EM waves are converted into complex voltages by the antenna feeds. These complex voltages are then fed into a correlator, which outputs pairwise correlations of the voltages between antennas  $p$  and  $q$ , as shown in Figure 1.5.

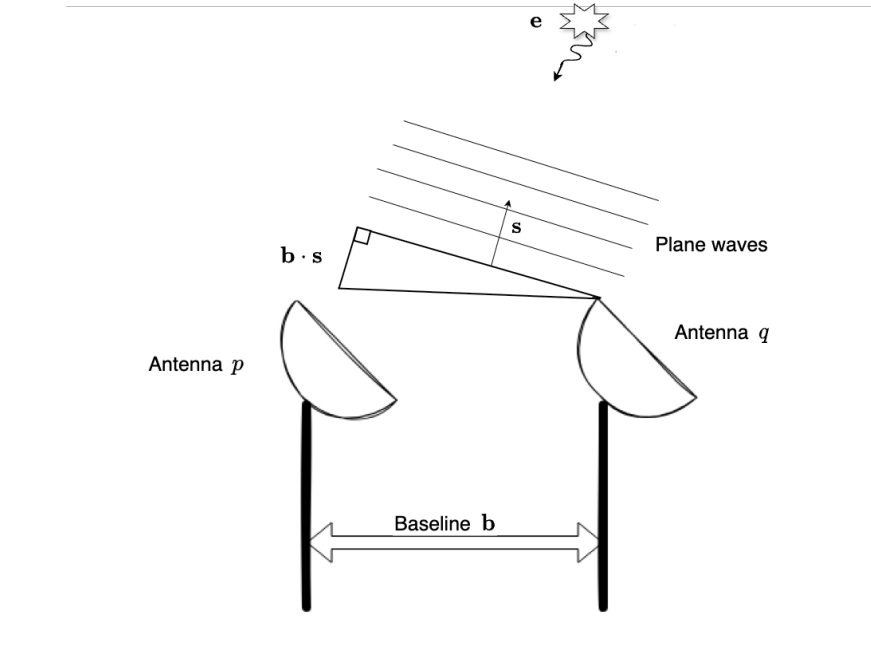


FIGURE 1.2: Two antennas,  $p$  and  $q$ , are observing a radio source. The separation between the two antennas is represented by the baseline vector  $\mathbf{b}$ . Due to the geometric delay, the signal reaches antenna  $q$  before it arrives at antenna  $p$ . This delay corresponds to the projection of the baseline  $\mathbf{b}$  onto the observation direction  $\mathbf{s}$ .

Moreover, the correlator averages these measurements over a finite period and bandwidth. Averaging of the measurements over a time interval and bandwidth reduces noise and filters out frequencies greater than a few hertz, effectively acting as a low-pass filter. The result from a correlator is what interferometers measure, known as complex visibilities. Interferometers do not measure the sky intensity directly, instead, they measure these complex visibilities. The sky brightness and the



visibility are related through the principles of Fourier transformation, as they are Fourier pairs. This relationship is known as the Van Cittert-Zernike (VCZ) theorem [39–41]. Each baseline in an interferometer probes a single spatial frequency or Fourier component, determined by the baseline's length and direction. In this way, the array samples the Fourier plane ( $uv$ -plane) at discrete points corresponding to the baseline configurations and observing geometry.

For instance, for a source located in the direction  $\mathbf{s}$ , where  $\mathbf{s}$  is a unit vector pointing toward the source, observed at a single frequency  $\nu$ , with a brightness given by  $I_\nu(\mathbf{s})$ , and an effective collecting area of  $A(\mathbf{s})$  (assuming same  $A$  for all of the antennas), the visibility is given by: [14, 40]:

$$V_{pq} = A(\mathbf{s})I_\nu(\mathbf{s})e^{-i\omega\tau}, \quad (1.4)$$

where  $\tau$  is the geometric delay defined as  $\mathbf{b} \cdot \mathbf{s}$ , with  $\mathbf{b}$  being the baseline vector, and  $\omega = 2\pi\nu$  is the angular frequency. Extending this to account for contributions from the entire sky  $\Omega$ , the visibility becomes:

$$V_{pq} = \int_{\Omega} A(\mathbf{s})I_\nu(\mathbf{s})e^{-i2\pi\nu(\mathbf{b} \cdot \mathbf{s})} d\Omega. \quad (1.5)$$

Incorporating a finite bandwidth  $\Delta\nu$ , the visibility is further generalized as:

$$V_{pq} = \int_{\Delta\nu} \int_{\Omega} A(\mathbf{s})I_\nu(\mathbf{s})e^{-i2\pi\nu(\mathbf{b} \cdot \mathbf{s})} d\Omega d\nu. \quad (1.6)$$

Let's define the baseline in a  $uvw$  coordinate system, where  $\hat{\mathbf{u}}$  points along the east,  $\hat{\mathbf{v}}$  points along the north, and  $\hat{\mathbf{w}}$  points along the phase tracking centre:

$$\mathbf{b}_\lambda = \frac{\mathbf{b}}{\lambda} = \lambda(u, v, w). \quad (1.7)$$

Let the sky positions be specified in the  $lmn$  coordinate system, where  $(l, m, n)$  are direction cosines with respect to the  $(u, v, w)$  axes:

$$\mathbf{s} = (l, m, n). \quad (1.8)$$

$\mathbf{s}$  is a unit vector, thus:

$$|\mathbf{s}| = 1 = l^2 + m^2 + n^2. \quad (1.9)$$

We then have:

$$\begin{aligned}\frac{\nu \mathbf{b} \cdot \mathbf{s}}{c} &= \frac{\mathbf{b} \cdot \mathbf{s}}{\lambda} = ul + vm + wn \\ &= ul + vm + w(\sqrt{1 - l^2 - m^2})\end{aligned}\quad (1.10)$$

and:

$$\begin{aligned}d\Omega &= \frac{dl dm}{n} \\ &= \frac{dl dm}{\sqrt{1 - l^2 - m^2}}.\end{aligned}\quad (1.11)$$

The complex visibility function can then be written as:

$$V_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} A(l, m) I_v(l, m) e^{-i2\pi(ul+vm+w(\sqrt{1-l^2-m^2}))} \frac{dl dm}{\sqrt{1 - l^2 - m^2}}. \quad (1.12)$$

The term  $d\nu$  disappears in Equation 1.12 because we assume a single frequency. The phase centre, which is the position on which the synthesized Field of View (FoV) is centred, is defined as a direction  $\mathbf{s}_0$ , where  $l = m = 0$ , and:

$$\frac{\nu \mathbf{b} \cdot \mathbf{s}_0}{c} = \frac{\mathbf{b} \cdot \mathbf{s}_0}{\lambda} = w. \quad (1.13)$$

The direction  $\mathbf{s}$ , relative to the phase centre is defined as:

$$\boldsymbol{\sigma} = \mathbf{s} - \mathbf{s}_0. \quad (1.14)$$

Thus, for a source located in the direction  $\mathbf{s}$ , the complex visibility is given by:

$$V_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} A(l, m) I_v(l, m) e^{-i2\pi(ul+vm+w(\sqrt{1-l^2-m^2}-1))} \frac{dl dm}{\sqrt{1 - l^2 - m^2}}. \quad (1.15)$$

This gives:

$$\begin{aligned}
V_{pq} &= A(0)I_v(0)e^{-i2\pi(0)} \\
&= A(0)I_v(0) = \text{constant},
\end{aligned} \tag{1.16}$$

at the phase centre.

For small fields of view ( $l, m \ll 1$ ) and small  $w$ , we approximate  $w(\sqrt{1-l^2-m^2}-1) \approx -\frac{1}{2}(l^2+m^2)w$ . This simplifies the visibility function to:

$$V_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{A(l, m)}{\sqrt{1-l^2-m^2}} I_v(l, m) e^{-i2\pi(ul+vm)} dl dm. \tag{1.17}$$

For an unresolved or point source, the intensity distribution is independent of  $l, m$ , such that  $I(l_0, m_0) = I_0$ . The corresponding visibility function is then given by:

$$V_{pq} = I_0 e^{-i2\pi(ul_0+vm_0)}, \tag{1.18}$$

where  $I_0$  is the total intensity. The visibility amplitude is then:

$$\begin{aligned}
|V_{pq}| &= |I_0 e^{-i2\pi(ul_0+vm_0)}| \\
&= |I_0| \cdot |e^{-i2\pi(ul_0+vm_0)}| \\
&= I_0.
\end{aligned} \tag{1.19}$$

Thus, for a point source, the visibility amplitude remains constant. For a resolved or extended source, the visibility function is given by:

$$V_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} I(l, m) e^{-i2\pi(ul+vm)} dl dm. \tag{1.20}$$

The integral sums contributions from different parts of the source, each with varying phases that partially cancel one another. As a result, the visibility amplitude  $|V_{pq}|$  becomes dependent on the baseline length. For short baselines (small  $u, v$  values), the amplitude remains large. On the other hand, for long baselines (large  $u, v$  values), the amplitude is significantly reduced and approaches zero, i.e.,  $|V_{pq}| \rightarrow 0$ .

In Figure 1.3, we plot the visibility amplitude as a function of baseline length for both a point source and an extended source, shown in Figure 1.4, each with an intensity of 1 Jy. For the point source, the visibility amplitude remains constant

across all baseline lengths. In contrast, for the extended source, the visibility amplitude decays rapidly as the baseline length increases.

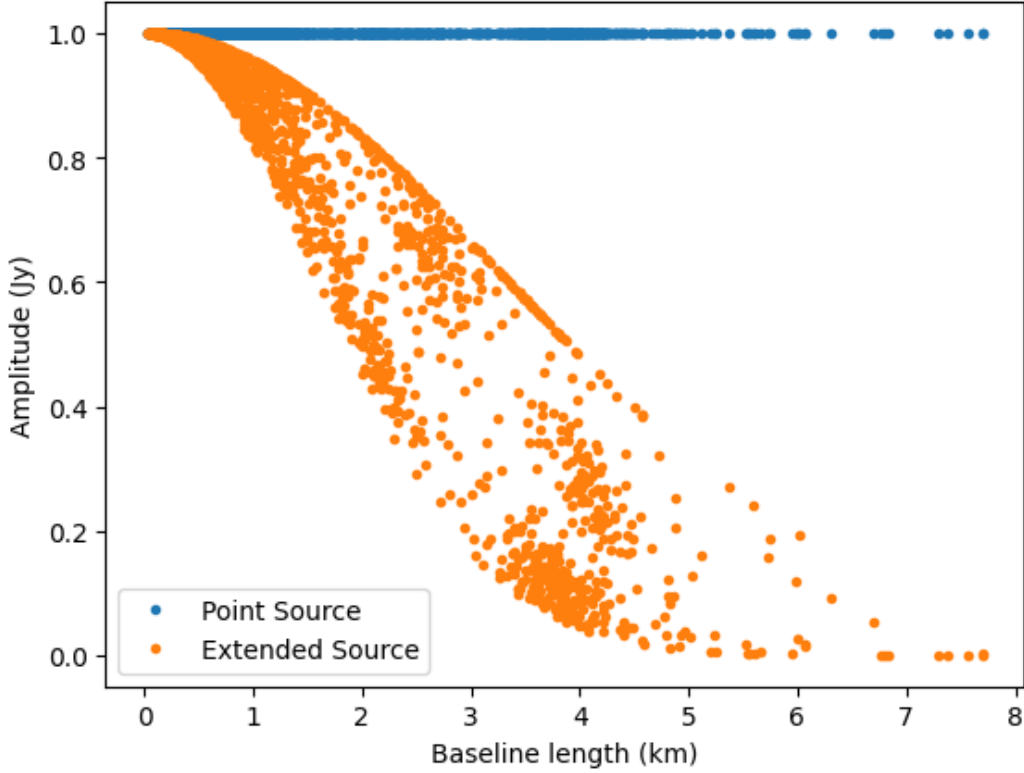


FIGURE 1.3: Visibility amplitude as a function of baseline length for both a point and an extended source at the phase centre, each with an amplitude of 1 Jy. For the point source, the visibility amplitude remains constant across different baseline lengths. In contrast, the visibility amplitude for the extended source decreases rapidly with increasing baseline length. The data was simulated for the MeerKAT telescope without any noise.

The relationship between the sky brightness and the visibility in Equation 1.17 is known as the Van Cittert-Zernike (VCZ) theorem [39–41]. It can alternatively be articulated within the framework of the Radio Interferometric Measurement Equation (RIME). This formalism was originally developed by Hamaker et al. [42].

### 1.2.3 The Radio Interferometric Measurement Equation

In the context of a 2-element interferometer comprising antennas  $p$  and  $q$ , each antenna is equipped with two receptors,  $a$  and  $b$ , to capture the complete polarization information. These receptors convert the electromagnetic signal into complex voltages, enabling each antenna to measure the following [43]:

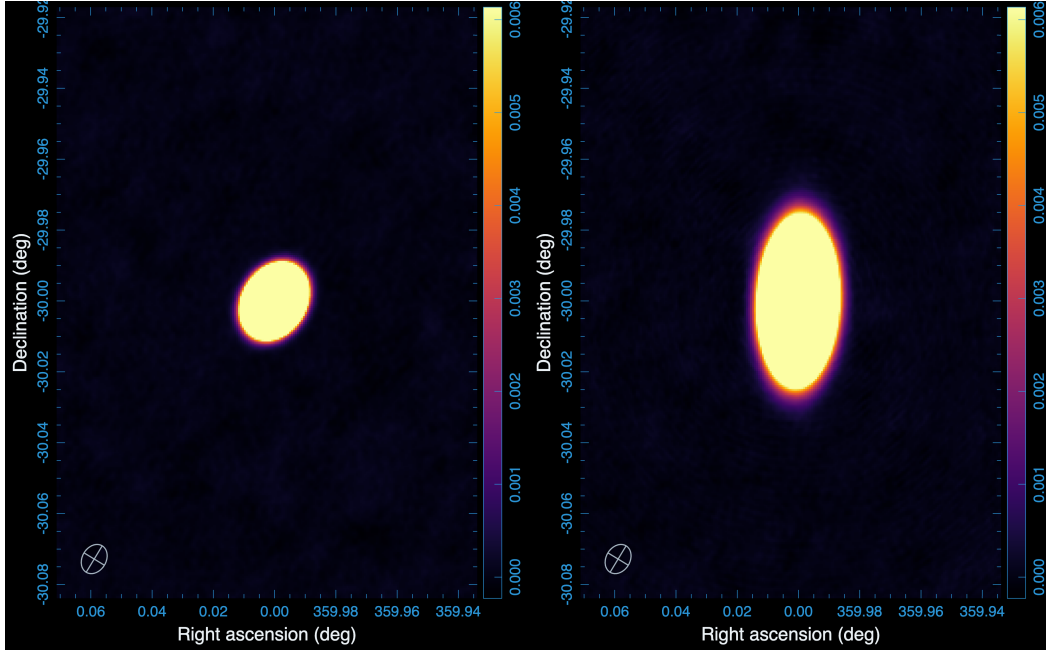


FIGURE 1.4: The source models used in Figure 1.3. The left panel displays the point source, while the right panel illustrates the extended source.

$$\begin{aligned}\mathbf{v}_p &= \begin{pmatrix} v_{pa} \\ v_{pb} \end{pmatrix} \\ \mathbf{v}_q &= \begin{pmatrix} v_{qa} \\ v_{qb} \end{pmatrix}.\end{aligned}\tag{1.21}$$

The complex voltages are subsequently fed into a correlator, which, in turn, generates four pairwise correlations, as illustrated in Figure 1.5:

$$\begin{aligned}v_{aa} &= \langle v_{pa} v_{qa}^* \rangle \\ v_{ab} &= \langle v_{pa} v_{qb}^* \rangle \\ v_{ba} &= \langle v_{pb} v_{qa}^* \rangle \\ v_{bb} &= \langle v_{pb} v_{qb}^* \rangle.\end{aligned}\tag{1.22}$$

The  $\langle \cdot \rangle$  operator denotes that these correlations are averaged over a finite time and

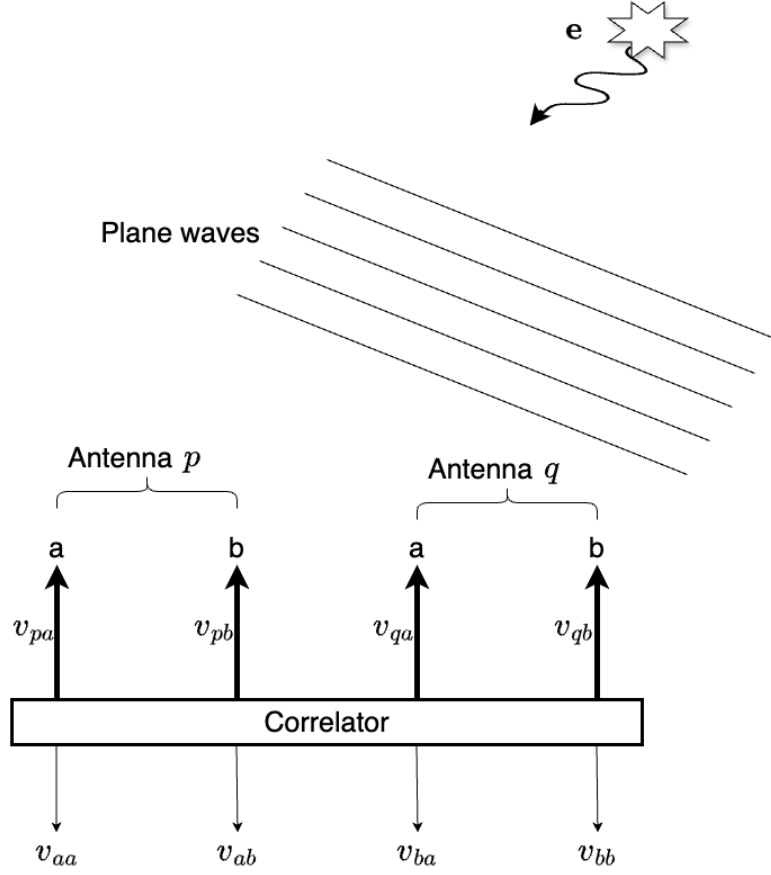


FIGURE 1.5: The signal, denoted by  $e$ , is converted into complex voltages by antenna receptors  $a$  and  $b$ . Antenna receptor  $a$  measures the horizontally polarized component of the signal, while receptor  $b$  measures the vertically polarized signal. The correlator then computes four pairwise correlations.

bandwidth, and the  $*$  symbol represents the complex conjugate operation. These four correlations are then used to form the visibility matrix:

$$\begin{aligned}
 V_{pq} &= 2 \begin{pmatrix} v_{aa} & v_{ab} \\ v_{ba} & v_{bb} \end{pmatrix} \\
 &= 2 \left\langle \begin{pmatrix} v_{pa} \\ v_{pb} \end{pmatrix} \begin{pmatrix} v_{qa}^* & v_{qb}^* \end{pmatrix} \right\rangle \\
 &= 2 \langle \mathbf{v}_p \mathbf{v}_q^H \rangle.
 \end{aligned} \tag{1.23}$$

Where  $^H$  denotes the Hermitian conjugate transpose operation, and the factor of 2 is a matter of convention as explained by Smirnov [43]. The complex voltages can

be connected to the EM field vector  $\mathbf{e}$  in the following way:

$$\begin{aligned}\mathbf{v}_p &= \mathbf{J}_p \mathbf{e} \\ \mathbf{v}_q &= \mathbf{J}_q \mathbf{e}.\end{aligned}\tag{1.24}$$

In equation 1.24, the Jones matrix  $\mathbf{J}$  accounts for the effects that distort the signal  $\mathbf{e}$  as it travels from the source to the antennas. Each antenna has its own unique Jones matrix, representing the different paths the signal takes to reach each antenna. The visibility matrix is then obtained as follows:

$$\begin{aligned}V_{pq} &= 2\langle (\mathbf{J}_p \mathbf{e})(\mathbf{J}_q \mathbf{e})^H \rangle \\ &= 2\langle \mathbf{J}_p \mathbf{e} \mathbf{e}^H \mathbf{J}_q^H \rangle \\ &= \mathbf{J}_p 2\langle \mathbf{e} \mathbf{e}^H \rangle \mathbf{J}_q^H.\end{aligned}\tag{1.25}$$

It is important to note that the Jones matrix is placed outside the averaging operator. This positioning reflects our assumption that the Jones matrix remains constant over the averaging interval. This assumption has significant implications, which we will discuss further.

The quantity  $\langle \mathbf{e} \mathbf{e}^H \rangle$  represents the source brightness distribution, denoted by  $\mathbf{B}$ , and it is related to the Stokes parameters as follows [14, 39, 40]:

$$\begin{aligned}\mathbf{B} &= 2\langle \mathbf{e} \mathbf{e}^H \rangle = 2 \begin{pmatrix} e_x e_x^* & e_x e_y^* \\ e_y e_x^* & e_y e_y^* \end{pmatrix} \\ &= \begin{pmatrix} I + Q & U + iV \\ U - iV & I - Q \end{pmatrix},\end{aligned}\tag{1.26}$$

where  $x$  and  $y$  denote the two orthogonal polarization directions in the sky reference frame. Note that the antenna feed polarization indices  $a, b$  are transformed into the sky polarization basis  $x, y$  through the parallactic angle Jones matrix. This gives us the RIME of a single-point source:

$$V_{pq} = \mathbf{J}_p \mathbf{B} \mathbf{J}_q^H.\tag{1.27}$$

However, the sky is not made up of a point source but instead consists of a continuous brightness distribution. To generalize the RIME to the entire sky, we incorporate a dependence of the brightness distribution  $\mathbf{B}$  on the observation direction  $\mathbf{s}$ :

$$V_{pq} = \iint_{4\pi} \mathbf{J}_p(\mathbf{s}) \mathbf{B}(\mathbf{s}) \mathbf{J}_q^H(\mathbf{s}) d\mathbf{s}, \quad (1.28)$$

where we integrate over the entire FoV denoted as  $\Omega = 4\pi$ . Using this formulation, we can modify the variable in the integration:

$$\begin{aligned} V_{pq} &= \int_l \int_m \mathbf{J}_p(l, m) \mathbf{B}(l, m) \mathbf{J}_q^H(l, m) \frac{d\mathbf{s}}{dl dm} dl dm \\ &= \int_l \int_m \mathbf{J}_p(l, m) \mathbf{B}(l, m) \mathbf{J}_q^H(l, m) \frac{1}{n} dl dm. \end{aligned} \quad (1.29)$$

The term  $\frac{1}{n}$  arises as a result of switching the integration volume from  $d\mathbf{s}$  to  $dl dm$ . We can express the Jones matrix  $\mathbf{J}_p$  as a Jones chain, which consists of multiple Jones matrices. Each of these matrices represents a specific propagation effect:

$$\mathbf{J}_p = \mathbf{J}_{p,n} \mathbf{J}_{p,n-1} \cdots \mathbf{J}_{p,1}. \quad (1.30)$$

In this chain, the rightmost term corresponds to the effects that distort the signal at the source, while the leftmost term corresponds to the effects that distort the signal at the antennas. The Jones terms can be further categorized based on whether they are Direction Dependent Effects (DDE) - effects dependent on  $\mathbf{s} = (l, m)$ , or Direction-Independent Effects (DIE) - effects independent of  $\mathbf{s} = (l, m)$ .

The geometric delay term, represented as  $K_p$ , is always included in the Jones chain. This term accounts for the delay in signal reception between the two antennas. As shown in Figure 1.2, the signal arrives at antenna  $q$  before antenna  $p$ . The phase term can be expressed as follows:

$$K_p = e^{-2\pi i(\mathbf{u}_{p\lambda} \cdot \mathbf{s})} \quad (1.31)$$

In Equation 1.31,  $\mathbf{u}_{p\lambda}$  denotes the antenna position of antenna  $p$  in the  $uvw$  coordinate system expressed in wavelengths. Since  $K_p$  is a scalar, it commutes with all



other operators, enabling us to rearrange it freely. Therefore, we can simplify the expression as follows:

$$\begin{aligned}
 K_{pq} &= K_p K_q^H \\
 &= e^{-2\pi i((\mathbf{u}_{p\lambda} - \mathbf{u}_{q\lambda}) \cdot \mathbf{s})} \\
 &= e^{-2\pi i(ul + vm + wn)},
 \end{aligned} \tag{1.32}$$

where  $u, v$ , and  $w$  correspond to the components of the baseline vector  $\mathbf{b}_\lambda = \mathbf{u}_p - \mathbf{u}_q$ .

The visibility phase varies with the baseline length and the distance from the phase centre. This phase variation is crucial because the complex exponential term in the visibility equation serves as the Fourier kernel. As a result, the measured visibilities can be interpreted as samples of the Fourier Transform of the sky brightness distribution. Without this spatially varying phase relationship, interferometers would be unable to recover positional information through inverse Fourier transforms, making aperture synthesis imaging impossible.

To mitigate smearing effects, discussed in detail in Section 1.2.4, the fringe stopping technique is employed, involving the insertion of an artificial delay in the correlator within the signal chain. The phase term needed for insertion in the correlator can be determined by considering a wavefront arriving from direction  $\mathbf{s}_0$  (phase centre), and the phase term can be expressed as follows:

$$e^{-2\pi i(\mathbf{b}_\lambda \cdot \mathbf{s}_0)} = e^{-2\pi i(w)}. \tag{1.33}$$

The resulting  $K$ -term is:

$$\begin{aligned}
 K_{pq} &= e^{-2\pi i(ul + vm + wn)} e^{-2\pi i(w)} \\
 &= e^{-2\pi i(ul + vm + w(n-1))}.
 \end{aligned} \tag{1.34}$$

The Jones term is then expressed as follows:

$$\mathbf{J}_p = \mathbf{G}_p \mathbf{E}_p(l, m) K_p(l, m), \tag{1.35}$$

where  $\mathbf{G}_p$  represents the direction-independent Jones term, while  $\mathbf{E}_p(l, m)$  is the direction-dependent Jones term. The visibility function then becomes:

$$V_{pq} = \mathbf{G}_p \left( \int_l \int_m \frac{1}{n} \mathbf{E}_p(l, m) \mathbf{B}(l, m) \mathbf{E}_q^H(l, m) e^{-2\pi i(ul+vm+w(n-1))} dl dm \right) \mathbf{G}_q^H. \quad (1.36)$$

To obtain a 2D-Fourier transform, the  $w$ -term is split into two per-antenna terms. Since  $w = w_p - w_q$ , we have:

$$W_p = \frac{1}{\sqrt{n}} e^{w_p(n-1)}. \quad (1.37)$$

This term is considered as a DDE and it is incorporated into  $\mathbf{E}_p$ . Thus, we now have:

$$V_{pq} = \mathbf{G}_p \left( \int_l \int_m \mathbf{B}_{pq}(l, m) e^{-2\pi i(ul+vm)} dl dm \right) \mathbf{G}_q^H. \quad (1.38)$$

This is the 2D Fourier Transform. The  $pq$  indices in  $\mathbf{B}_{pq}$  indicate that each baseline observes a distinct portion of the sky.  $\mathbf{B}_{pq}$  is referred to as the apparent sky, expressed as  $\mathbf{B}_{pq} = \mathbf{E}_p \mathbf{B} \mathbf{E}_q$ . Assuming that each baseline consistently observes the same apparent sky  $\mathbf{B}_{app}$  throughout the observation, this can be expressed as:

$$V_{pq} = \mathbf{G}_p \left( \int_l \int_m \mathbf{B}_{app}(l, m) e^{-2\pi i(ul+vm)} dl dm \right) \mathbf{G}_q^H. \quad (1.39)$$

This equation can also be expressed more concisely as:

$$V_{pq} = \mathbf{G}_p X_{pq} \mathbf{G}_q^H. \quad (1.40)$$

This is the RIME for the full sky scenario. As noted earlier, both direction-independent and direction-dependent gain errors vary with time and frequency. These errors affect the observed visibilities, requiring a calibration process to remove their effects. Calibration involves estimating and correcting the gain error terms. Notably, methodologies for calibrating direction-independent effects and direction-dependent effects were introduced by Smirnov [44, 45].

### 1.2.4 Smearing Effects

When deriving the RIME, it is assumed that the Jones terms remain constant during the averaging interval. However, in practice, these terms vary with time and frequency. For instance, the phase  $K$ -term in Equation 1.34 depends on both the baseline vector and the frequency. Moreover, since the baseline vector in the  $uvw$  coordinate system changes over time due to the Earth's rotation, the  $K$ -term becomes time-dependent. Averaging such a term leads to a reduction in the brightness or intensity amplitude because a complex quantity with a rotating complex phase is being averaged. This amplitude reduction is known as time and bandwidth smearing in the image plane and as decorrelation in the visibility plane [46]. Smearing effects result in the broadening of the observed source. Bandwidth smearing causes a radial broadening and reduction in the amplitude of a source, which increases with the observational spectral resolution. Time smearing causes azimuthal broadening and reduction in amplitude, which increases with the observational temporal resolution. Time smearing is also dependent on the elliptical tracks that each baseline traces along the  $uv$ -plane. Baselines move faster along the time axis than the frequency axis, making time smearing more pronounced than frequency smearing.

Time smearing, which causes azimuthal broadening, is more severe for sources at high declinations due to faster baseline movement. Both time and bandwidth smearing lead to source broadening and amplitude reduction, though the total integrated flux density remains constant. The  $K$ -term, which depends on baseline length and the source's position relative to the phase centre, worsens smearing for longer baselines and sources away from the phase centre. As illustrated in Figure 1.6, the amplitude is highest, and the source remains compact at the phase centre, while smearing causes broadening and amplitude reduction away from it.

In Figure 1.7, point sources with a brightness of 1 Jy are positioned at varying distances from the phase centre, simulated for the MeerKAT telescope at 1 GHz. Without smearing effects, the recovered amplitude fluctuates around 1 Jy, primarily due to thermal noise. However, with smearing effects, the amplitude rapidly decreases as the distance from the phase centre increases. At this frequency, the MeerKAT telescope's Field of View (FoV) is approximately 1.6 deg, with a radial FoV of about 0.78 deg. The amplitude decay is faster beyond the field of view but does not immediately drop to zero, indicating that sources outside the field of view can still be detected and may contaminate measurements.

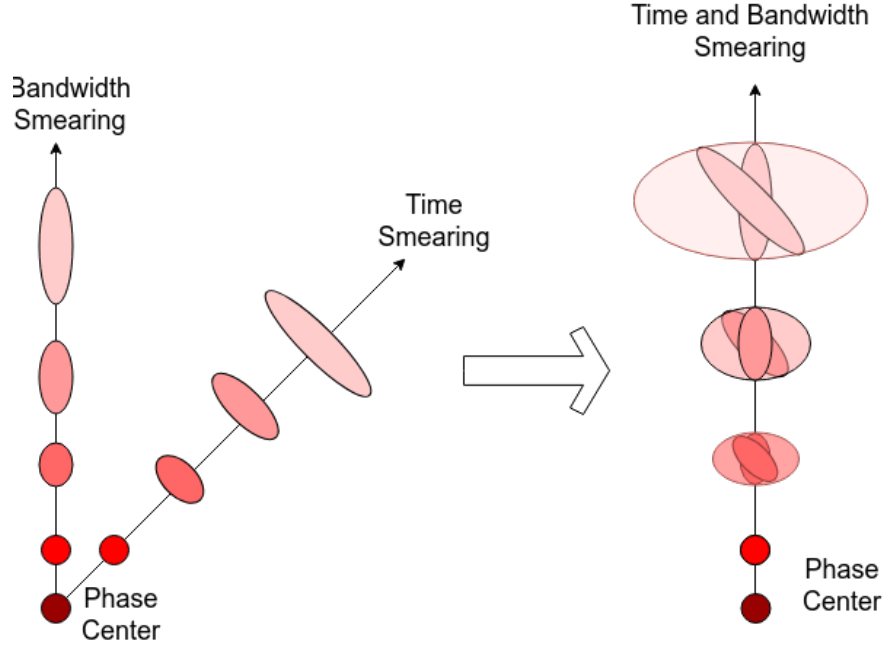


FIGURE 1.6: Time and bandwidth smearing effects. The darker shades represent higher amplitudes, while the lighter shades indicate lower amplitudes. The smearing effect worsens with increasing distance from the phase centre, resulting in reduced peak amplitude and broadened source appearance.

Figure 1.8 shows the smearing effects in the image plane for a source located 2 deg away from the phase centre, with a brightness of 1 Jy. The left panel illustrates the compact source without smearing, whereas the right panel shows the smeared source, with its brightness spread out.

Furthermore, Figure 1.9 illustrates how smearing (decorrelation) effects depend on baseline length for a source located 2 deg away from the phase centre. Without smearing effects, the recovered amplitude remains nearly constant across all baselines. However, when smearing is present, the amplitude decreases as the baseline length increases.

This behaviour confirms what is known; longer baselines experience greater angular speed in the  $uv$ -plane, resulting in more severe smearing. Figure 1.10 illustrates the  $uv$ -coverage for a 4-hour observation using the MeerKAT telescope. Notably, long baselines at the edges display faster angular motion and trace longer tracks, while short baselines remain concentrated near the centre with slower movement.

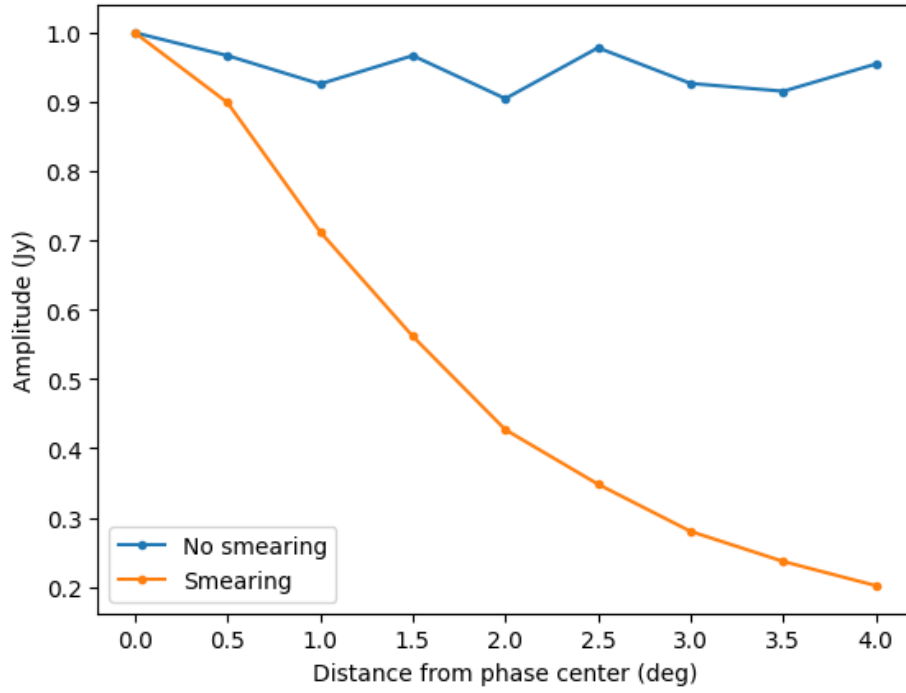


FIGURE 1.7: Smearing effects as a function of source position. The data was simulated using the MeerKAT telescope over a 4-hour observation period with an integration time of 20 seconds, at a frequency of 1 GHz with a channel bandwidth of 4 MHz. The sky model consists of 9 point sources, each with a flux of 1 Jy.

Having established a clear understanding of the fundamentals of radio interferometry, the mathematical framework underlying visibility measurements, and the smearing effects that arise from averaging the visibility function, which results in blurred source emissions and reduced source amplitude, we turn our attention to addressing the challenges of data storage and processing through data compression. The following section explores both traditional and newly proposed radio interferometric data compression strategies.

### 1.3 Compression of Visibility Data

The correlated visibilities sampled by antennas  $p$  and  $q$  are averaged in the correlator over finite time and frequency intervals,  $\Delta t$  and  $\Delta \nu$ , respectively. Post-correlation, the visibility data can be further averaged to reduce its volume. The large size of visibility datasets primarily arises from high temporal and spectral resolutions and the number of baseline configurations. High temporal resolution is

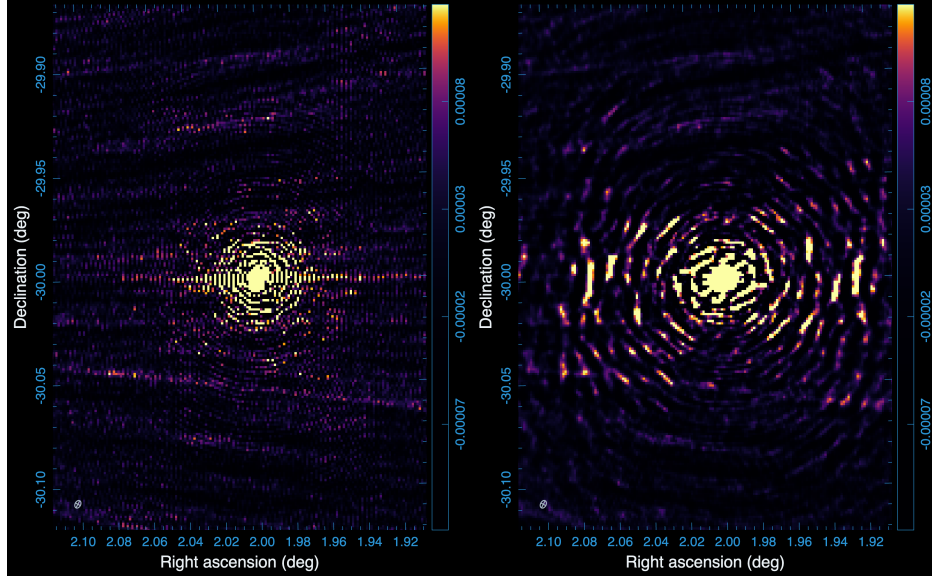


FIGURE 1.8: Smearing effects in the image plane for a source located 2 deg away from the phase centre, with a brightness of 1 Jy. The left panel shows the source without smearing effects, and the right panel shows the smeared source.

necessary for time-dependent calibration, while high spectral resolution is essential for spectral line studies and for mitigating man-made Radio Frequency Interference (RFI) [44]. Additionally, these high resolutions are required to limit smearing to acceptable levels.

Modern radio interferometers operate with extremely high temporal and spectral resolutions, generating massive data volumes that are expensive to store and process. The size of visibility data scales as follows [47]:

$$\text{data size} = n_b \times \frac{T_{\text{obs}}}{\Delta t} \times \frac{B_w}{\Delta \nu} \times n_{\text{pol}} \times 32 \text{ bits}, \quad (1.41)$$

where  $n_b$  is the number of baselines,  $T_{\text{obs}}$  is the total observing time, and  $\Delta t$  is the temporal resolution (integration time).  $B_w$  is the total bandwidth,  $\Delta \nu$  is the frequency resolution (channel width), and  $n_{\text{pol}}$  is the number of polarizations. For radio data, each complex visibility is stored as a 32-bit floating point value [48]. High temporal and spectral resolutions correspond to small values of  $\Delta t$  and  $\Delta \nu$ , which contribute to the large data sizes.

To address the storage and processing challenges posed by these massive datasets, various data compression techniques have been proposed and implemented.

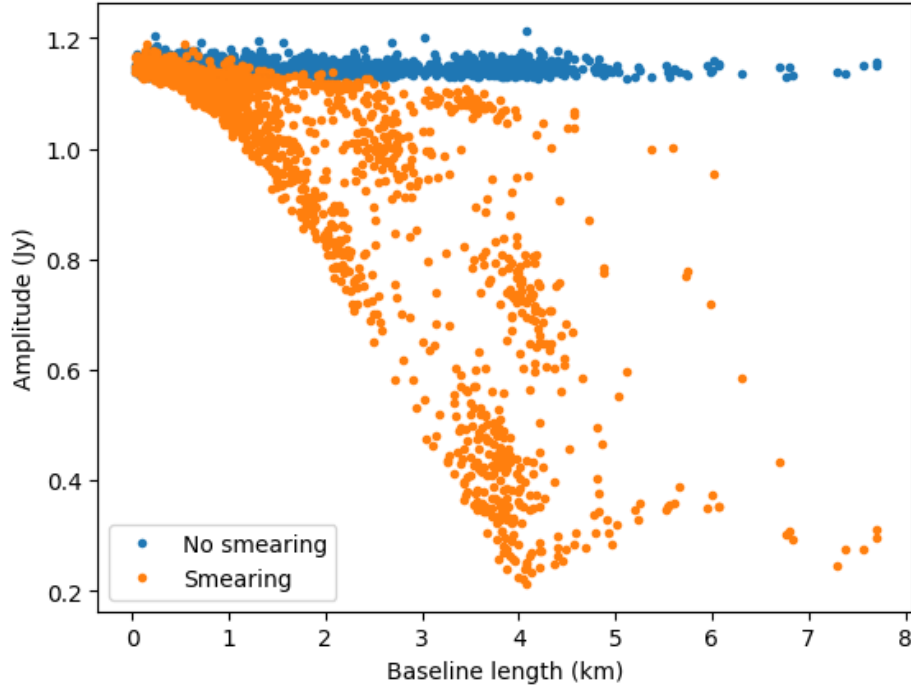


FIGURE 1.9: Smearing (decorrelation) effects as a function of baseline length. The data was simulated using the MeerKAT telescope over a 4-hour observation period with an integration time of 20 seconds, at a frequency of 1 GHz with a channel bandwidth of 4 MHz. The sky model consists of 1 point source 2 deg away from the phase centre, with a flux of 1 Jy.

### 1.3.1 Simple Averaging

Simple averaging is a basic and straightforward data compression technique [49]. It involves computing the arithmetic mean of the visibility data for a baseline  $pq$ ,  $V_{pq}$ , over a time interval  $\Delta t$  and frequency interval  $\Delta \nu$ .

In simple averaging, all baselines are averaged over equal time and frequency intervals  $\Delta t \times \Delta \nu$ . However, this uniform approach is not ideal because the fringe rate, or interference pattern, changes more rapidly for longer baselines. As a result, longer baselines need more frequent sampling to avoid smearing effects. Figure 1.11 illustrates this relationship between baseline length and the required averaging interval. With simple averaging, an intermediate interval is typically chosen, as shown in the figure. This interval is smaller than what shorter baselines require, which benefits them, but larger than what longer baselines need, leading to smearing effects for the longer baselines [46].

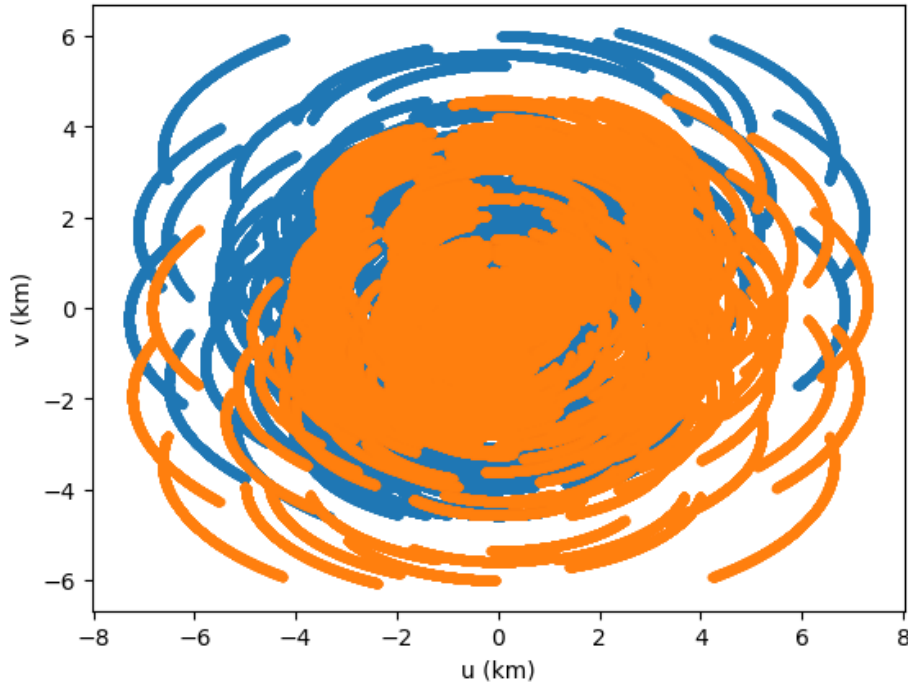


FIGURE 1.10:  $uv$ -coverage of the MeerKAT telescope for a 4-hour observation at a declination of 30 degrees South. An interferometer  $uv$ -coverage is given by the  $uv$  points (blue in the figure) and their complex conjugate (orange in the figure). The longer baselines, located at the edges, exhibit greater angular speed compared to the shorter baselines, which are concentrated towards the centre. This difference in angular speed highlights the varying data sampling rates across different baselines.

Figure 1.12 illustrates the smearing effects caused by time and frequency averaging. The dataset is simulated for the MeerKAT telescope, with a total observing time of 4 h, sampled every second at a frequency of 1.4 GHz, using 64 channels and a frequency resolution of 50 kHz. The sky distribution consists of 9 point sources, each with a brightness of 1 Jy.

The left panel of Figure 1.12 shows the effects of time averaging over various intervals, while the right panel demonstrates the impact of frequency averaging. In both cases, increasing the averaging interval leads to a greater reduction in the recovered brightness amplitude, with the effect being more pronounced for sources located away from the phase centre.

Furthermore, Figure 1.12 highlights that amplitude decay occurs more gradually for frequency averaging. This is because the baselines move more rapidly along the time axis than the frequency axis, leading to more pronounced smearing along the



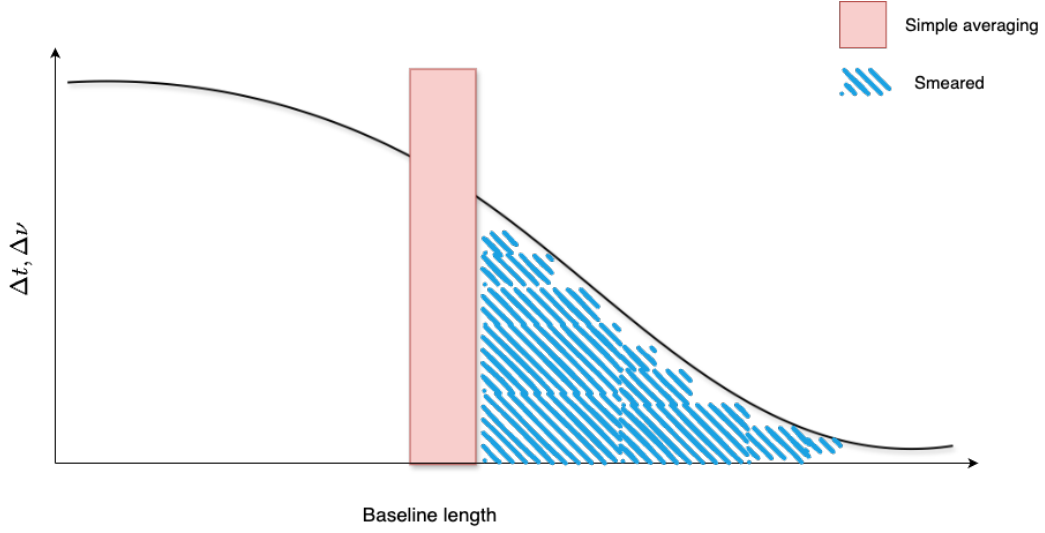


FIGURE 1.11: Simple averaging formalism. An intermediate interval is selected for averaging across all the baselines.

time domain. These smearing effects, which arise from time and frequency averaging, exhibit a non-convolutional nature in the image plane. This means they vary with position across the FoV and cannot be simply represented as a convolution of the Point Spread Function (PSF), which is the interferometer's response to a point source, with the true brightness distribution. Consequently, they cannot be mitigated through calibration and persist even after simple deconvolution. Deconvolution is a process that removes the PSF from the observed brightness distribution by assuming that the distortion is uniform across the image.

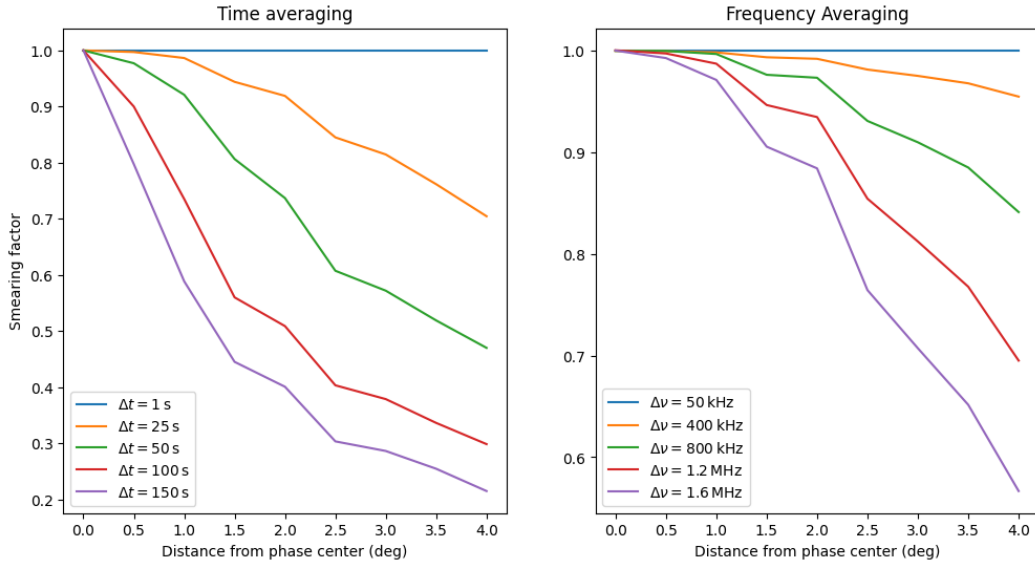


FIGURE 1.12: The effects of averaging visibility data over time and frequency. The data is simulated for the MeerKAT array with a total observing time of 4h, sampled every second at frequency of 1.4GHz, with 64 channels and a frequency resolution of 50kHz, resulting in a total bandwidth of 3.2MHz. The sky model consists of 9 point sources placed at various positions, each with a flux of 1Jy. The smearing factor is computed as the ratio between the averaged data and the high-resolution unaveraged data.

### 1.3.2 Baseline Dependent Averaging

In Figure 1.11, we observed that the averaging interval is baseline-dependent. However, simple averaging often employs an intermediate interval, which introduces smearing on long baselines due to their requirement for frequent sampling and shorter averaging intervals. To mitigate this issue, Cotton proposed Baseline Dependent Averaging (BDA) [50] as a solution.

In interferometer arrays with centrally condensed uv-coverage, such as the MeerKAT, the majority of the data originates from shorter baselines (see Figure 1.10 and 1.13). These shorter baselines can be aggressively averaged without compromising data fidelity. By averaging shorter baselines over longer intervals and longer baselines over shorter intervals, as illustrated in Figure 1.14, BDA achieves high compression rates with minimal smearing effects.

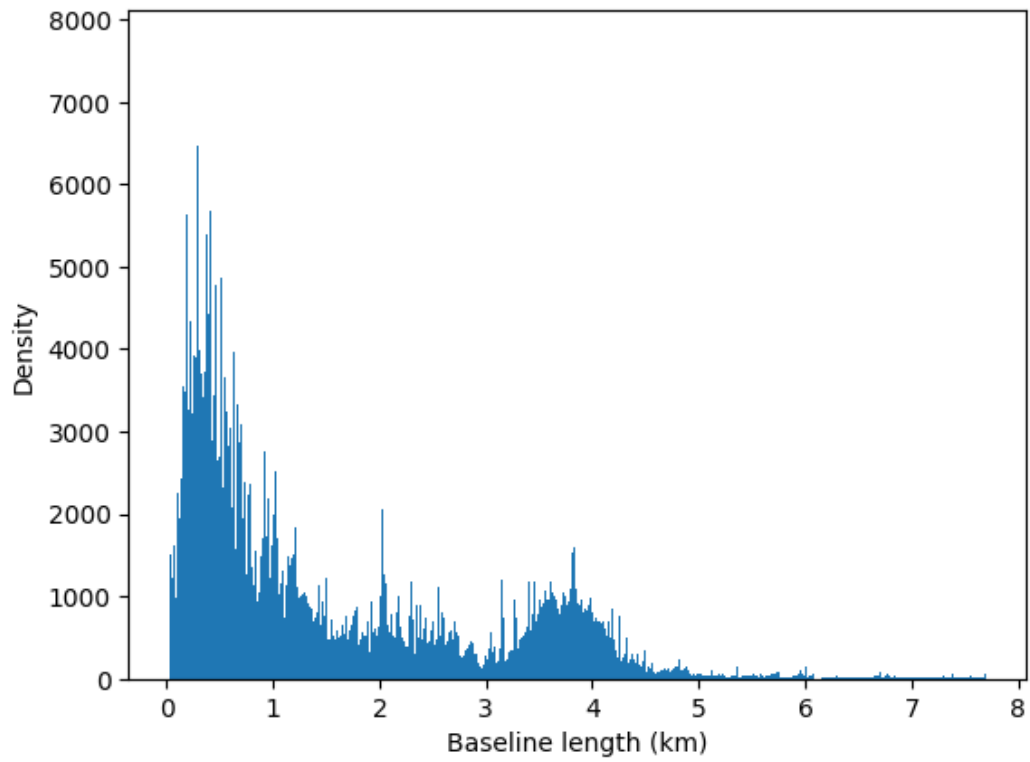


FIGURE 1.13: Baseline distribution for the MeerKAT array over a 4 h observation.

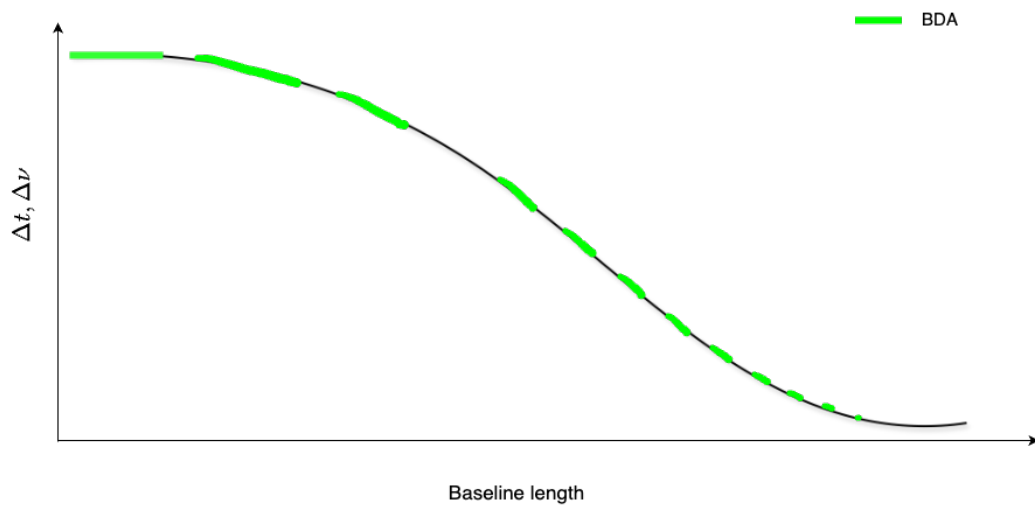


FIGURE 1.14: BDA formalism. The averaging interval is baseline-dependent, with shorter baselines being more aggressively averaged and longer baselines undergoing less averaging.

BDA has been extensively explored in the literature. For example, Cotton [51] investigated the effects of averaging visibility data using BDA and demonstrated a significant reduction in data volume with minimal smearing. Wijnholds et al. [52] demonstrated that BDA could reduce visibility data volume for the SKA1-mid telescope by over 80% along the time axis, and Deng [53] et al. further supported these findings. Additionally, Atemkeng et al. [54, 55] demonstrated that combining BDA with window functions effectively suppresses sources outside the field of interest, reducing smearing and unwanted signals while maintaining compression rates similar to simple averaging.

However, despite its benefits, BDA has significant drawbacks. For example, it alters the data structure, making it incompatible with the standard format in the field, the Measurement Set (MS) [48] format. The MS is the standardized format for storing radio interferometric data. Since BDA changes the time-frequency grid, baselines will have different numbers of time slots and channels after averaging, which disrupts the uniformity required by the MS format. Additionally, data averaged using BDA cannot be easily processed with standard tools.

### 1.3.3 Singular Value Decomposition

While BDA achieves high compression rates with minimal smearing effects by leveraging the baseline distribution, it has drawbacks that make it less favourable. Alternative approaches have been proposed by Rajeswaran and Winberg [56], who investigated lossless compression algorithms on SKAO data, such as SZIP, GZIP, the LZ4 filter, LZ4, and the Fully Adaptive Prediction Error Coder (FAPEC). Masui et al. [57] explored a lossless compression method that optimized data precision by discarding noisy bits. Offringa [58] introduced Dynamical Statistical Compression (Dysco), a lossy compression technique specifically designed for complex, noise-dominated interferometric visibility data. This method first normalizes groups of visibilities to have similar statistical distributions and then applies non-linear quantization.

A promising alternative to BDA for data compression was introduced by Atemkeng et al. [8], who proposed using SVD to compress visibility data. SVD is a mathematical technique that decomposes a matrix into three components. The SVD of an  $M \times N$  complex matrix  $V$  is expressed as [59, 60]:

$$V = U\Lambda W^\dagger, \quad (1.42)$$

where  $\dagger$  represents the complex transpose operator. In this decomposition,  $U$  is an  $M \times N$  complex unitary matrix, and  $W$  is an  $N \times N$  complex unitary matrix, orthonormal to  $U$ .  $\Lambda$  is an  $N \times N$  diagonal matrix, with its diagonal entries  $\eta_i = \Lambda_{ii}$ , which are non-negative real numbers known as singular values. These singular values are the square roots of the eigenvalues of  $V^\dagger V$  or  $VV^\dagger$ , and they are ordered in descending magnitude. The rank of the complex matrix  $V$  is determined by the number of non-zero singular values. The entries of  $U$ , denoted by  $u_1, \dots, u_M$ , are the eigenvectors of  $VV^\dagger$  and are referred to as left singular vectors. Similarly, the entries of  $W$ , denoted by  $w_1, \dots, w_N$ , are the eigenvectors of  $V^\dagger V$ , known as the right singular vectors. This decomposition can also be expressed as:

$$V = \sum_{i=0}^r \eta_i u_i w_i^\dagger, \quad (1.43)$$

where  $r = \min(M, N)$  is the rank of  $V$ , representing the number of singular values. For a baseline  $pq$ , with  $M$  time slots and  $N$  frequency channels, the visibility data matrix can be represented as:

$$V_{pq} = \begin{bmatrix} V_{pq\,t_1v_1} & \cdots & V_{pq\,t_1v_N} \\ \vdots & \ddots & \vdots \\ V_{pq\,t_Mv_1} & \cdots & V_{pq\,t_Mv_N} \end{bmatrix}$$

Atemkeng et al. proposed utilizing SVD to find a low-rank approximation of  $V_{pq}$  without significant information loss. However, they acknowledged that this approach is challenging due to the inherent correlations between the entries of  $V_{pq}$ , which prevent perfect reconstruction through a low-rank approximation. The low-rank approximation of  $V_{pq}$  is given by:

$$V_{pq,n} = \sum_{i=0}^n \eta_{pqi} u_{pqi} w_{pqi}^\dagger, \quad (1.44)$$

where  $n \leq r$  is the reduced rank. The contribution of each singular value  $\eta_i$  to the reconstruction of  $V_{pq}$  is proportional to its magnitude. Larger singular values capture most of the features of the original data, while smaller singular values primarily capture noise and can be discarded without losing significant information.

Atemkeng et al. observed that the first singular value, being the largest, captures nearly all the features of the data at the phase centre. However, as the distance from the phase centre increases, the singular values decay more rapidly on shorter baselines and more slowly on longer baselines. This suggests that a greater number of singular values are required to accurately represent the data away from the phase centre on longer baselines.

This observation led to the combination of low-rank approximation with BDA formalism. By using a low-rank approximation, where the same number of singular values are retained across all baselines (referred to as simple SVD), and combining this with BDA (referred to as Baseline-Dependent Singular Value Decomposition (BDSVD)), where the number of retained singular values,  $n_{pq}$ , depends on the baseline, the authors were able to achieve high data compression rates. This approach not only discarded noise but also minimized smearing in the Field of View (FoV) and outperformed both simple averaging and BDA.

The choice of domain, whether visibility or image, does not affect the number of singular values required for compression, as the singular values remain consistent in both domains, yielding the same result. In the image domain, the compressed dirty image is represented by:

$$\begin{aligned}\hat{I}_{pq}^d &= \mathcal{F}^{-1}\{V_{pq,n}\} \\ &= \sum_{i=0}^n \eta_{pqi} I_{pqi}^d,\end{aligned}\tag{1.45}$$

where  $\eta_{pqi} I_{pqi}^d$  corresponds to the component of the sky observed by baseline  $pq$ , with the singular value  $\eta_{pqi}$  quantifying the contribution strength of  $\eta_{pqi} I_{pqi}^d$  to the image reconstruction. For BDSVD, the reduced rank  $n_{pq}$  is baseline-dependent. The component  $I_{pqi}^d$  can be expressed as:

$$I_{pqi}^d = u'_{pqi} w'^{\dagger}_{pqi},\tag{1.46}$$

in accordance with Parseval's theorem. The full dirty image is obtained by summing over all baselines:

$$\begin{aligned}
\hat{I}^d &= \sum_{pq} \hat{I}_{pq}^d \\
&= \sum_{pq} \left( \sum_{i=1}^n \eta_{pqi} I_{pqi}^d \right) \\
&= \sum_{pq} \left( \sum_{i=1}^n \eta_{pqi} u'_{pqi} w'^{\dagger}_{pqi} \right).
\end{aligned} \tag{1.47}$$

In this thesis, we will focus exclusively on compressing the data in the visibility plane.

SVD has been widely applied across various fields for signal processing, image restoration, and data compression [61]. In radio astronomy, Offringa et al. [62] explored using SVD to remove RFI from observational data. Additionally, Briggs [63] explored the use of SVD to decompose the PSF system matrix, truncating small singular values that amplify noise. This stabilizes the inversion process, enabling a robust deconvolution that results in high-quality images.

Unlike BDA, BDSVD does not modify the time-frequency grid structure of the data, ensuring compatibility with the standard MS format. Furthermore, the data can be processed further using standard tools, making BDSVD a desirable compression method. It achieves high compression rates with minimal smearing effects. However, BDSVD is significantly more computationally expensive than both simple averaging and BDA. While simple averaging and BDA scale as  $\mathcal{O}(N_{vis})$ , where  $N_{vis}$  is the number of visibilities, BDSVD scales as  $\mathcal{O}(N_{vis}r)$ , where  $r$  is the rank of the data. To effectively handle this computational complexity, BDSVD needs to be implemented in a parallel fashion.

We have examined traditional averaging techniques used to reduce visibility data volumes in radio interferometry and briefly introduced SVD-based compression techniques, which show promise for transforming visibility data compression. In the subsequent sections and chapters, we will delve deeper into the effectiveness of these techniques, focusing on their ability to preserve data fidelity and image quality during compression.

## 1.4 Problem Statement

The volume of visibility data in radio interferometry increases dramatically with high temporal and spectral resolution, as well as the growing number of baselines used in modern arrays. This rapid data growth results in high storage and processing costs, particularly for large-scale radio interferometers like the MeerKAT, the Australian Square Kilometre Array Pathfinder (ASKAP), the Low Frequency Array (LOFAR), the Murchison Widefield Array (MWA), and the upcoming Square Kilometre Array Observatory (SKAO), which generate unprecedented data volumes.

To address these challenges, averaging techniques are commonly used to compress data and reduce its volume. However, these methods result in signal loss, causing smeared or blurred source emissions and reduced source amplitudes in radio interferometric images. Furthermore, the extent of this smearing is baseline-dependent, as the signal phase depends on baseline length. Specifically, longer baselines are more affected than shorter ones. Although BDA addresses this issue by applying variable averaging intervals, it modifies the time-frequency grid structure, making the data incompatible with the standard MS format used in radio interferometry.

A promising alternative is the SVD-based data compression method proposed by Atemkeng et al [8]. This approach exploits the inherent structure of visibility data to achieve high compression rates while preserving temporal and spectral resolution. However, further exploration is needed to understand its effectiveness and limitations across various observational scenarios.

## 1.5 Research Aims

This thesis builds on the work of Atemkeng et al. [8] by evaluating the effectiveness of SVD-based compression in reducing the volume of visibility data while maintaining data fidelity and image quality. The primary aim is to investigate the Baseline-Dependent Singular Value Decomposition (BDSVD) compression method across a range of observational scenarios, with a focus on understanding how it is influenced by different sky distributions. This will provide insights into the potential and limitations of BDSVD as a viable solution for managing the growing data volumes in radio interferometry.



Although the analysis is centred on the MeerKAT telescope, the methodologies and findings can be adapted for other modern and future interferometers.

## Chapter 2

# Research Methodology

This research explores the feasibility of using SVD to compress radio interferometric data to address the challenge of large data volumes generated by modern radio interferometers. The primary goal is to evaluate the impact of compression on data fidelity, particularly in the image plane. For this study, we utilized simulated data, with all simulations conducted using tools provided by the `stimela` [64] framework.

### 2.1 Simulations

We simulated the visibility data using `simms` [65] tool. `simms` uses the `casatasks` [66] library to generate synthetic visibility data. The generated data are output as CASA Tables called a Measurement Set (MS) [48], a standardized format for storing radio interferometric data. The simulated MS represents an observation conducted with the MeerKAT array. This observation spans a total duration of 2.5 h, with an integration time of 1 s. The observing frequency is set at 1.4 GHz, with a channel width of 50 kHz, a total bandwidth of 1.2 MHz, and 24 channels.

At the observing frequency of 1.4 GHz, the FoV of the MeerKAT array, defined by the primary beam's full width at half maximum (FWHM), extends to a radial distance of approximately 0.55 deg. The first and second sidelobes at this frequency are located at radial distances of around 1.32 deg and 2.25 deg, respectively.

Following the simulation of the MS, the visibility data were populated with sky models using the `meqtrees` [67] tool. The simulations began with a sky model containing a 1 Jy point source at the phase centre. Additional simulations were performed for a point source off-centre around the second sidelobe. The study then

progressed to more realistic sky models, incorporating both point and extended (Gaussian) sources with varying brightness levels.

The above sky distributions were simulated at a declination of  $-30$  deg. Additional simulations were performed for a 1 Jy off-centre point source at declinations of 0 deg,  $-60$  deg, and  $-90$  deg.

To further evaluate the effectiveness of SVD across diverse observational scenarios, low Signal to Noise Ratio (SNR) sky distributions were simulated for an off-centre point source. Importantly, all simulations assumed perfectly calibrated data with no flagged data, and included thermal noise computed using an System Equivalent Flux Density (SEFD) value of 442 Jy.

## 2.2 Data Compression

Following the simulation of visibility data for various sky models, the data matrix  $V_{pq}$ , for each baseline  $pq$ , with rank  $r$ , is compressed through the following steps:

- Rank determination: Identify the rank  $r$  of the data matrix  $V_{pq}$ .
- Decomposition: Decompose the data matrix  $V_{pq}$  into the matrices  $U$ ,  $\Lambda$ , and  $W^\dagger$  using the `dask.array.linalg.svd()` function. This function is well-suited for handling large datasets, and efficiently managing memory and computation.
- Truncation: To achieve lossy compression, truncate the matrices  $U$ ,  $\Lambda$ , and  $W^\dagger$  by selecting the desired number of singular values,  $n$ , for the reconstruction of  $V_{pq}$ . Singular values beyond  $n$  are effectively set to zero, eliminating unwanted data components. The truncated matrices represent the compressed data.
- Reconstruction: Using the truncated matrices, reconstruct the visibility matrix  $V_{pq}$  as a low-rank approximation  $V_{pq,n}$ .

The second aspect of our analysis is the study of individual singular values. In this case, only one singular value  $\Lambda_{ii}$  is selected, and the low-rank approximation is given by:

$$V_{pq,i} = \eta_{pqi} u_{pqi} w_{pqi}^\dagger \quad (2.1)$$

In these simulations, only the Stokes  $I$  term was used, meaning that we have the correlations  $XX = YY = I/2$ .

## 2.3 Statistical Analysis of Residuals

Statistical analyses were performed in the visibility plane on the residual data, which is defined as the difference between the high and low-rank data. Metrics such as mean, standard deviation, skewness, kurtosis, and the  $\chi^2$  statistic were computed. The  $\chi^2$  was calculated between the residual data and the noise only data. These statistics were computed using the `scipy.stats` library.

The  $\chi^2$  statistic was normalized by the number of visibilities:

$$\chi_{nvis}^2 = \frac{\chi^2}{n_b \times M \times N \times n_{pol}}, \quad (2.2)$$

where  $n_b$  is the number of baselines,  $M$  is the number of time slots,  $N$  is the number of frequency channels, and  $n_{pol}$  is the number of correlations.

## 2.4 Imaging

For imaging the visibility data, we used the `wsclean` [68] software. Natural weighting was applied to minimize biases that might arise from visibility weights. Averaging modifies the Point Spread Function (PSF), and following Briggs' [63] suggestion that SVD can be used to decompose the Point Spread Function (PSF) matrix and truncate small singular values that amplify noise, thereby enabling robust deconvolution that improves image quality, we chose to work with dirty images. This approach allowed us to evaluate whether compressing the data using SVD could effectively perform deconvolution and reduce sidelobe effects.

## 2.5 Compression Factor

The size of the original data matrix  $V_{pq}$  is given by  $M \times N$ . The compressed stored data is the truncated matrices which collectively have a size of  $M \times n + n \times n + n \times N = n(M + N + 1)$ . The compression factor, defined as the ratio between the number of data points of the compressed data (low-rank) to the uncompressed data (high-rank), is then computed as:

$$C_f = \frac{M \times N}{n(M + N + 1)}, \quad (2.3)$$

where  $n$  is the reduced rank. To account for the fact that the singular values are real while the left and right singular matrices aren't, a factor of 0.5 is used instead to give:

$$C_f = \frac{M \times N}{n(M + N + 0.5)}. \quad (2.4)$$

## Chapter 3

# Results and Discussion

In this chapter, we analyze the impact of various observational scenarios on Baseline-Dependent Singular Value Decomposition (BDSVD) compression using the MeerKAT telescope.

### 3.1 Individual Singular Values

This section examines the individual singular values in detail. The goal is to analyze how different sky distributions affect the features captured by these singular values. This analysis also provides insights into the characteristics of low-power singular values. For this study, the data has a rank of 24, corresponding to 24 singular values. This is because the data for each baseline consists of 24 frequency channels and spans  $2.5 \text{ h} \times 3600 \text{ s} = 9000$  time slots, resulting in a matrix with dimensions such that its rank is given by  $\min(M, N) = 24$ .

#### 3.1.1 Phase centre

We consider a sky distribution with a 1 Jy point source located at the phase centre. At the phase centre, no smearing effects occur, and the first singular value captures all the data features. Figure 3.1 shows the mean weights of the singular values for this scenario. Baselines are grouped by length: short ( $<1 \text{ km}$ ), medium ( $1 \text{ km} < \text{length} < 3 \text{ km}$ ), and long ( $>3 \text{ km}$ ). The mean weight for each group is computed as the average of the singular value weights, with error bars indicating standard deviations. This figure demonstrates that, for a 1 Jy point source at the phase centre, the first singular value dominates, with subsequent values being smaller

across all baselines. This confirms that the first singular value effectively encapsulates all data features, consistent with the negligible decoherence expected at the phase centre. Consequently, the data can be fully represented by the first singular value, allowing compression from rank 24 to rank 1, achieving a compression factor of approximately 50 (Table 3.1).

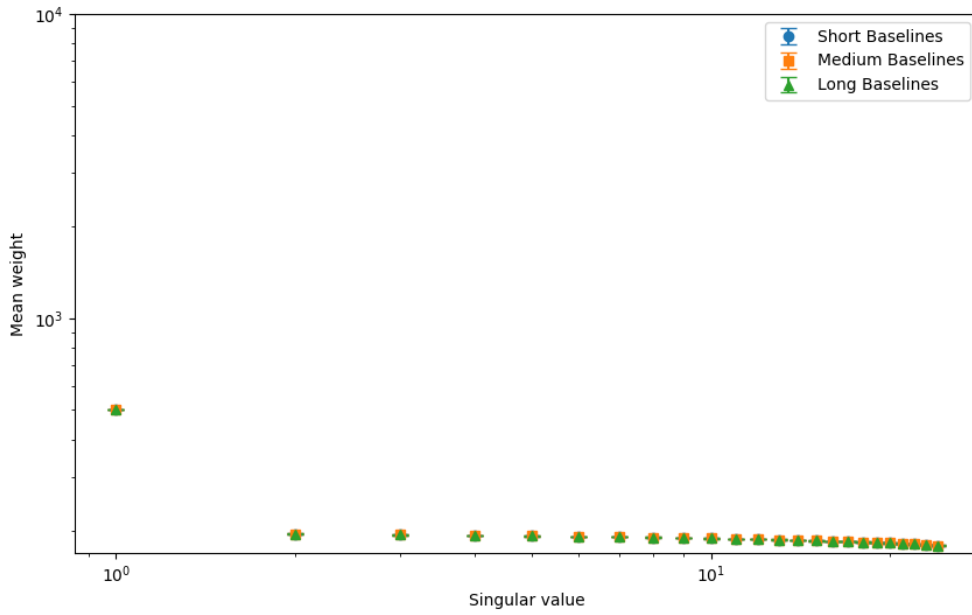


FIGURE 3.1: Singular values and their corresponding weights for a 1 Jy point source at the phase centre.

Figure 3.2 provides further insight into the data compression process using various singular values. The left panel compares the original data (full rank) with the data compressed using different singular values. The right panel shows the corresponding residuals, which highlight the differences between the original and compressed data.

The peak brightness in the original data image is 0.999 907 Jy. Using only the first singular value for compression successfully recovers  $0.999\,893 \pm 2.012\,62 \times 10^{-6}$  Jy, which corresponds to approximately 99.9986% of the original flux. The uncertainty of  $2.012\,62 \times 10^{-6}$  Jy represents the Root Mean Square (RMS) extracted from the residual of the first singular value. This result demonstrates the dominance of the

first singular value in capturing the features of the data. In contrast, subsequent singular values contribute negligibly to the representation of the data, with recovered peak brightness values falling below the RMS. These higher-order singular values primarily capture noise rather than meaningful signal components. Notably, the residual images from higher-order singular values reveal the source structure, unlike the residual image of the first singular value, which closely resembles random noise. This indicates that the first singular value effectively isolates the primary signal, leaving behind residuals that closely resemble random noise.

The accompanying statistics table in Figure 3.2 confirms this observation. The residual distribution for the first singular value is consistent with the noise distribution, indicating that its residuals are dominated by random fluctuations. Conversely, the residual distributions for higher-order singular values deviate from the noise distribution. Specifically, the mean, standard deviation and  $\chi^2_{nvis}$  values of these residuals are relatively larger, suggesting the presence of the signal components.

The mean and standard deviation values for the residual statistics in Figure 3.2 and subsequent figures appear notably large due to their computation in the visibility plane, where noise is uncorrelated. In the visibility plane, each visibility measurement contributes independent thermal noise:

$$\sigma_{vis} = \frac{SEFD}{\sqrt{2 * \Delta\nu \Delta t}}, \quad (3.1)$$

where  $SEFD$  is the System Equivalent Flux Density of the two antennas involved,  $\Delta\nu$  is the channel width, and  $\Delta t$  is the integration time. In this case, the antennas share an  $SEFD$  of 442 Jy, leading to a thermal noise level of approximately:

$$\sigma_{vis} = \frac{442 \text{ Jy}}{\sqrt{2 * 50 \text{ kHz} * 1 \text{ s}}} = 1.39773 \text{ Jy}, \quad (3.2)$$

for each baseline. Figure 3.3 illustrates the corresponding power profile of the noise in the visibility plane, revealing the significant scatter of power values across different baselines, leading to the observed large mean and standard deviation values.

In contrast, the image plane transformation changes the noise characteristics. Through the Fourier transform, the visibility noise becomes correlated. The noise is effectively averaged over the number of visibilities contributing to each pixel, scaling according to the relationship:



$$\sigma_{img} = \frac{\sigma_{vis}}{\sqrt{N_{vis}}}, \quad (3.3)$$

where  $N_{vis}$  represents the number of visibilities contributing to each image pixel [69]. This averaging substantially reduces the noise level in the image plane compared to the visibility plane.

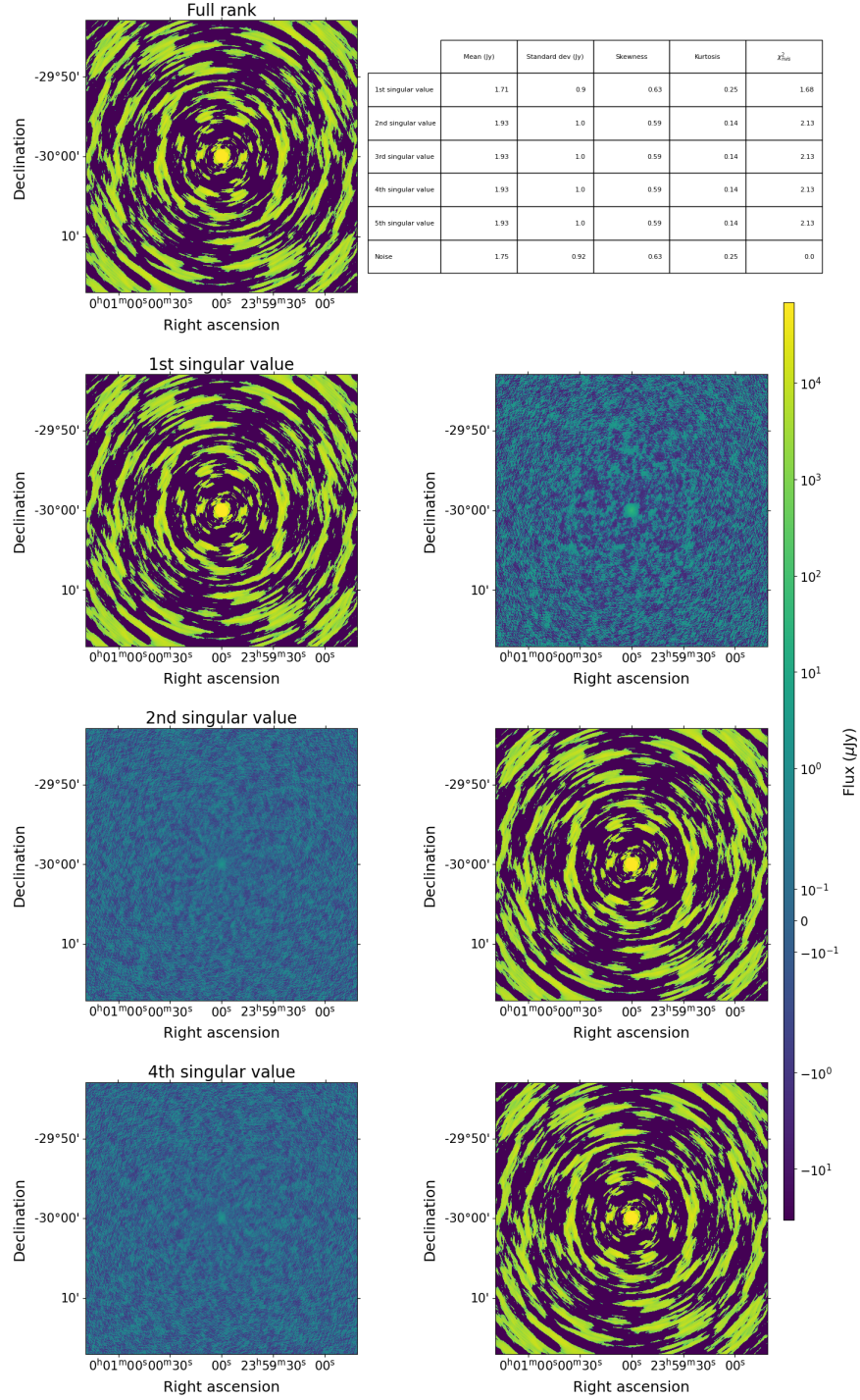


FIGURE 3.2: Images of the original data (left panel) for a point source at the phase centre alongside images of the data reconstructed using different singular values. The right panel shows the corresponding residuals, highlighting the differences between the original and compressed data. Additionally, the accompanying table contains the residual statistics.

In conclusion, for the sky distribution with a bright point source at the phase

centre, the first singular value alone is sufficient to fully represent the data across all baselines.

TABLE 3.1: Compression factors. The data sizes are given in units of GB

| No. Singular values | Compression factor | Data size (GB) |
|---------------------|--------------------|----------------|
| Full rank (24)      | 1.0                | 6.49           |
| 20                  | 2.4                | 2.70           |
| 15                  | 3.2                | 2.03           |
| 12                  | 3.98               | 1.63           |
| 9                   | 5.32               | 1.22           |
| 8                   | 5.98               | 1.09           |
| 7                   | 6.84               | 0.95           |
| 6                   | 7.98               | 0.81           |
| 5                   | 9.58               | 0.68           |
| 4                   | 11.96              | 0.54           |
| 3                   | 15.96              | 0.41           |
| 2                   | 23.94              | 0.27           |
| 1                   | 47.86              | 0.14           |

### 3.1.2 Away from the Phase Centre

For a sky distribution with a 1 Jy point source located around 2.25 deg off-centre, our findings are largely consistent with the previous case. Figure 3.4 demonstrates that the first singular value remains much larger than the subsequent singular values, indicating its dominance in capturing the data features. In contrast, the subsequent singular values primarily capture the noise features.

Figure 3.5 confirms this observation. When compressing the data using only the first singular value,  $0.957091 \pm 1.83464 \times 10^{-5}$  Jy is recovered, corresponding to approximately 99.8966% of the peak brightness, compared to 0.958082 Jy in the original data image. This demonstrates the first singular value's ability to capture most of the data features while leaving behind noise features. However, compared to the previous case (point source at the phase centre), the second singular value captures some meaningful components here. For example, the second singular value recovers approximately  $1.61327 \times 10^{-4} \pm 1.83464 \times 10^{-5}$  Jy (which is around

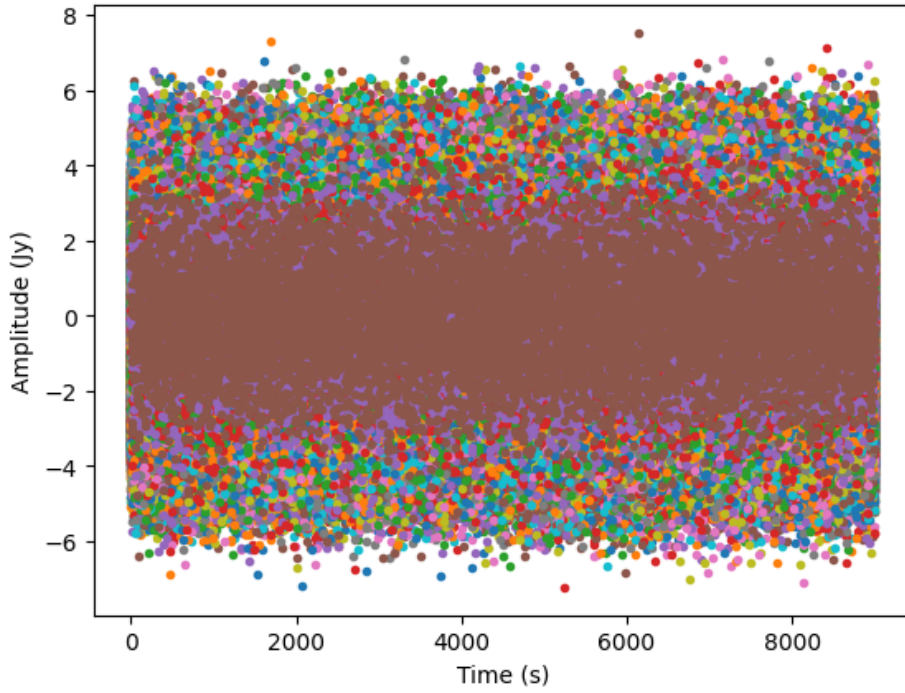


FIGURE 3.3: Noise power profile in the visibility plane for different baselines, given by different colours.

$1.68386 \times 10^{-2}\%$ ) of the peak brightness. This contribution, though small, is significant compared to the negligible contributions from subsequent singular values, whose recovered peak brightness falls below the RMS value. As such, these higher-order singular values do not capture any meaningful components but instead represent noise features. The second singular value captures meaningful components in this case due to the decoherence experienced away from the phase centre.

The accompanying statistics table in Figure 3.5 reveals further insights. The residual distribution for the first singular value is consistent with the noise distribution, confirming its noise-like nature. Conversely, the residual distributions for higher-order singular values deviate from the noise distribution, indicating the presence of the signal components in these residuals. Specifically, their means, standard deviations, and  $\chi^2_{nvis}$  values are relatively larger, all of which suggest the presence of the original signal components.

Thus, the first singular value captures the majority of the data features across all

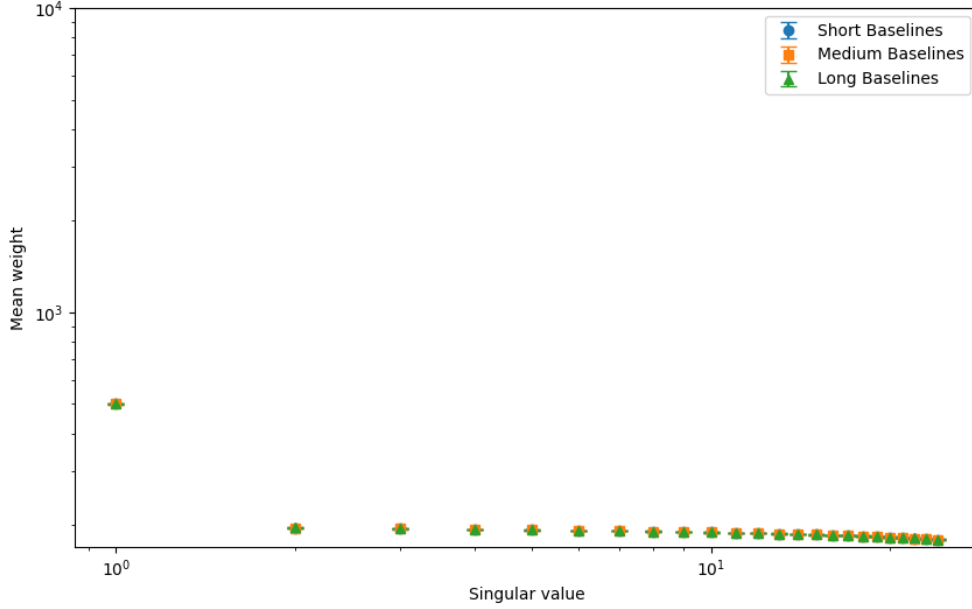


FIGURE 3.4: Singular values and their corresponding weights for a 1 Jy point source 2.25 deg off-centre.

baselines, capturing the primary signal while leaving noise in its residuals. Higher-order singular values primarily capture noise, as evidenced by their residuals revealing the original data features. Consequently, the data can be effectively compressed using only the first singular value, reducing the rank from 24 to 1 and achieving a compression factor of approximately 50 on short and medium baselines. However, on long baselines, the inclusion of the first two singular values is necessary to eliminate the 0.1% smearing, resulting in a compression factor of around 24.



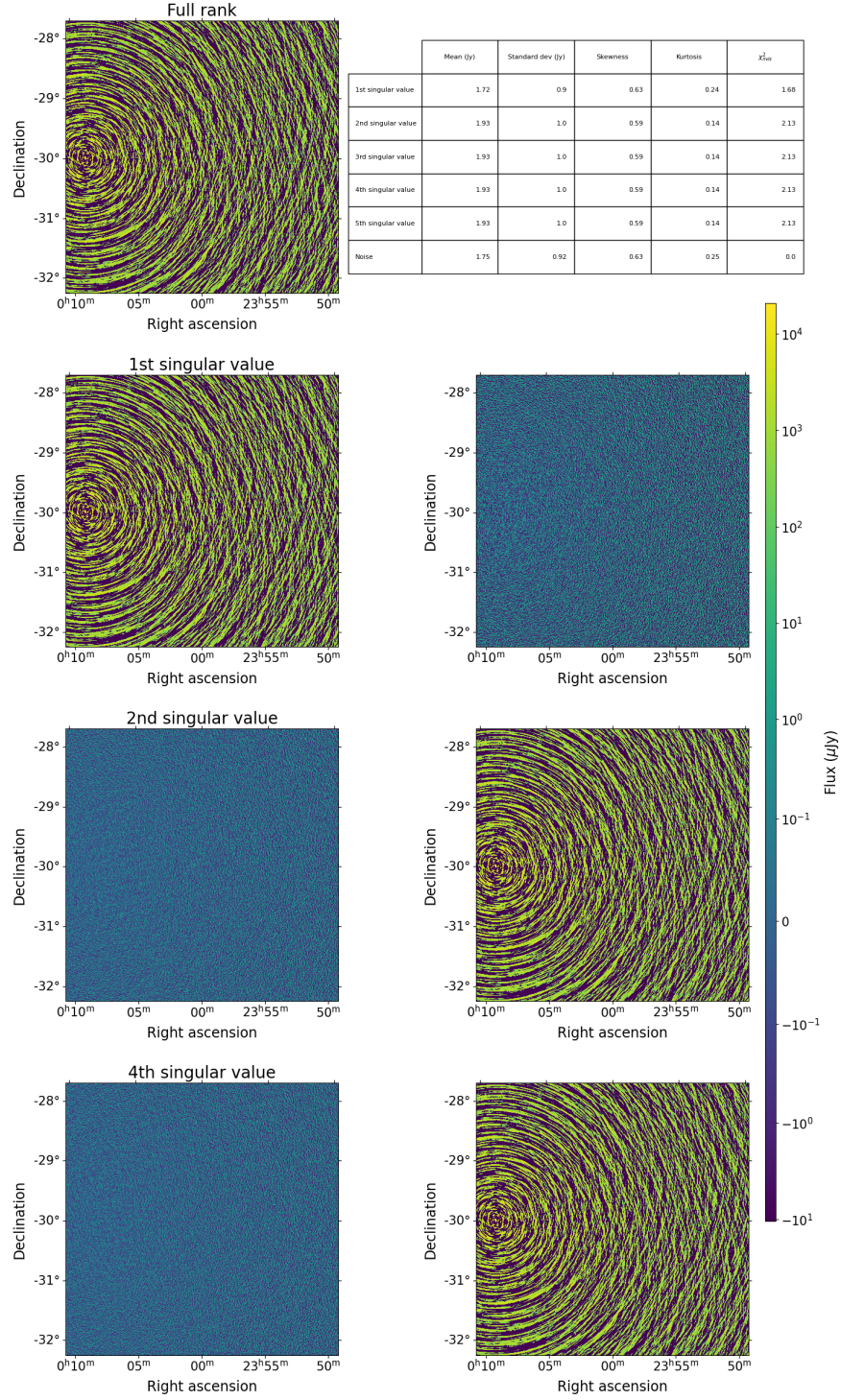


FIGURE 3.5: The left panel presents the original data images along with the compressed data images for the sky distribution with a point source 2.25 deg off-centre. The right panel presents the residual statistics table along with the residual images.

### 3.1.3 Point Sources

We consider a sky distribution consisting of multiple point sources with varying brightness levels. In this scenario, the first singular value remains dominant, capturing most of the data features. However, the second singular value also plays a significant role, particularly on medium and long baselines. Figure 3.6 illustrates that the singular value distribution on short baselines resembles that of the previous cases. On longer baselines, however, the second singular value exhibits a larger weight due to its contribution in capturing additional data features.

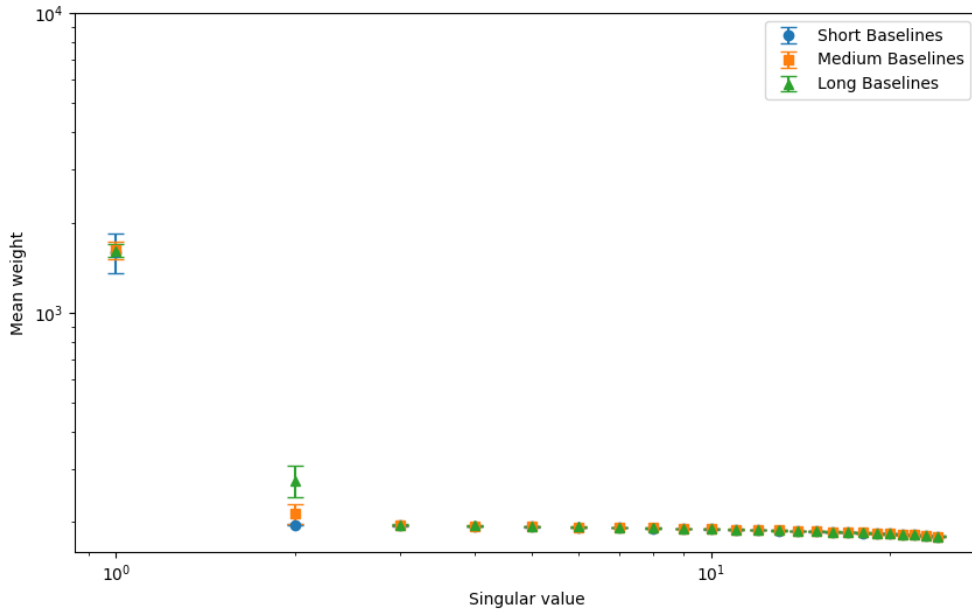


FIGURE 3.6: Singular values and their corresponding weights for the sky distribution with point sources.

When compressing the data using only the first singular value across all baselines, the recovered peak brightness is  $8.31269 \times 10^{-1} \pm 8.81299 \times 10^{-5} \text{ Jy}$ , corresponding to approximately 98.3540% of the original peak brightness ( $8.45181 \times 10^{-1} \text{ Jy}$ ) for a source at the edge of the Field of View (FoV), which is 1.1 deg for the MeerKAT telescope at 1.4 GHz. This reduction in recovered peak brightness is attributed to the inability of the first singular value to fully capture features away from the phase centre, which are instead captured by the second singular value.

Figure 3.7 demonstrates that the residual image of the compressed data using only the first singular value no longer consists solely of noise, as observed in previous cases. Instead, it includes features not captured by the first singular value, particularly near the edges of the field. These features are captured by the second singular value, which recovers approximately  $1.39485 \times 10^{-2} \pm 8.81299 \times 10^{-5}$  (around 1.65036%) of the peak brightness. Near the phase centre, the first singular value successfully captures all data features. However, its effectiveness diminishes toward the edges of the field due to the expected decoherence away from the phase centre. Singular values beyond the second contribute minimally to the data features, recovering peak brightness below the RMS value. These higher-order singular values primarily capture noise features rather than signal components.

The accompanying statistics table in Figure 3.7 confirms these observations. The residual distribution for the data compressed using the first singular value is consistent with the noise distribution, indicating that its residuals are dominated by random fluctuations. In contrast, the residual distributions for higher-order singular values deviate from the noise distribution. Their means, standard deviations, and  $\chi^2_{\text{vis}}$  are relatively larger compared to the noise distribution, supporting the presence of original data features in these residuals.

Thus, on short baselines, the first singular value fully captures all data features, enabling compression to rank 1 with a factor of approximately 50. However, on medium and long baselines, the second singular value is essential for accounting for features away from the phase centre. When the data is compressed using only the first singular value across all baselines, this results in around 2% smearing and a recovery of approximately 98% of the peak brightness for sources at the edge of the field. In this case, the data can be fully represented by using the first two singular values on medium and long baselines, achieving a compression factor of approximately 24.



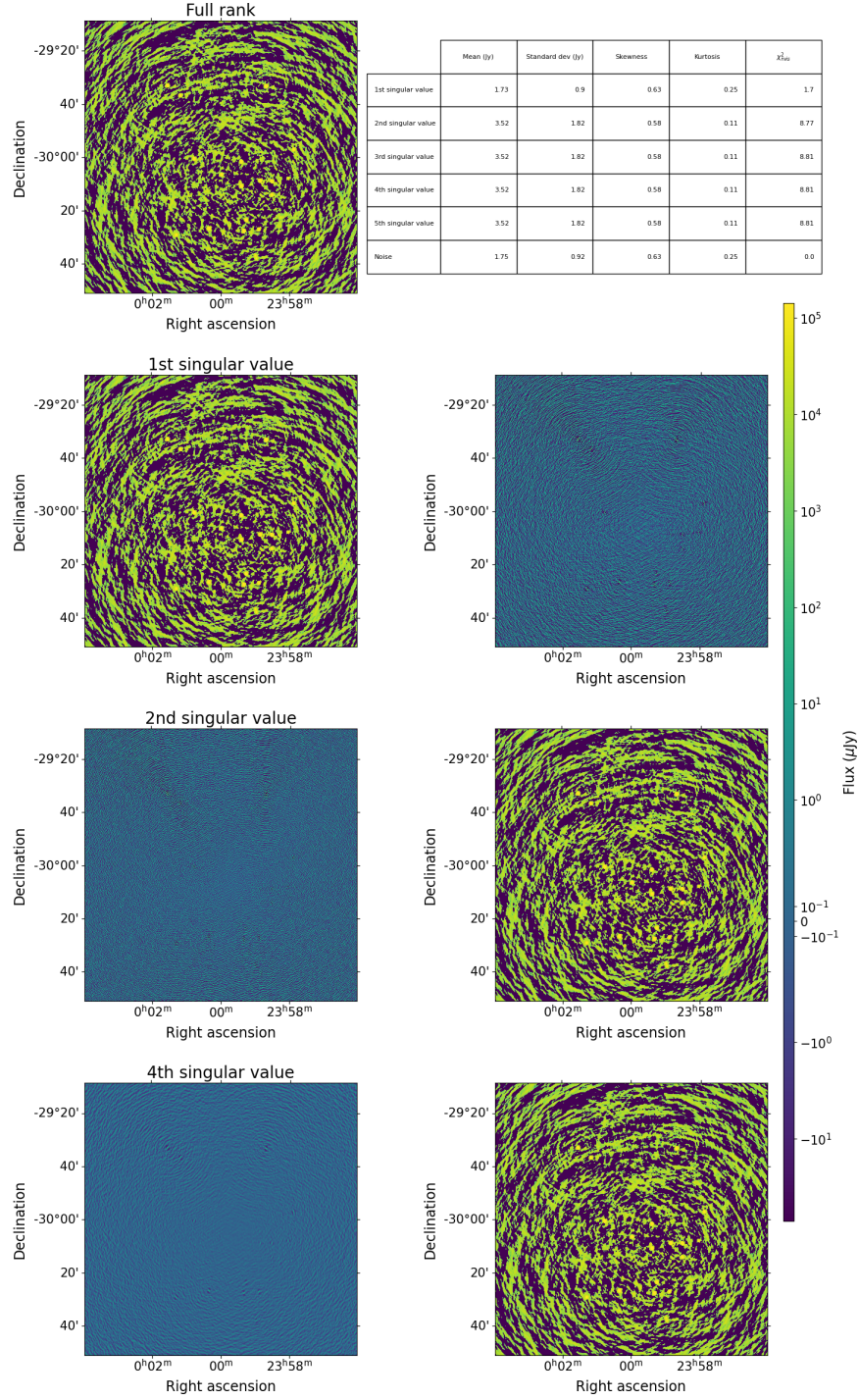


FIGURE 3.7: The left panel presents the original data images along with the compressed data images for the sky distribution with multiple point sources. The right panel presents the residual statistics table along with the residual images.

### 3.1.4 Point and Extended Sources

In this scenario, we analyze a sky distribution comprising both point and extended sources with varying brightness levels. Similar to the previous case, the first singular value dominates on short baselines, with subsequent singular values being smaller, as shown in Figure 3.8. On medium and long baselines, however, the first singular value carries less weight compared to short baselines, while the second singular value exhibits a larger weight. This indicates that the second singular value captures significant data features for these baselines. This behaviour arises because fringe rates change more rapidly on longer baselines, requiring frequent sampling and distributing the data features among multiple singular values.

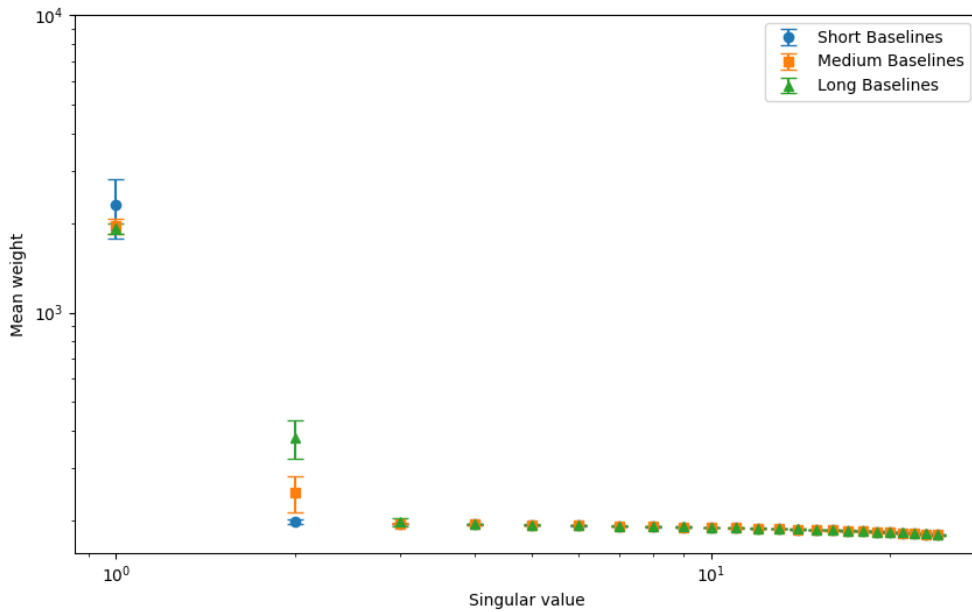


FIGURE 3.8: Singular values and their corresponding weights for the sky distribution with both point and extended sources.

For a source at the edge of the FoV, compressing the data using only the first singular value results in a recovered peak brightness of approximately  $2.80953 \times 10^{-1} \pm 1.66982 \times 10^{-4} \text{Jy}$ , corresponding to about 96.9696% of the original peak brightness ( $2.89733 \times 10^{-1} \text{Jy}$ ). This is significantly lower than in previous cases.

The residual image of the first singular value, shown in Figure 3.9, reveals features not captured by this singular value, particularly near the edges of the image. These features are instead captured by the second singular value, which recovers approximately  $8.66226 \times 10^{-3} \pm 1.66982 \times 10^{-4}$  Jy, or about 2.98974% of the original peak brightness, a substantially higher proportion compared to the previous case. Furthermore, the third singular value recovers approximately  $2.34975 \times 10^{-4} \pm 1.66982 \times 10^{-4}$  Jy, or around  $8.11007 \times 10^{-2}\%$ . The singular values beyond the third recover a peak brightness below the RMS, indicating that they do not capture any meaningful components but instead capture noise features.

The power profiles horizontally across the centre of the images in Figure 3.10 demonstrate that higher-order singular values capture features away from the phase centre, that are not captured by the first singular value.

The accompanying statistics table in Figure 3.9 demonstrates that the residual distribution for the first singular value is consistent with the noise distribution, reflecting the noise nature of the residual. For higher-order singular values, the significant deviations in mean and standard deviation from the noise distribution highlight the presence of original data features. This is supported by the slight right-skewness, heavy-tailedness, and relatively larger  $\chi^2_{nvis}$  values of the higher-order residual distributions.

These results demonstrate that while the first singular value effectively fully captures the data features on short baselines, it fails to do so on longer baselines, leading to approximately 3% smearing. On medium and long baselines, data features are distributed among additional singular values, with the second singular value capturing most of these features, especially at the edges of the field. Beyond this, higher-order singular values primarily capture noise, as evidenced by their residual images revealing the original data features.

Thus, while rank-1 compression is effective on short baselines, achieving a compression factor of approximately 50, more singular values are required on medium and long baselines to eliminate smearing. This reduces the achievable compression ratio on longer baselines.

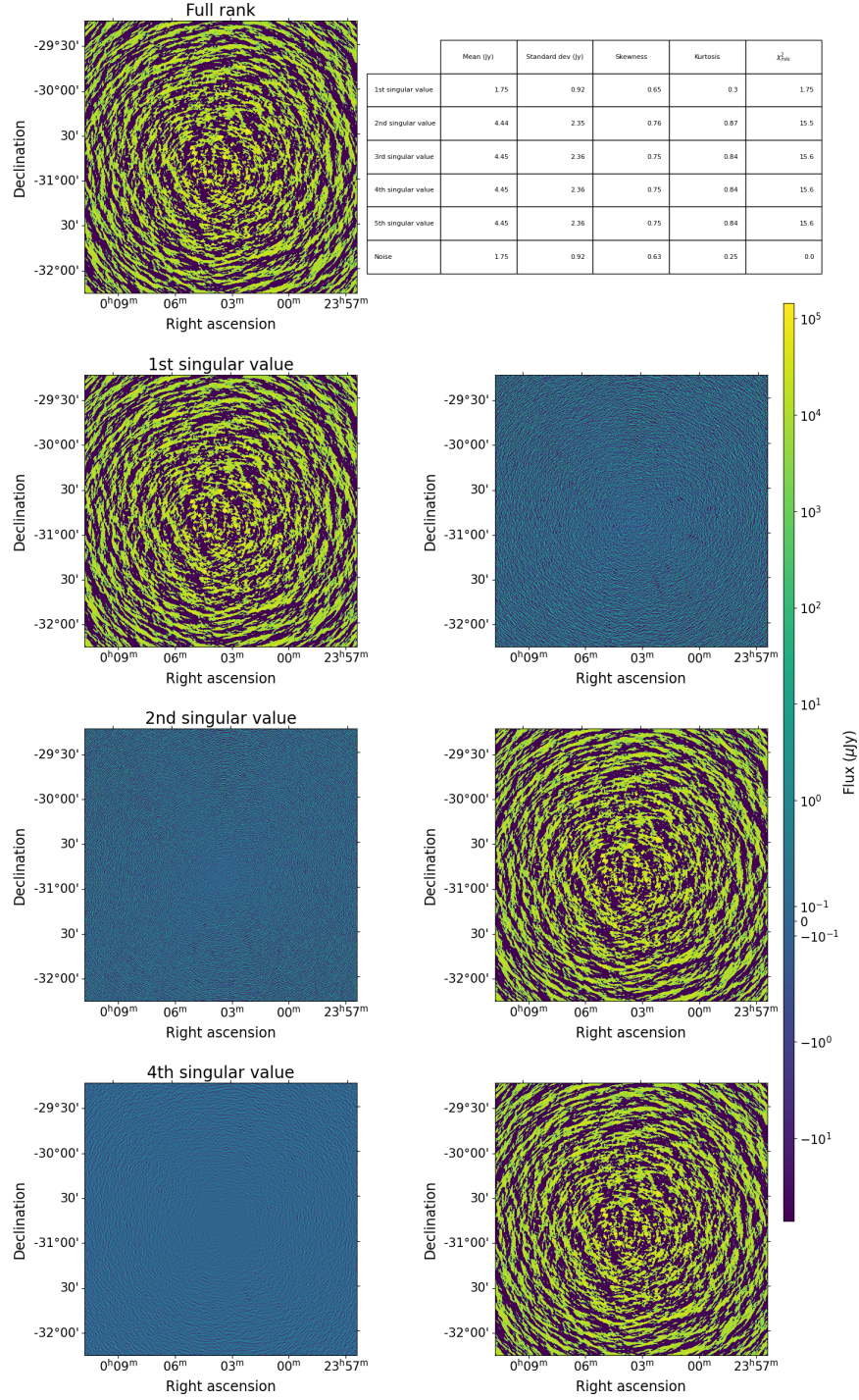


FIGURE 3.9: The left panel presents the original data along with the compressed data for the sky distribution with point and extended sources. The right panel presents the residual statistics table along with the residual images.

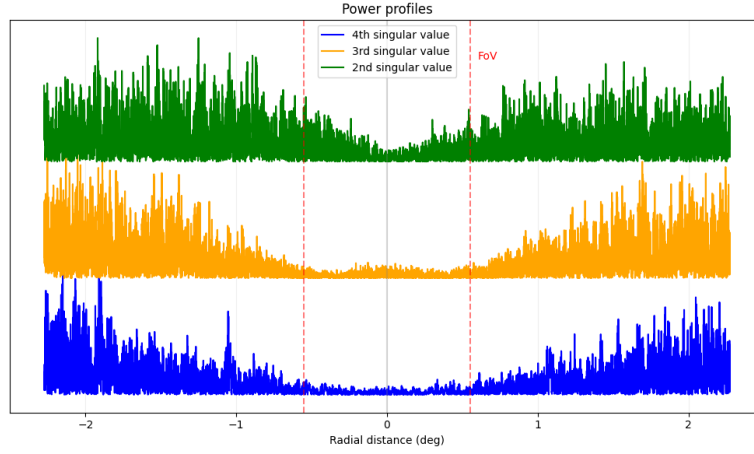


FIGURE 3.10: Power profiles horizontally across the images in Figure 3.9. The power profiles show that the higher-order singular values capture features away from the phase centre.

### 3.1.5 Signal to Noise Ratio (SNR)

In this case, we investigate the impact of SNR on BDSVD compression. For this analysis, we consider a sky distribution with a point source with SNR of around 100 (flux =  $100\sigma_{img}$ ), at a distance of 0.4 deg. This represents a low SNR. Additional SNR's are provided in Section 6. Figure 3.11, which displays the mean weights of the singular values, reveals a change in their distribution compared to previous cases. All singular values exhibit similar weights across all baselines, and the first singular value no longer dominates. This indicates that for a sky distribution with low SNR or a faint source, data features are evenly distributed among all singular values. Consequently, the first singular value does not capture all data features and cannot solely represent the data. Additional singular values are therefore required for accurate representation.

Figure 3.12 illustrates the original data image alongside the compressed data images and corresponding residuals images. When compressing the data using only the first singular value, approximately  $2.96172 \times 10^{-4} \pm 5.15278 \times 10^{-5}$  Jy of the peak brightness is recovered, corresponding to about 4.55803% of the original peak brightness ( $6.49781 \times 10^{-3}$  Jy). The second singular value recovers around



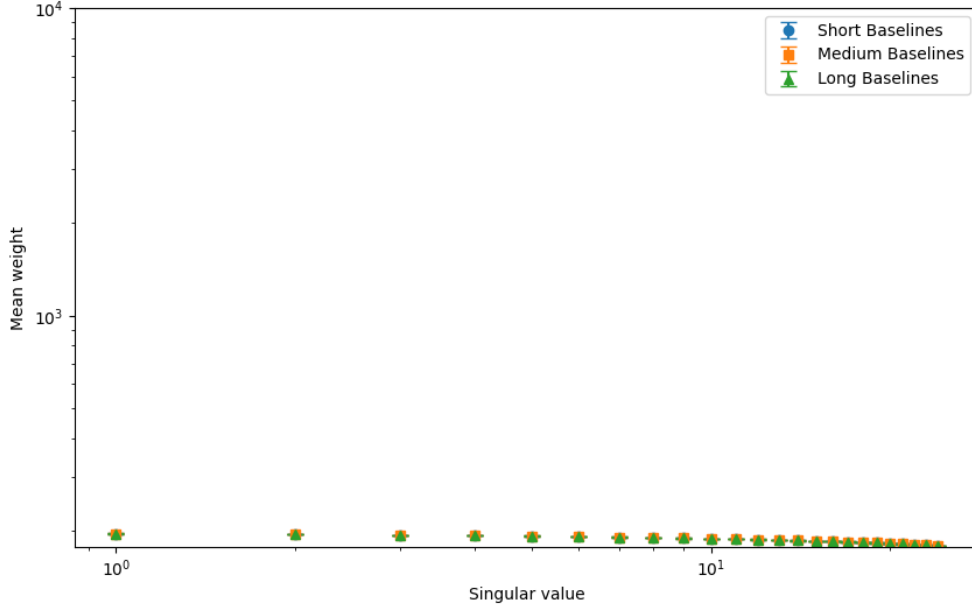


FIGURE 3.11: Singular values and their corresponding weights for the sky distribution with a point source with an SNR of 100.

$3.10928 \times 10^{-4} \pm 5.15278 \times 10^{-5}$  Jy, or about 4.78513% of the original peak brightness. These small and nearly equal contributions further suggest that the data features are evenly distributed across all singular values. For example, even the twentieth (20th) singular value (out of twenty-four) captures approximately  $2.42229 \times 10^{-4} \pm 5.15278 \times 10^{-5}$  Jy, or about 3.72785% of the original peak brightness. Each singular value contributes roughly 4% of the total peak brightness ( $100\%/24 \approx 4\%$ ).

Since the first singular value no longer dominates and recovers only about 4.55803% of the data, its residual contains signal components, as shown in Figure 3.12. This behaviour is consistent across all singular values, as they equally capture data features. The accompanying residual statistics table in Figure 3.12 provides further insights. The residual distributions for all singular values match, suggesting that they equally capture the features and therefore have matching residuals.

In conclusion, for a sky distribution consisting of a faint source, the first singular value alone is insufficient to represent the data. This is because the data features are evenly distributed among all the singular values, with each capturing a significant

amount of the data features. As a result, effective compression is not possible in this scenario.

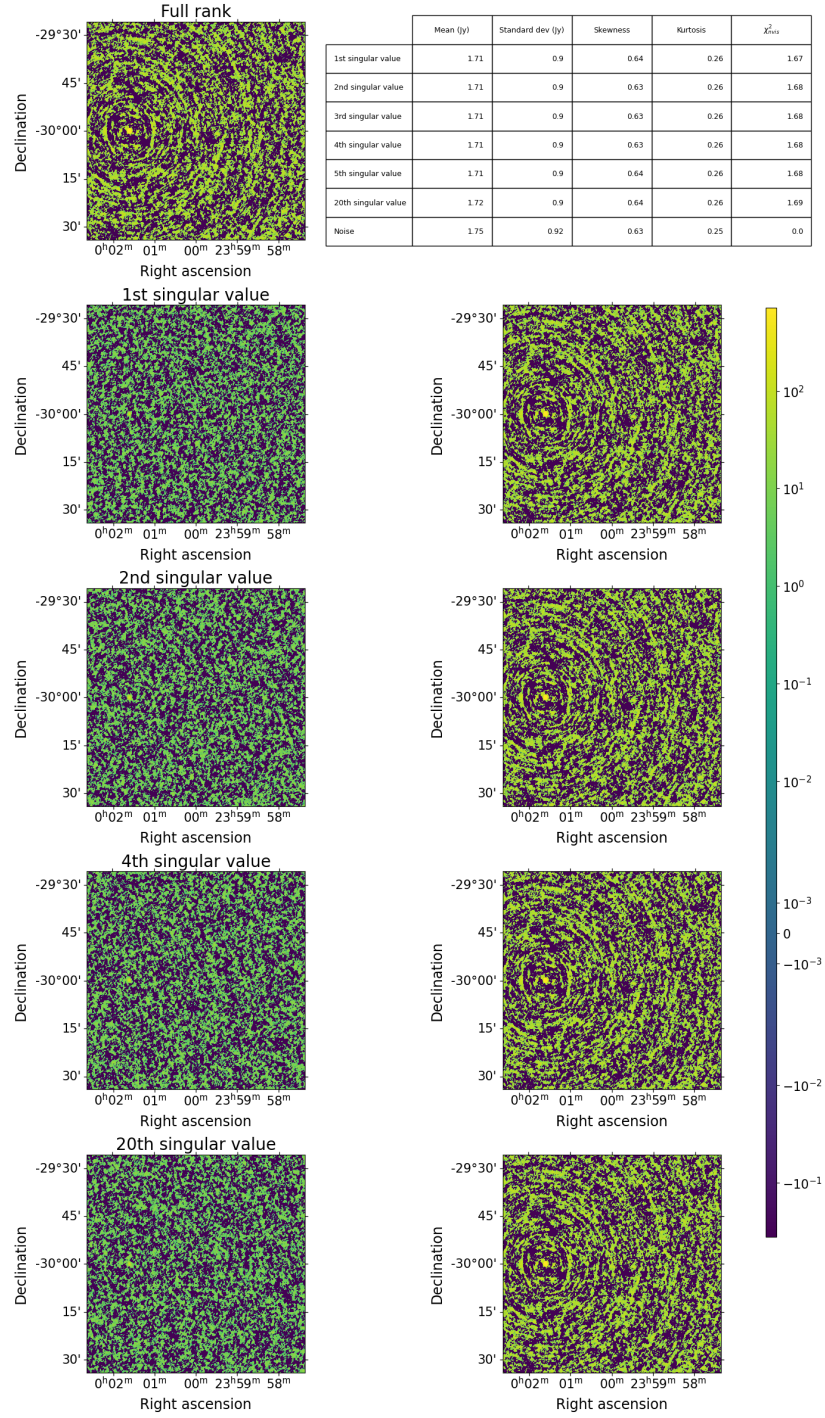


FIGURE 3.12: The left panel presents the original data images along with the compressed data images for the sky distribution with a point source 0.4 deg off-centre with an SNR of 100. The right panel presents the residual statistics table along with the residual images.

### 3.1.6 Point Sources with a total flux of 10 Jy

We consider a sky model consisting of 100 point sources with varying brightness levels, where the lowest SNR is approximately 5 ( $\text{flux} = 5\sigma_{\text{img}}$ ), the highest SNR is around 10,000 ( $\text{flux} = 10000\sigma_{\text{img}}$ ), and the total flux is 10 Jy. The data is simulated for a 1.25 h observation with an integration time of 4 s and a channel width of 1 MHz. Figure 3.13 presents the mean weights of the singular value for this distribution. From the singular value distribution, it is evident that on short baselines, the first four singular values dominate and capture most of the features. On medium baselines, the first eight singular values dominate, while subsequent singular values contribute significantly less. On long baselines, approximately the first 12 singular values (half of the total) dominate and capture meaningful features

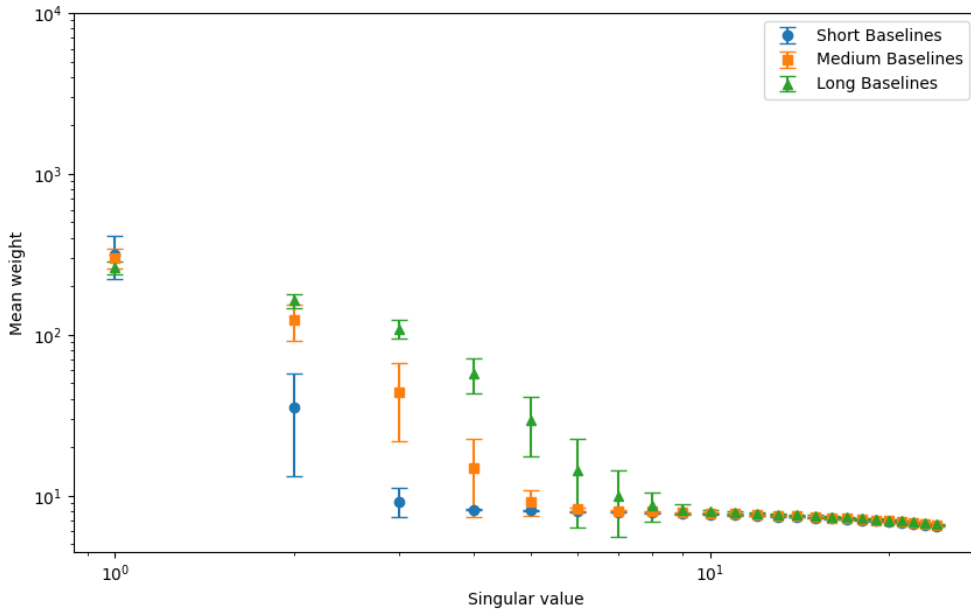


FIGURE 3.13: Singular values and their corresponding weights for the sky distribution with 100 point sources with a total flux of 10 Jy.

The first singular value captures the majority of the features but is insufficient on its own. For a source in Figure 3.14 at the edge of the FoV, it successfully recovers approximately  $6.78002 \times 10^{-2} \pm 9.95074 \times 10^{-4}$  Jy, corresponding to about 54.4263% of the original peak brightness ( $1.24573 \times 10^{-1}$  Jy). The second singular



value contributes significantly, recovering approximately  $2.08976 \times 10^{-2} \pm 9.95074 \times 10^{-4}$  Jy, or about 16.7755% of the original peak brightness. The third singular value adds further contribution, recovering around  $1.1442 \times 10^{-2} \pm 9.95074 \times 10^{-4}$  Jy, or approximately 9.185%. Additionally, the eighth singular value recovers about  $3.72390 \times 10^{-3} \pm 9.95074 \times 10^{-4}$  Jy, corresponding to roughly 2.98934%. Beyond the twelfth singular value, the recovered peak brightness falls below the RMS value, indicating that these singular values primarily capture noise features.

The accompanying statistics table in Figure 3.14 shows that the residual distribution for the first singular value closely aligns with the noise distribution, as it captures the majority of the data features and leaves primarily noise in its residuals. In contrast, higher-order singular values exhibit significant deviations in mean and standard deviation from the noise distribution, indicating the presence of predominantly original data features. This is further supported by the large  $\chi_{nvis}^2$  values associated with the higher-order residual distributions.

Thus, for this sky distribution, data features are distributed among many singular values across all baseline lengths. The first singular value alone is insufficient to fully represent the data, as it captures only about 50% of the features. Additional singular values are necessary to achieve a complete representation of the data.

### 3.1.7 Source Direction

Thus far, we have explored sky distributions with sources positioned at a declination of  $-30$  deg, which corresponds to the MeerKAT telescope's phase centre. In this case, we shift our focus to a 1 Jy point source positioned around 0.4 deg away from the phase centre at different declinations. This allows us to investigate how the source or field direction impacts BDSVD compression. Here, we focus on a source at a declination of  $-90$  deg, with additional declinations provided in Section 6.

Figure 3.15 displays the mean weights of the singular values for different baselines. The singular value distribution is consistent with the patterns observed in the first two cases; the first singular value dominates, being significantly larger than subsequent singular values across all baseline lengths. This suggests that the first singular value effectively captures all data features while subsequent singular values primarily capture noise features.

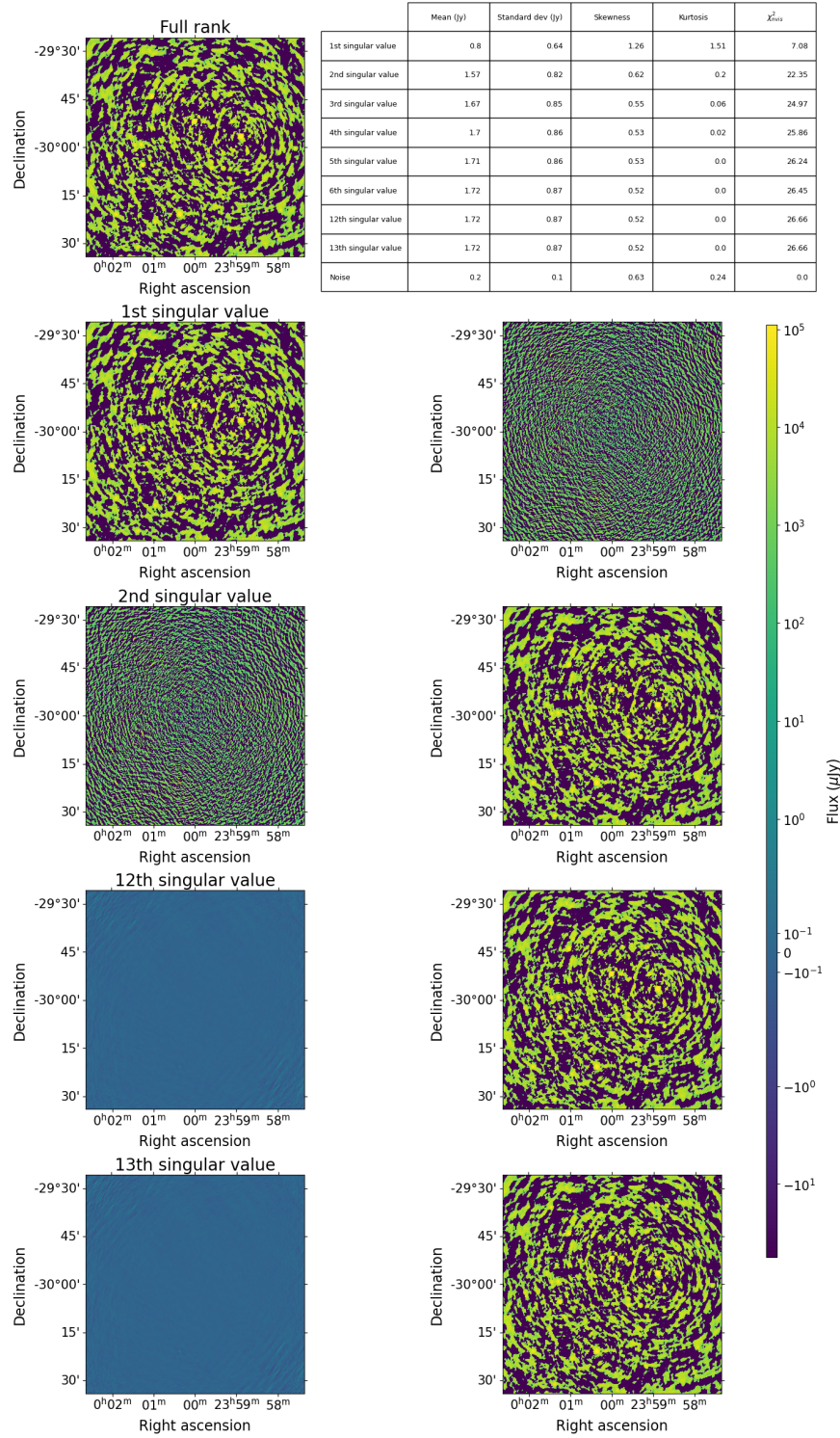


FIGURE 3.14: The left panel presents the original data along with the compressed data for the sky distribution with 100 point sources with a total flux of 10 Jy. The right panel presents the residual statistics table along with the residual images.

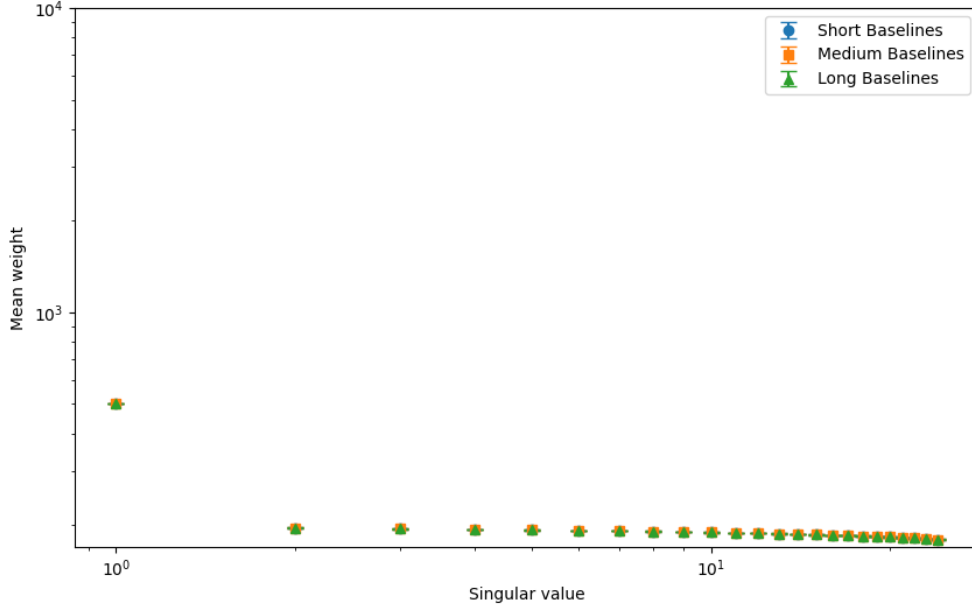


FIGURE 3.15: Singular values and their corresponding weights for the sky distribution with a point source at a declination of  $-90$  deg.

This is confirmed in Figure 3.16. When compressing the data to only the first singular value, the recovered peak brightness is  $9.99977 \times 10^{-1} \pm 2.49027 \times 10^{-6}$  Jy, or approximately 99.9975% of the original peak brightness (1.0 Jy). This indicates that the first singular value effectively fully captures the data features, leaving behind noise features. The higher-order singular values contribute minimally and primarily capture the noise features, recovering peak brightness below the RMS value.

At high declinations, we anticipated more smearing due to baselines moving faster, which would require more singular values, particularly on long baselines. However, consistent with the case of the point source at the phase centre, the first singular value remains sufficient to represent the data. This indicates that the source or field direction does not impact the BDSVD compression.

The accompanying statistics table in Figure 3.16 demonstrates that the residual distribution for the first singular value is consistent with the noise distribution, confirming its noise-like nature. In contrast, the residual distributions for higher-order singular values deviate from the noise distribution, indicating the presence of original data features. These residuals have relatively larger mean, standard deviation,

and  $\chi^2_{nvis}$  values, reinforcing the conclusion that they contain signal components.

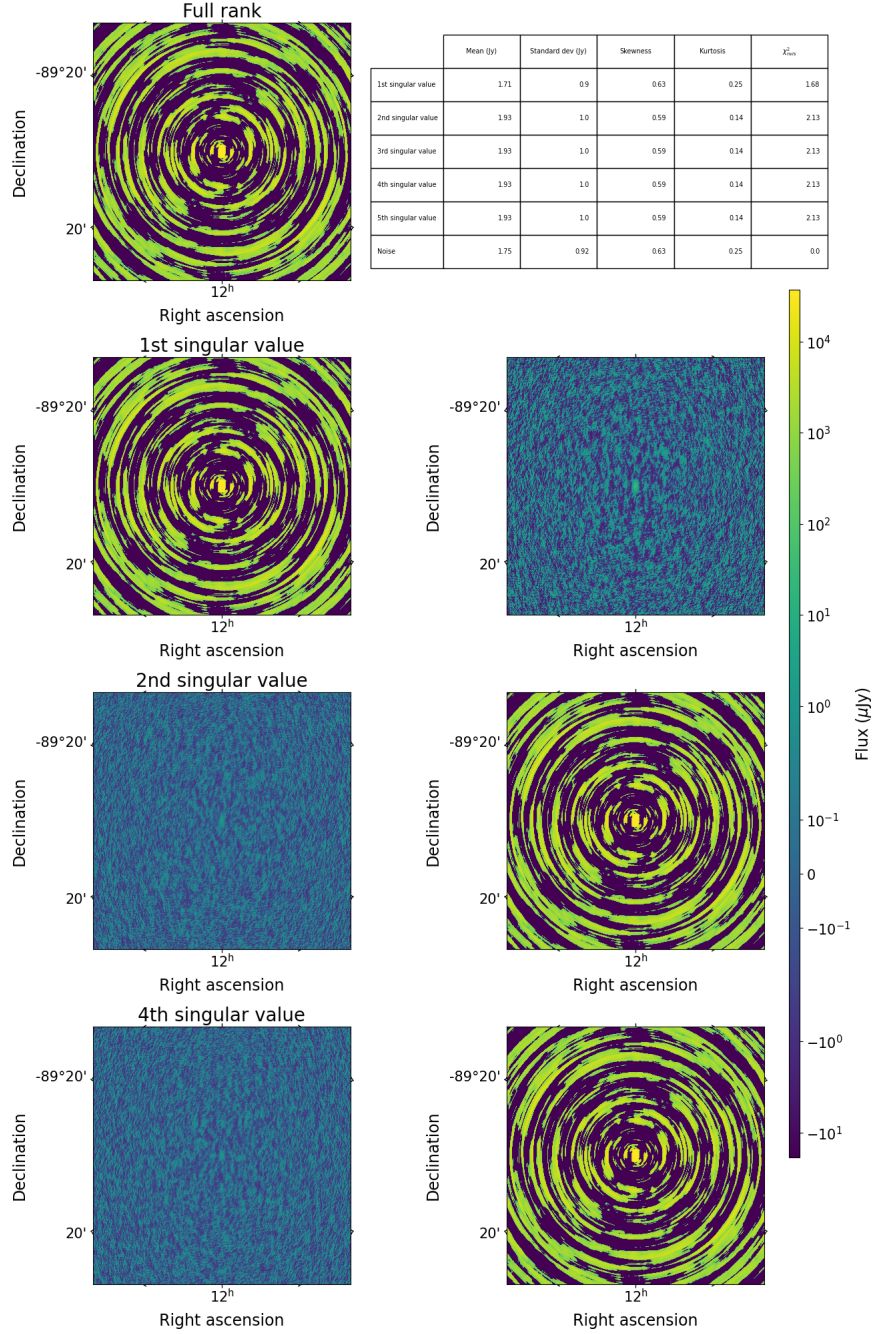


FIGURE 3.16: The left panel presents the original data images along with the compressed data images for the sky distribution with an off-centre point source at a declination of  $-90$  deg. The right panel presents the residual statistics table along with the residual images.

Thus, the first singular value alone is sufficient to represent the data across all baselines, effectively fully capturing the data features while leaving noise in its

residuals. Higher-order singular values primarily capture noise features, as evidenced by their residual images revealing the original data features. Consequently, the data can be effectively compressed to rank 1, achieving a compression factor of approximately 50 even at high declinations. This demonstrates that BDSVD compression is independent of source direction.

## 3.2 Cumulative Singular Values

In the previous section, we examined how different sky distributions influence the features captured by individual singular values. For a point source, the first singular value was shown to fully encapsulate the data features, making it sufficient for data representation. However, in scenarios with multiple sources or low SNR, higher-order singular values also play a role in representing the data. In this section, we investigate the cumulative contribution of singular values to determine the number required to fully capture the data features for each sky distribution.

### 3.2.1 Phase Centre

For a sky distribution consisting of a point source located at the phase centre, the first singular value effectively fully encapsulates the data features, recovering 99.9986% of the peak brightness. Inclusion of additional singular values does not offer any benefit in terms of feature recovery or data fidelity. Similarly, the residual images in Figure 3.17 reveal no noticeable differences when more singular values are included.

The residual statistics presented in Figure 3.17 indicate that adding more singular values decreases the difference between the residual and the noise distribution. This is suggested by the skewness and kurtosis values aligning closely with those of the noise, as well as the progressively smaller  $\chi^2_{\text{nois}}$  values.

Thus, for this sky distribution, the first singular value is sufficient to represent the data without introducing smearing effects, achieving a compression factor of 50. Additional singular values are unnecessary across all baselines.



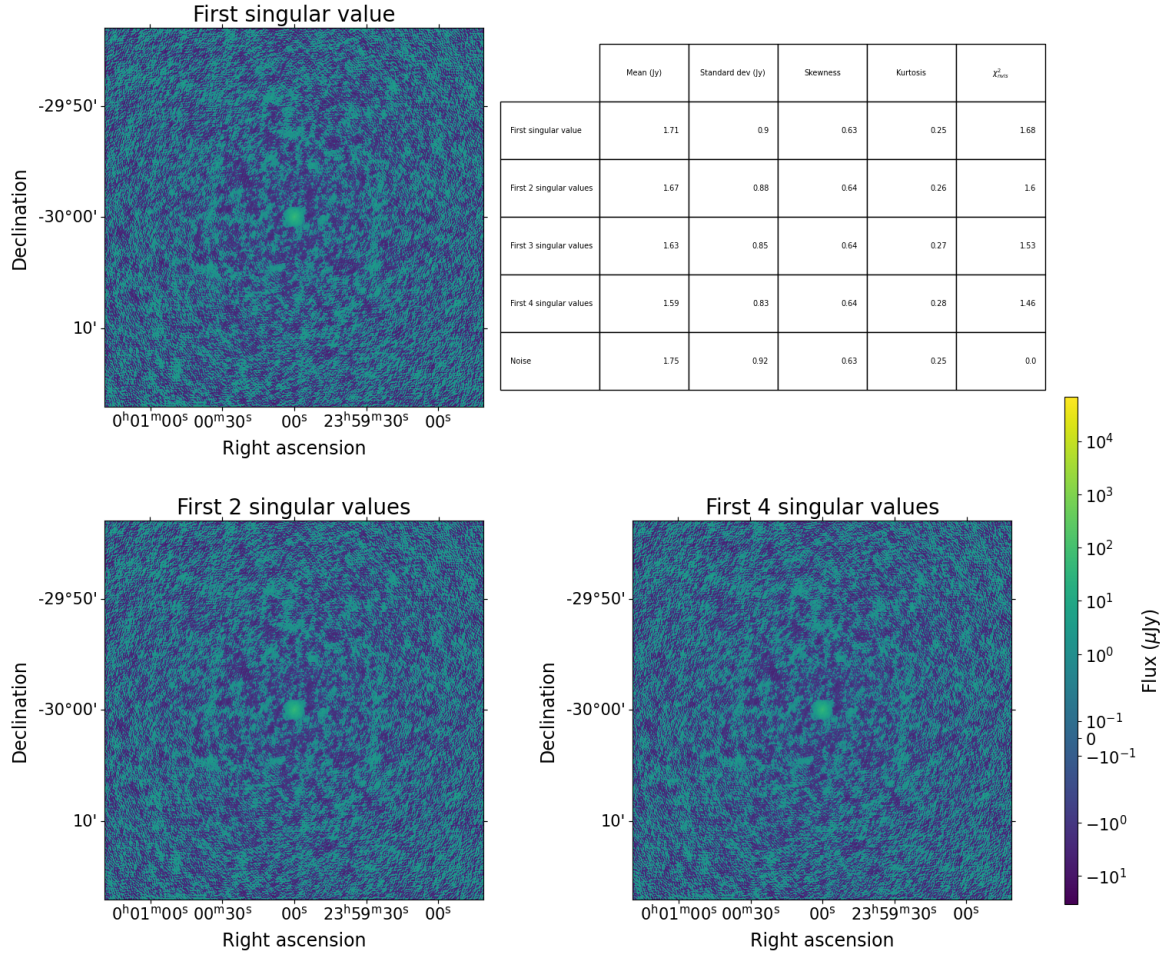


FIGURE 3.17: Residual statistics and images after reconstructing the data for the sky distribution with a point source at the phase centre using the first  $n$  singular value.

### 3.2.2 Away from the Phase centre

This case closely mirrors the previous one. The first singular value captures the data features almost entirely, recovering approximately  $9.57091 \times 10^{-1} \pm 1.83464 \times 10^{-5}$  Jy, or about 99.8966% of the original peak brightness ( $9.58082 \times 10^{-1}$  Jy). Including the second singular value increases the recovered peak brightness to  $9.57459 \times 10^{-1} \pm 1.83464 \times 10^{-5}$  Jy, corresponding to approximately 99.9350%. The inclusion of additional singular values beyond this point does not offer any benefit in terms of feature recovery or data fidelity.

The residual statistics presented in Figure 3.18 indicate that adding more singular values decreases the difference between the residual and the noise distribution.

This is suggested by the skewness and kurtosis values aligning closely with those of the noise, as well as the progressively smaller  $\chi_{nvis}^2$  values.

Thus, for this sky distribution, the data can be effectively compressed using only the first singular value across all baselines, with minimal smearing of around 0.1%. To further eliminate smearing, particularly on long baselines, the second singular value must be included, resulting in a compression factor of 24.

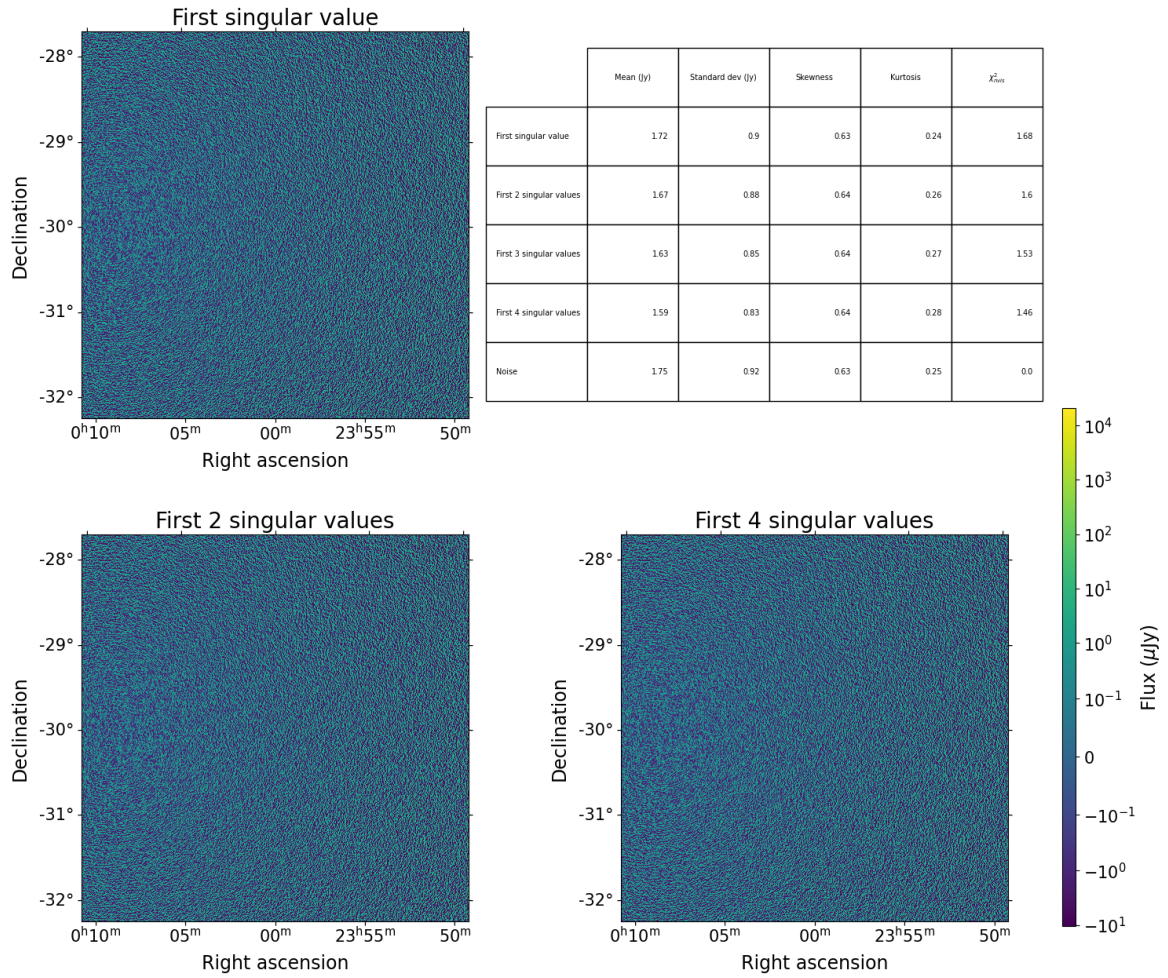


FIGURE 3.18: Residual statistics and images after reconstructing the data for the sky distribution with a point source 2.25 deg off-centre using the first  $n$  singular value.

### 3.2.3 Point Sources

For a sky distribution consisting of point sources, the first singular value captures the majority of the data features, recovering approximately  $8.312\,69 \times 10^{-1} \pm 8.812\,99 \times$

$10^{-5}$  Jy, or about 98.3540% of the original peak brightness ( $8.45181 \times 10^{-1}$  Jy). This results in approximately 2% smearing for sources at the edge of the FoV, away from the phase centre. This limitation arises because the first singular value struggles to capture features away from the phase centre, as illustrated in Figure 3.10. These features are instead captured by higher-order singular values. Therefore, including additional singular values is necessary to eliminate smearing on medium and long baselines. This process can be likened to widening a boxcar function in the image plane, enabling the full capture of features away from the phase centre.

By adding the second singular value, the recovered peak brightness increases to  $8.44619 \times 10^{-1} \pm 8.81299 \times 10^{-5}$  Jy, corresponding to about 99.9336% of the original peak brightness. Including more singular values beyond this point offers little benefit in terms of feature recovery and data fidelity. For example, incorporating the first seven singular values recovers  $8.44847 \times 10^{-1} \pm 8.81299 \times 10^{-5}$  Jy, or about 99.9605%, which represents a little improvement over using just the first two singular values.

The residual statistics presented in Figure 3.19 align with previous findings. Adding singular values decreases the difference between the residual and the noise distribution. This is suggested by the skewness and kurtosis values aligning closely with those of the noise, as well as the progressively smaller  $\chi_{nvis}^2$  values.

Thus, for this sky distribution, the first two singular values are sufficient for effective compression on medium and long baselines, achieving a compression factor of approximately 24, while on short baselines, the first singular value alone remains adequate. This highlights the baseline-dependent nature of the compression. Using the first two singular values results in a recovery of approximately 99.9336%, with smearing reduced to about  $6.63974 \times 10^{-2}\%$ .

### 3.2.4 Point and Extended Sources

For this sky distribution, the first singular value recovers approximately  $2.80953 \times 10^{-1} \pm 1.66982 \times 10^{-4}$  Jy, corresponding to about 96.9696% of the original peak brightness ( $2.89733 \times 10^{-1}$  Jy) for a source at the edge of the FoV. This results in a smearing of about 3.03044%. In this case, data features are distributed among higher-order singular values, with these components predominantly capturing features away from the phase centre.



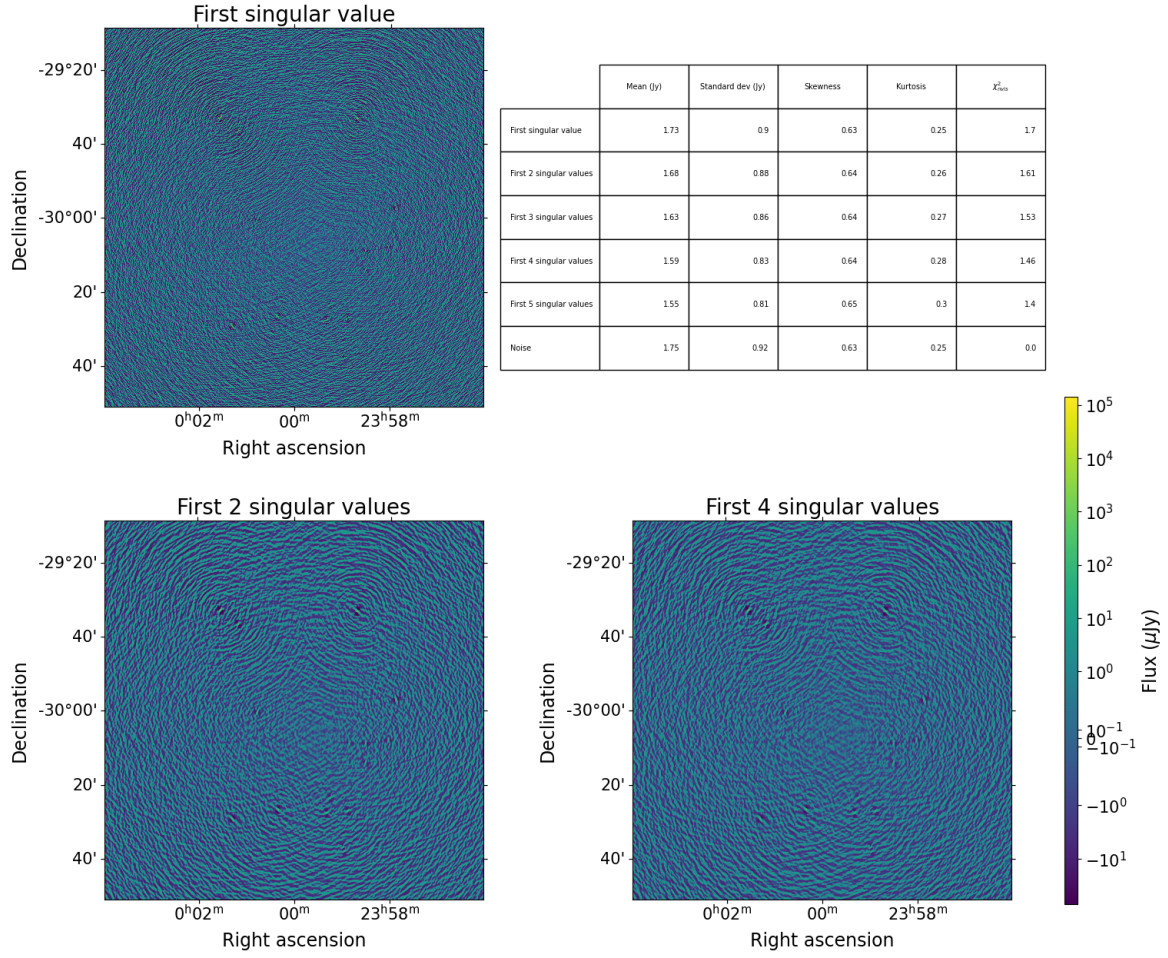


FIGURE 3.19: Residual statistics and images after reconstructing the data for the sky distribution with multiple point sources using the first  $n$  singular value.

By adding the second singular value, additional features are captured, increasing the recovered peak brightness to approximately  $2.89297 \times 10^{-1} \pm 1.66982 \times 10^{-4} \text{Jy}$ , or around 99.8494%. Including more singular values further increases the recovered peak brightness, demonstrating that incorporating additional singular values is analogous to widening a boxcar function in the image plane, enabling the capture of features away from the phase centre. For example, using the first three singular values increases the recovered peak brightness to  $2.89510 \times 10^{-1} \pm 1.66982 \times 10^{-4} \text{Jy}$ , corresponding to about 99.9231% of the original peak brightness.

However, including more singular values beyond the third does not have significant benefits in terms of feature recovery and data fidelity. For instance, using the first eight singular values recovers  $2.89607 \times 10^{-1} \pm 1.66982 \times 10^{-4} \text{Jy}$ , or around

99.9564% of the original peak brightness, which represents a little improvement over using the first three singular values.

The residual statistics presented in Figure 3.20 show that the residual distribution becomes increasingly consistent with the noise distribution as the  $\chi^2_{nvis}$  values progressively decrease.

Thus, for this case, the first singular value is sufficient to represent the data on short baselines, achieving a compression factor of approximately 50. On medium and long baselines, the first three singular values are sufficient to represent the data, recovering around 99.9231% of the peak brightness for a source at the edge of the field. This results in a compression factor of approximately 16 on medium and long baselines. Adding more singular values beyond the first three provides only modest improvements in data recovery.

### 3.2.5 SNR

For a sky distribution with a point source with low SNR ( $\text{flux} = 100\sigma_{img}$ ), the first singular value no longer dominates and fails to capture the majority of the data features. Instead, the data features are distributed equally among all singular values. In this case, compressing the data using only the first singular value recovers only  $2.96172 \times 10^{-4} \pm 5.15278 \times 10^{-5}$  Jy, corresponding to approximately 4.55803% of the original peak brightness ( $6.49781 \times 10^{-3}$  Jy). Adding the second singular value increases the recovered peak brightness to  $6.071 \times 10^{-4} \pm 5.15278 \times 10^{-5}$  Jy, or about 9.34316%.

To fully capture the data features and recover all of the peak brightness, all the singular values are required, as they equally contribute to the representation of the data. Table 3.2 summarizes the recovered peak brightness when compressing the data using the first  $n$  singular values.

Figure 3.21 presents the residual statistics and images. The statistics indicate that including the first twelve singular value reduces the difference between the residual distribution and the noise distribution, suggesting that the residual is predominantly noise. However, beyond the first twelve singular values, the residual distribution deviates from the noise distribution, indicating a modification in its structure.

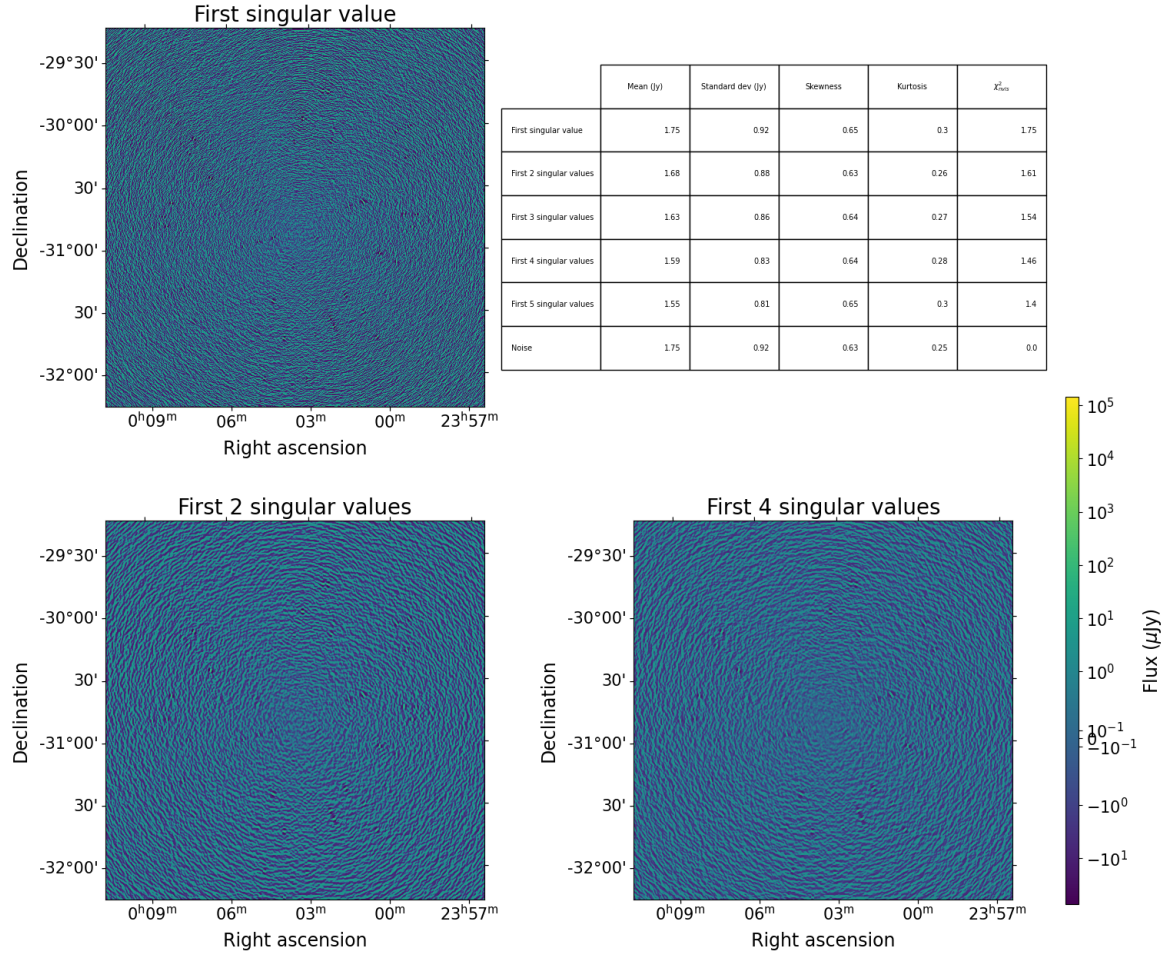


FIGURE 3.20: Residual statistics and images after reconstructing the data for the sky distribution with point and extended sources using the first  $n$  singular value.

TABLE 3.2: Recovered peak brightness for a sky distribution with SNR of 100.

| Singular Values | Recovered Peak Brightness $\pm 5.15278 \times 10^{-5}$ (Jy) | Recovered Peak Brightness (%) |
|-----------------|-------------------------------------------------------------|-------------------------------|
| First           | $2.96172 \times 10^{-4}$                                    | 4.55803                       |
| First 2         | $6.071 \times 10^{-4}$                                      | 9.34316                       |
| First 3         | $9.07106 \times 10^{-4}$                                    | 13.9602                       |
| First 4         | $1.19735 \times 10^{-3}$                                    | 18.4270                       |
| First 12        | $3.43791 \times 10^{-3}$                                    | 52.9088                       |
| First 18        | $5.01965 \times 10^{-3}$                                    | 77.2515                       |
| First 20        | $5.51076 \times 10^{-3}$                                    | 84.8095                       |
| First 22        | $6.02015 \times 10^{-3}$                                    | 92.6490                       |
| First 23        | $6.26934 \times 10^{-3}$                                    | 96.4839                       |
| All 24          | $6.49781 \times 10^{-3}$                                    | 100                           |



Thus, for a sky distribution consisting of a faint source, BDSVD proves ineffective. This is because all singular values are necessary to fully capture the data features and recover the source intensity, leaving no opportunity for meaningful compression.

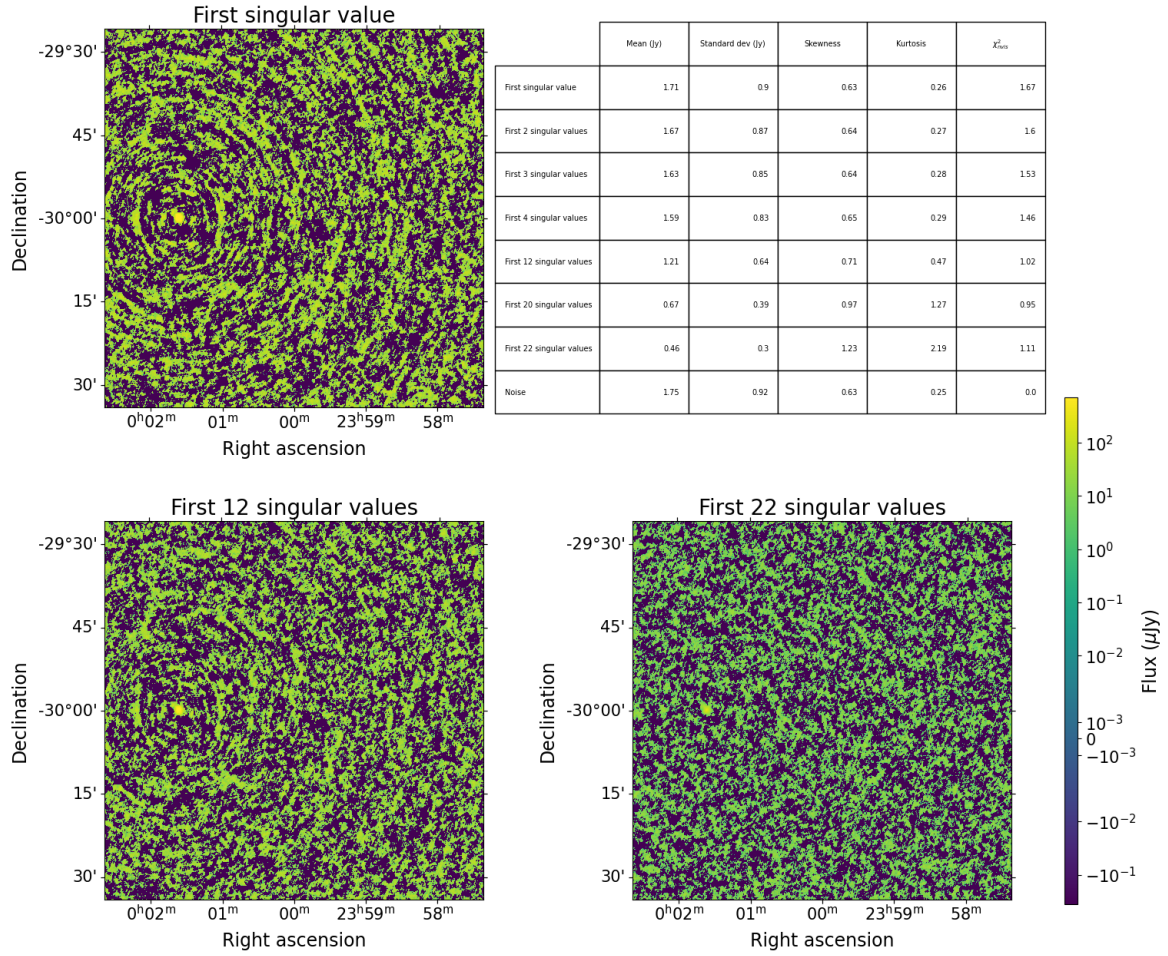


FIGURE 3.21: Residual statistics and images after reconstructing the data for the sky distribution with a point source off-centre with an SNR of 100 using the first  $n$  singular value.

### 3.2.6 Point Sources with a total flux of 10 Jy

As shown in Figure 3.13, multiple singular values dominate across all baselines and contribute to capturing data features. Consequently, the first singular value recovers only approximately  $6.780\,02 \times 10^{-2} \pm 9.950\,74 \times 10^{-4}$  Jy, corresponding to about 54.4263% of the original peak brightness ( $1.245\,73 \times 10^{-1}$  Jy) for a source at the edge

of the FoV. Adding the second singular value increases the recovered peak brightness to  $8.86829 \times 10^{-2} \pm 9.95074 \times 10^{-4}$  Jy, or about 71.1897%. To fully capture the features and achieve complete recovery of the brightness, additional singular values must be included.

Table 3.3 summarizes the recovered peak brightness when compressing the data using the first  $n$  singular values. It is evident that the first twelve singular values effectively fully capture the features and recover the peak brightness, recovering around  $1.24512 \times 10^{-1} \pm 9.95074 \times 10^{-4}$  Jy, corresponding to about 99.9516%. Including more singular values beyond this point results in only marginal improvements in feature recovery and is therefore unnecessary. For instance, using the first twenty-three singular values increases the recovered peak brightness to only  $1.24568 \times 10^{-1} \pm 9.95074 \times 10^{-4}$  Jy, or around 99.9965%.

TABLE 3.3: Recovered peak brightness for a sky distribution with multiple sources with a total flux of 10 Jy.

| Singular Values | Recovered Peak Brightness $\pm 9.95074 \times 10^{-4}$ (Jy) | Recovered Peak Brightness (%) |
|-----------------|-------------------------------------------------------------|-------------------------------|
| First           | $6.78002 \times 10^{-2}$                                    | 54.4263                       |
| First 2         | $8.86829 \times 10^{-2}$                                    | 71.1897                       |
| First 3         | $1.00125 \times 10^{-1}$                                    | 80.3747                       |
| First 4         | $1.06565 \times 10^{-1}$                                    | 85.5444                       |
| First 5         | $1.12066 \times 10^{-1}$                                    | 89.96                         |
| First 12        | $1.24512 \times 10^{-1}$                                    | 99.9516                       |
| All 24          | $1.24573 \times 10^{-1}$                                    | 100                           |

The residual statistics presented in Figure 3.22 indicate that adding more singular values decreases the difference between the residual and the noise distribution. This is suggested by the mean and standard deviation values aligning closely with those of the noise, as well as the progressively smaller  $\chi^2_{nvis}$  values.

Thus, the data can be effectively compressed using the first four singular values on short baselines, the first eight singular values on medium baselines, and the first twelve singular values on long baselines. This approach achieves a minimum compression factor of four on long baselines.

### 3.2.7 Source Direction

For the case where the source direction has been changed to a declination of  $-90$  deg, the first singular value recovers a peak brightness of  $9.99977 \times 10^{-1} \pm 2.49027 \times$

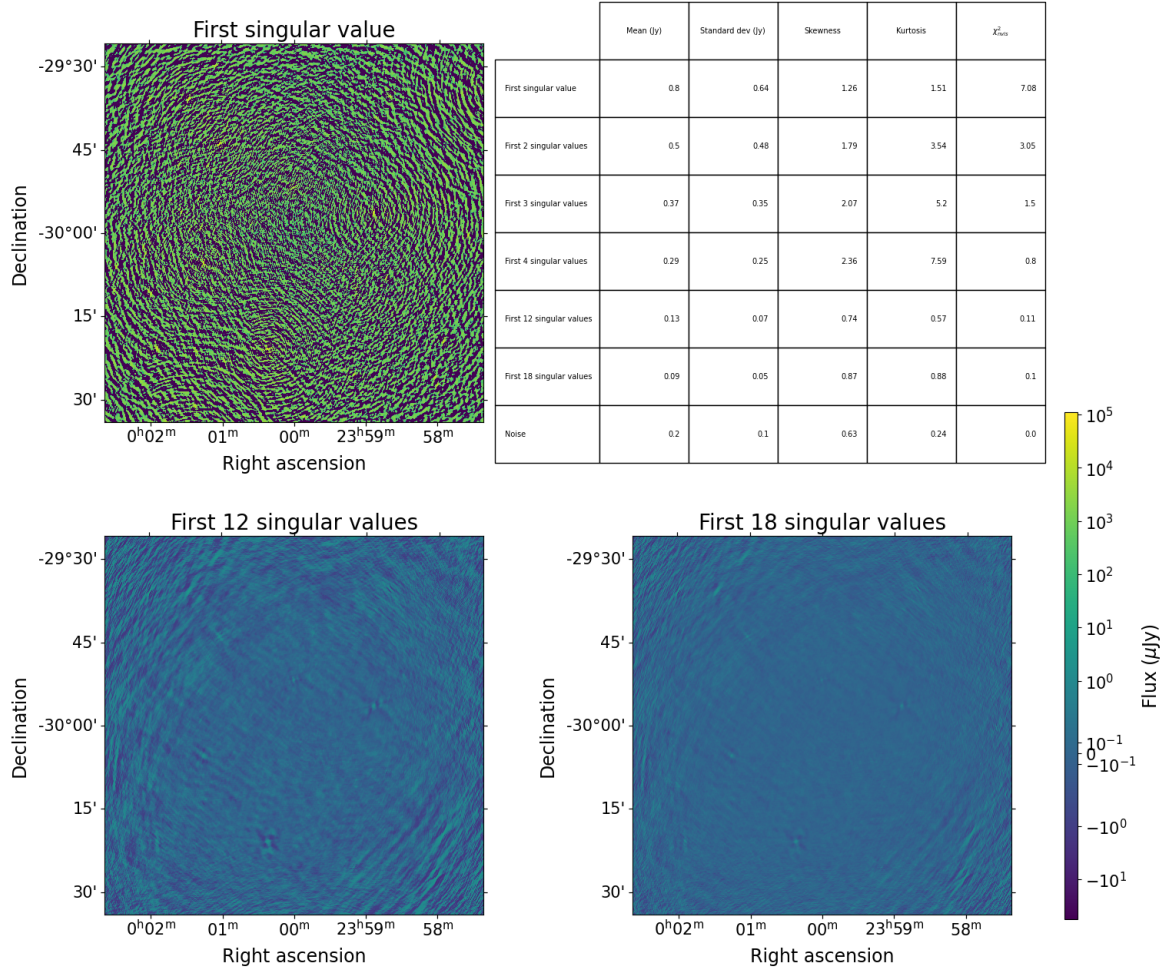


FIGURE 3.22: Residual statistics and images after reconstructing the data for the sky distribution with 100 point sources with a total flux of 10 Jy using the first  $n$  singular value.

$10^{-6}$  Jy, corresponding to approximately 99.9975% of the original peak brightness (1.0 Jy). The inclusion of additional singular values beyond this point does not offer any benefit in terms of feature recovery or data fidelity. Figure 3.23 presents the residual images and statistics after compressing the data using the first  $n$  singular values.

For this sky distribution, the first singular value is sufficient to fully represent the data.

Through the presentation of our results, we have shown that BDSVD is highly effective for bright sources, as only the first few singular values are required to capture the data features and recover the source brightness. However, BDSVD performs poorly for faint sources or those with low SNR. In these cases, all singular

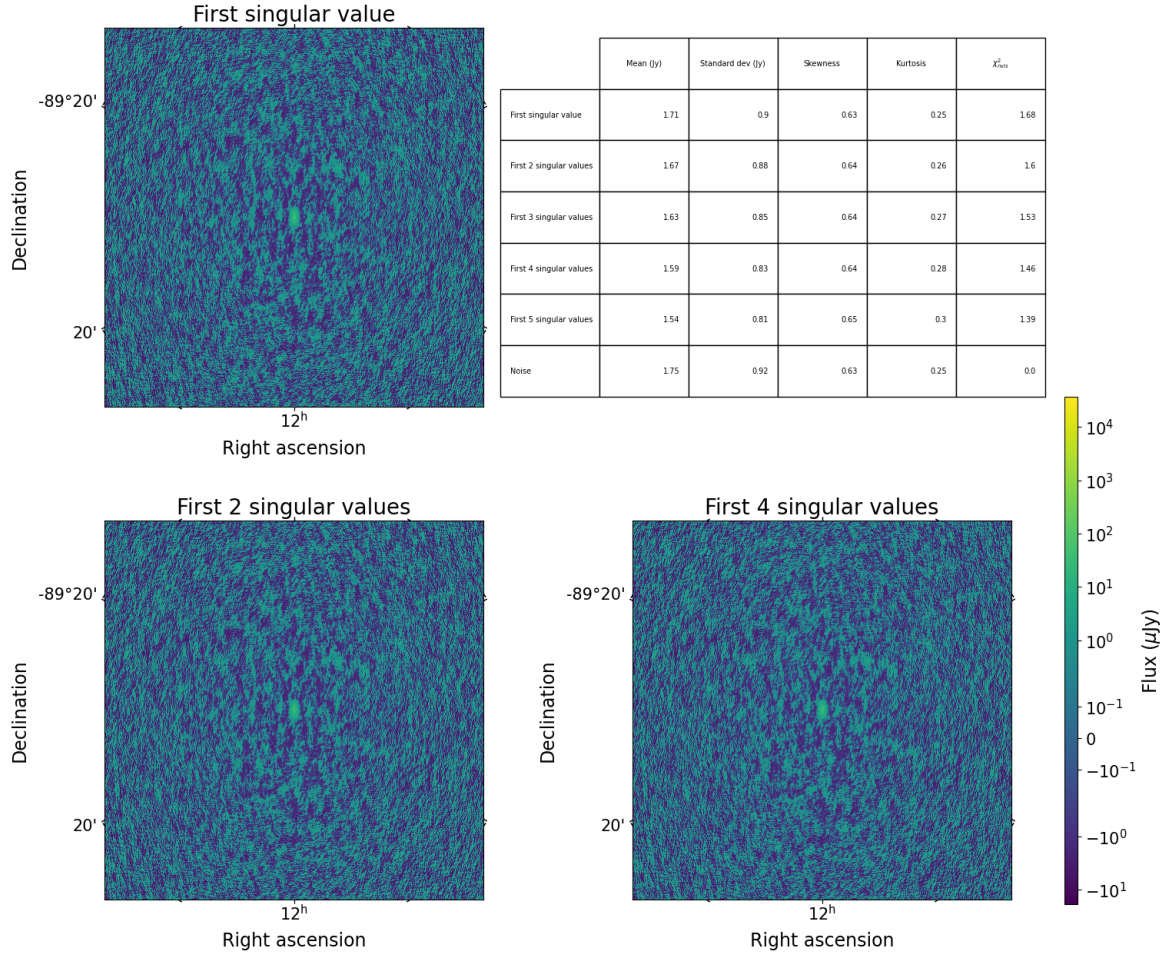


FIGURE 3.23: Residual statistics and images after reconstructing the data for the sky distribution with an off-centre point source at a declination of  $-90^\circ$  using the first  $n$  singular value.

values are needed to fully represent the data and recover the brightness, rendering the compression ineffective. This limitation significantly restricts the practical applicability of BDSVD in real-world observational scenarios.

Table 3.4 summarizes the BDSVD compression performance across the scenarios investigated in this section. It lists the recommended number of singular values ( $n$ ), the percentage of source brightness recovered, and the resulting compression factor for each case.

Section 6 presents additional results, including performance across different SNR scenarios and source directions, for an off-centre point source. These results show that BDSVD performs poorly in low-SNR conditions, requiring all singular



TABLE 3.4: Summary of BDSVD compression performance across different scenarios

| <b>Description</b>                  | <b>Recommended <math>n</math></b> | <b>Recovery (%)</b> | <b>Comp. factor</b> |
|-------------------------------------|-----------------------------------|---------------------|---------------------|
| Bright point source at phase centre | 1                                 | $> 99.90$           | 50                  |
| Bright point source off-centre      | 1                                 | $> 99.80$           | 50                  |
| Multiple point sources              | 2–3                               | $> 99.90$           | 24                  |
| Point and extended sources          | 3–4                               | $> 99.90$           | 16                  |
| Low SNR (SNR = 100)                 | 24                                | $\approx 100$       | 0                   |
| Point sources, total flux 10 Jy     | 4–12                              | $> 99.90$           | 4                   |
| Source direction (Dec = $-90$ deg)  | 1                                 | $> 99.90$           | 50                  |

values to fully capture the data. Furthermore, the consistent compression performance across various source directions confirms that the method’s effectiveness is not affected by the source or field direction.

In the next chapter, we introduce our software package, which implements SVD-based techniques for compressing visibility data. Specifically, the package supports both simple SVD and BDSVD methods.



## Chapter 4

# Software Package

This thesis is complemented by a software package I developed, *visco* [70], designed for compressing visibility data using the Singular Value Decomposition (SVD) techniques proposed by Atemkeng et al. [8]. The package offers two primary functionalities: one for compressing and archiving visibility data, and another for decompressing the archived data.

To install *visco*, use the following command:

```
pip install visco
```

## 4.1 Compressing and Archiving Visibility Data

### 4.1.1 Simple SVD

There are two primary methods for compressing the data prior to archiving it. The first method, referred to as simple SVD, involves using the same number of singular values (or the same rank,  $n$ ) across all baselines  $pq$ . To use this option, the user can run the following command:

LISTING 4.1: The simple SVD command.

```
archival --ddid 0 --fieldid 0 --scan 1 --correlation XX,  
YY --column DATA --compressionrank 10 --outfilename  
archived-data.zarr MSPATH.ms
```

In Listing 4.1, the `archival` command is used to perform data compression. The parameters are as follows:

- `ddid` specifies the `DATA_DESC_ID` of the data in the Measurement Set (MS), with a default value of 0.
- `fieldid` specifies the `FIELD_ID`, with a default of 0.
- `scan` specifies the `SCAN_NUMBER`, with a default value of 1.
- `correlation` defines the data correlations to compress and store.
- `column` identifies the data column in the MS.
- `compressionrank` sets the rank or number of singular values to retain across all baselines.
- `outfilename` specifies the name of the zarr where the data will be archived.

Finally, the user must provide the full path to the MS file.

The compressed data is stored in a zarr file, along with all columns from the main table and subtables of the MS. A Zarr file is a chunked, compressed, and scalable storage format for large multidimensional arrays. Among these columns, the `FLAG` and `WEIGHT_SPECTRUM` columns contribute significantly to storage space apart from the data column.

In the original work by Atemkeng et al., only the data column was considered for compression, with no attention given to the `FLAG` and `WEIGHT_SPECTRUM` columns. Our software package addresses this limitation by implementing compression for these additional columns. Specifically, the `FLAG` column is compressed using the `numpy.packbits` function, which groups 8 boolean values into a single integer, thereby reducing its size by a factor of 8.

For the `WEIGHT_SPECTRUM` column, compression is achieved through SVD. If the user employs simple SVD, the same compression rank applied to the data column is also used for the `WEIGHT_SPECTRUM` column. In the case of BDSVD, the user can specify the desired rank for the `WEIGHT_SPECTRUM` column via the parameter `weightcr` in Listing 4.2, with a default value of 1.

### 4.1.2 Baseline-Dependent Singular Value Decomposition (BDSVD)

The second method is referred to as Baseline-Dependent Singular Value Decomposition (BDSVD), where the rank or number of singular values retained varies based

on the baseline length. This approach allows for more aggressive compression on shorter baselines by retaining fewer singular values while preserving greater detail on longer baselines by retaining more. To use this method, the user specifies the decorrelation parameter instead of `compressionrank`. The decorrelation parameter defines the minimum fraction of the signal that must be preserved during compression for each baseline. For instance, in Listing 4.2, setting decorrelation to 0.98 ensures that at least 98% of the signal is retained on each baseline.

LISTING 4.2: The BDSVD command.

```
archival --ddid 0 --fieldid 0 --scan 1 --correlation XX,
        YY --column DATA --decorrelation 0.98 --outfilename
        archived-data.zarr MS.ms
```

## 4.2 Data Decompression

The data must be decompressed before it can be used. The second functionality of `visco` enables the decompression of data from a `zarr` file back into an MS. To decompress the data, the user simply runs the following command:

LISTING 4.3: The data decompression command.

```
decompression --zarr-path ZARRPATH.zarr --output-column DATA
              --ms MSNAME.ms
```

In Listing 4.3, the `zarr-path` parameter specifies the full path to the `zarr` file containing the compressed data. The `output-column` parameter indicates the column in the MS where the reconstructed visibility data will be stored. Finally, the `ms` parameter defines the name of the new MS that will contain the decompressed and reconstructed data.

## Chapter 5

# Conclusions and Future Work

### 5.1 Conclusions

In this work, we built on the work of Atemkeng et al. [8], which proposed using Singular Value Decomposition (SVD) to compress radio interferometric data. Our investigation focused on evaluating the effectiveness of Baseline-Dependent Singular Value Decomposition (BDSVD) in compressing visibility data while preserving data fidelity and maintaining image quality across various observational scenarios. We have explored the potential and limitations of BDSVD as a practical solution for managing the growing data volumes in radio interferometry.

We compressed the data by decomposing the data matrix  $V_{pq,r}$  for baseline  $pq$  into three matrices:  $U$ ,  $\Lambda$ , and  $W^\dagger$ . The matrices were truncated to the desired number of components  $n$  or compression factor  $C_f$ , and the data was reconstructed as a low-rank approximation  $V_{pq,n}$ , representing the compressed data.

Our results demonstrated that for a sky model with a bright (1 Jy) point source located at the phase centre, the first component can effectively capture all the essential data features across all baselines. This allows the other components to be discarded, resulting in a compression factor of approximately 50. Additionally, we found this outcome to hold true for different source directions, indicating that source direction does not impact BDSVD compression.

Although source distance does influence BDSVD compression, the effect is not significant. For example, we analyzed a sky distribution with a 1 Jy source located at a distance of 2.25 deg, which corresponds to the second sidelobe of the MeerKAT telescope at a frequency of 1.4 GHz. Our results show that the first component can effectively represent the data across all baselines with minimal smearing of only

0.1%. This smearing can be eliminated by including the second component for long baselines.

For fields containing multiple sources, whether point or extended, we found that the first component accurately represents features near the phase centre but struggles to do so away from it, leading to a smearing of approximately 4% for sources at the edges of the FoV, which is roughly 1.1 deg for the MeerKAT telescope at a frequency of 1.4 GHz. At the edges, the inclusion of additional components on medium and long baselines is necessary to eliminate this smearing. Using 3–4 components, the data features can be effectively fully represented, achieving a compression factor of around 16 on medium and long baselines.

Additionally, we investigated the impact of Signal to Noise Ratio (SNR) on BDSVD compression. Our findings indicate that for sky distributions containing a faint source or a source with a low SNR, all components are necessary to represent the data without introducing smearing. As a result, BDSVD becomes ineffective in such scenarios.

Overall, BDSVD demonstrates the ability to achieve high compression rates while preserving temporal and spectral resolution, as well as data fidelity and image quality. However, it faces significant limitations when dealing with faint sources or sources that have a low SNR. This is particularly challenging because, in practice, most observations will involve sources within the SNR range where BDSVD struggles. As a result, its effectiveness may be compromised in real-world observational data. In such scenarios, it would be more advantageous to first denoise the data or to use Baseline Dependent Averaging (BDA) for compression.

A substantial amount of work is still required to further explore and understand the full capabilities and limitations of BDSVD.

## 5.2 Future Work

In this work, we explored Singular Value Decomposition (SVD)-based compression for radio interferometric visibility data, evaluating its potential and limitations for long-term archival across diverse observational scenarios. While our results demonstrate that SVD-based compression performs effectively in many cases, significant challenges arise when handling faint sources, which may limit its reliability in real-world applications. To address this limitation, a natural next step is to

validate the method using real observational datasets. Specifically, we aim to investigate SVD compression on visibility data from catalogs derived from source count measurements in large-scale radio continuum surveys. Surveys such as the NRAO VLA Sky Survey (NVSS), the deep VLA Sky Survey (VLASS), and the ultra-sensitive MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) offer ideal testbeds for this purpose. By analyzing these datasets, we can further evaluate the performance of SVD on faint source populations and complex sky distributions.

Another important avenue for future research is addressing datasets that contain flagged values, as applying SVD compression to such datasets can increase noise levels. Currently, one approach involves estimating and filling in values for flagged data points. Atemkeng et al. [8] are working on an efficient estimation method that uses interpolation to estimate the flagged values during compression. A promising alternative is to replace flagged values with corresponding model data values, which could provide a more accurate approximation. Developing an efficient and reliable strategy for handling flagged data is essential for the broader applicability of SVD compression.

Additionally, the original work by Atemkeng et al. [8] does not address storage-intensive columns such as FLAG and WEIGHT\_SPECTRUM. In our accompanying software package, the WEIGHT\_SPECTRUM column is currently compressed using SVD. Since this column has the same shape as the data column, it achieves a similar compression factor. However, this approach may not be optimal for handling the WEIGHT\_SPECTRUM, and in future work, we plan to develop a more robust method for compressing this column.

A significant drawback of SVD is its high computational complexity, which poses challenges for practical implementation. To address this limitation in the future, we plan to parallelize our software package. This will significantly reduce computation time, making SVD-based compression more efficient and feasible for broader applications.

## Chapter 6

# Appendix A

### 6.0.1 Signal to Noise Ratio (SNR)-1000

We consider a sky distribution with a point source located approximately 0.4 deg away from the phase centre with an SNR of 1000 ( $\text{flux} = 1000\sigma_{\text{img}}$ ). Figure 6.1 demonstrates that all singular values exhibit similar weights across all baselines. This indicates that data features are evenly distributed among all singular values. Consequently, the first singular value does not fully capture all data features and cannot solely represent the data (Figure 6.2 and 6.3).

For example, the first singular value recovers only about  $8.2457 \times 10^{-3} \pm 1.81433 \times 10^{-4} \text{Jy}$ , corresponding to approximately 12.6193% of the original peak brightness ( $6.53420 \times 10^{-2} \text{Jy}$ ). Subsequent singular values recover fewer features but still contribute significantly. For instance, the twentieth singular value recovers approximately  $1.39966 \times 10^{-3} \pm 1.81433 \times 10^{-4} \text{Jy}$ , or about 2.14205%. Therefore, all the singular values are required for data representation as they all capture the features. Table 6.1 summarizes the recovered peak brightness.

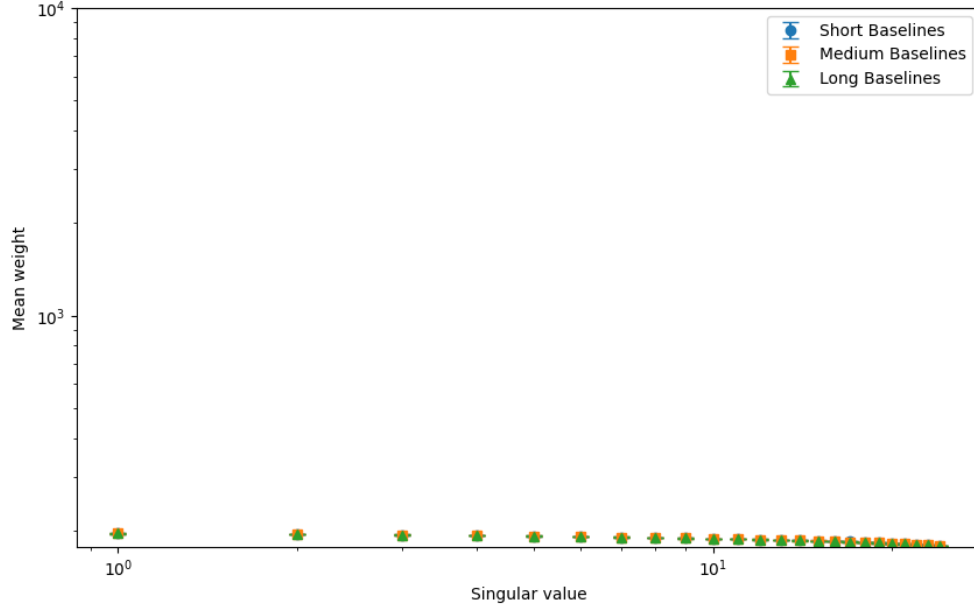


FIGURE 6.1: Singular values and their corresponding weights for the sky distribution with an off-centre point source with an SNR of 1000

TABLE 6.1: Recovered peak brightness for a sky distribution with SNR of 1000.

| Singular Values | Recovered Peak Brightness $\pm 1.81433 \times 10^{-4}$ (Jy) | Recovered Peak Brightness (%) |
|-----------------|-------------------------------------------------------------|-------------------------------|
| First           | $8.2457 \times 10^{-3}$                                     | 12.6193                       |
| First 2         | $1.43049 \times 10^{-2}$                                    | 21.8923                       |
| First 3         | $1.9417 \times 10^{-2}$                                     | 29.7159                       |
| First 4         | $2.36779 \times 10^{-2}$                                    | 36.2369                       |
| First 12        | $4.68009 \times 10^{-2}$                                    | 71.6245                       |
| First 18        | $5.75315 \times 10^{-2}$                                    | 88.0467                       |
| First 20        | $6.04267 \times 10^{-2}$                                    | 92.4776                       |
| First 22        | $6.30374 \times 10^{-2}$                                    | 96.4729                       |
| First 23        | $6.42401 \times 10^{-2}$                                    | 98.3136                       |
| All 24          | $6.53420 \times 10^{-2}$                                    | 100                           |



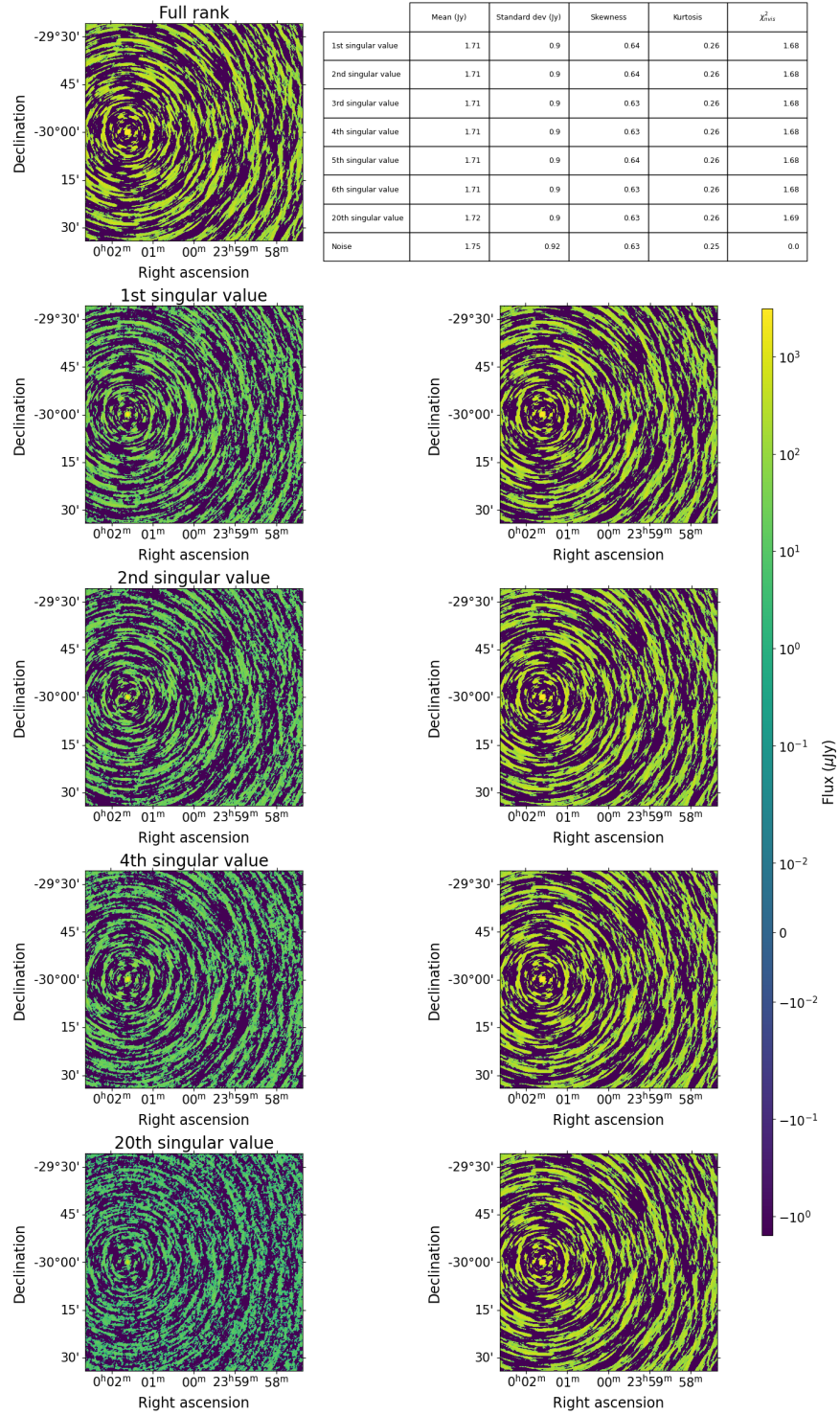


FIGURE 6.2: The left panel presents the original data along with the compressed data for the sky distribution with a point source at a distance of 0.4 deg with an SNR of 1000. The right panel presents the residual statistics table along with the residual images.

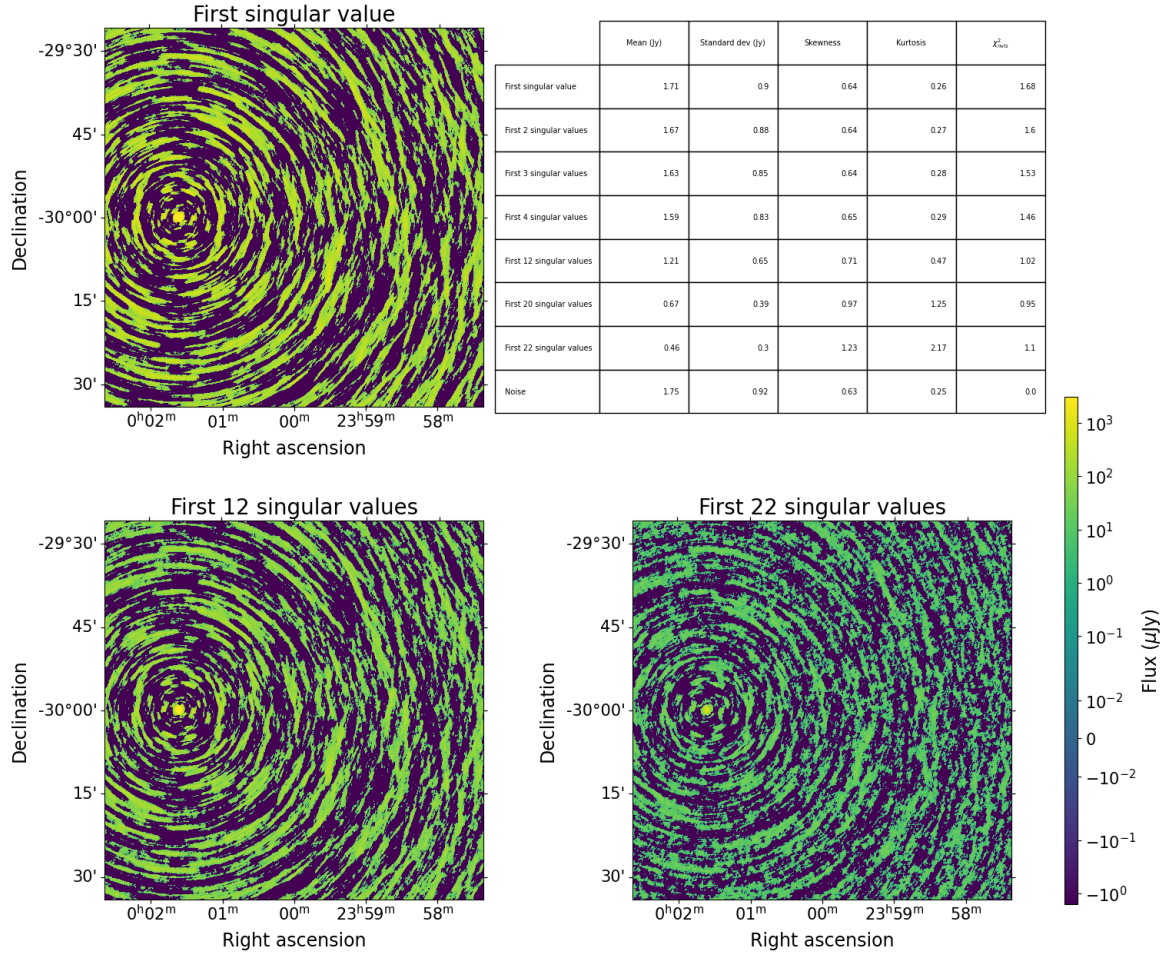


FIGURE 6.3: Residual statistics and images after reconstructing the data for the sky distribution with a point source at a distance of 0.4 deg with an SNR of 1000 using the first  $n$  singular value.

## 6.0.2 Signal to Noise Ratio (SNR)-50

We consider a sky distribution with a point source located approximately 0.4 deg away from the phase centre with an SNR of 50 (flux =  $50\sigma_{\text{img}}$ ). Figure 6.4 demonstrates that all singular values exhibit similar weights across all baselines. This indicates that data features are evenly distributed among all singular values. Consequently, the first singular value does not fully capture all data features and cannot solely represent the data (Figure 6.5 and 6.6).

For example, the first singular value recovers only about  $1.46341 \times 10^{-4} \pm 4.74444 \times 10^{-5}$  Jy, corresponding to approximately 4.5357% of the original peak brightness ( $3.22644 \times 10^{-3}$  Jy). Subsequent singular values recover an equal amount of the

peak brightness. For instance, the twentieth singular value recovers approximately  $1.244\,68 \times 10^{-4} \pm 4.744\,44 \times 10^{-5}$  Jy, or about 3.8578%. Therefore, all the singular values are required for the representation of the data. Table 6.2 summarizes the recovered peak brightness.

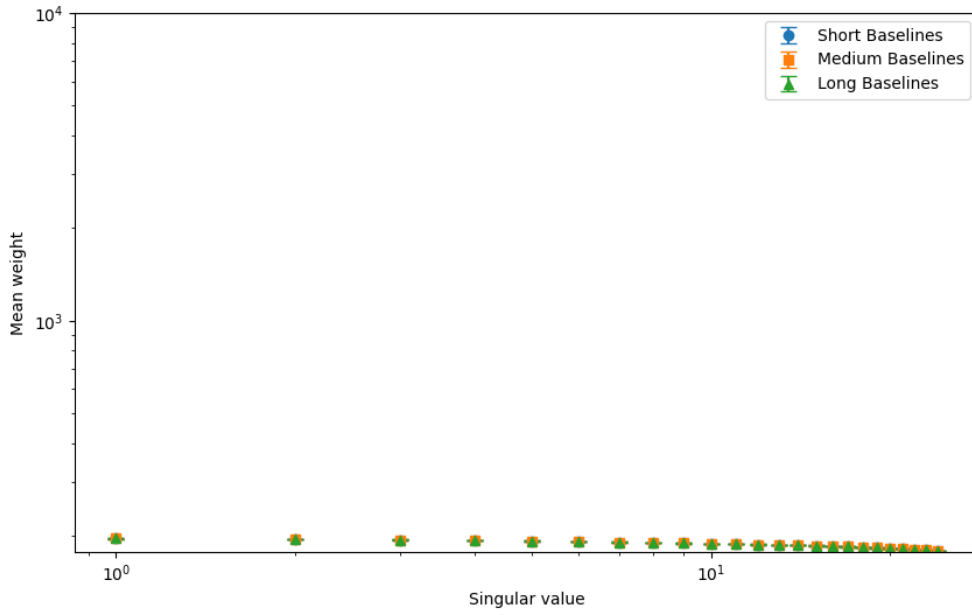


FIGURE 6.4: Singular values and their corresponding weights for the sky distribution with an off-centre point source with an SNR of 50

TABLE 6.2: Recovered peak brightness for a sky distribution with SNR of 50.

| Singular Values | Recovered Peak Brightness $\pm 4.744\,44 \times 10^{-5}$ (Jy) | Recovered Peak Brightness (%) |
|-----------------|---------------------------------------------------------------|-------------------------------|
| First           | $1.463\,41 \times 10^{-4}$                                    | 4.5357                        |
| First 2         | $2.822\,73 \times 10^{-4}$                                    | 8.7487                        |
| First 3         | $4.157\,73 \times 10^{-4}$                                    | 12.8864                       |
| First 4         | $5.558\,73 \times 10^{-4}$                                    | 17.2287                       |
| First 12        | $1.6907 \times 10^{-3}$                                       | 52.4014                       |
| First 18        | $2.473\,47 \times 10^{-3}$                                    | 76.6627                       |
| First 20        | $2.730\,04 \times 10^{-3}$                                    | 84.6147                       |
| First 22        | $2.987\,02 \times 10^{-3}$                                    | 92.5795                       |
| First 23        | $3.094\,38 \times 10^{-3}$                                    | 95.9069                       |
| All 24          | $3.226\,44 \times 10^{-3}$                                    | 100                           |

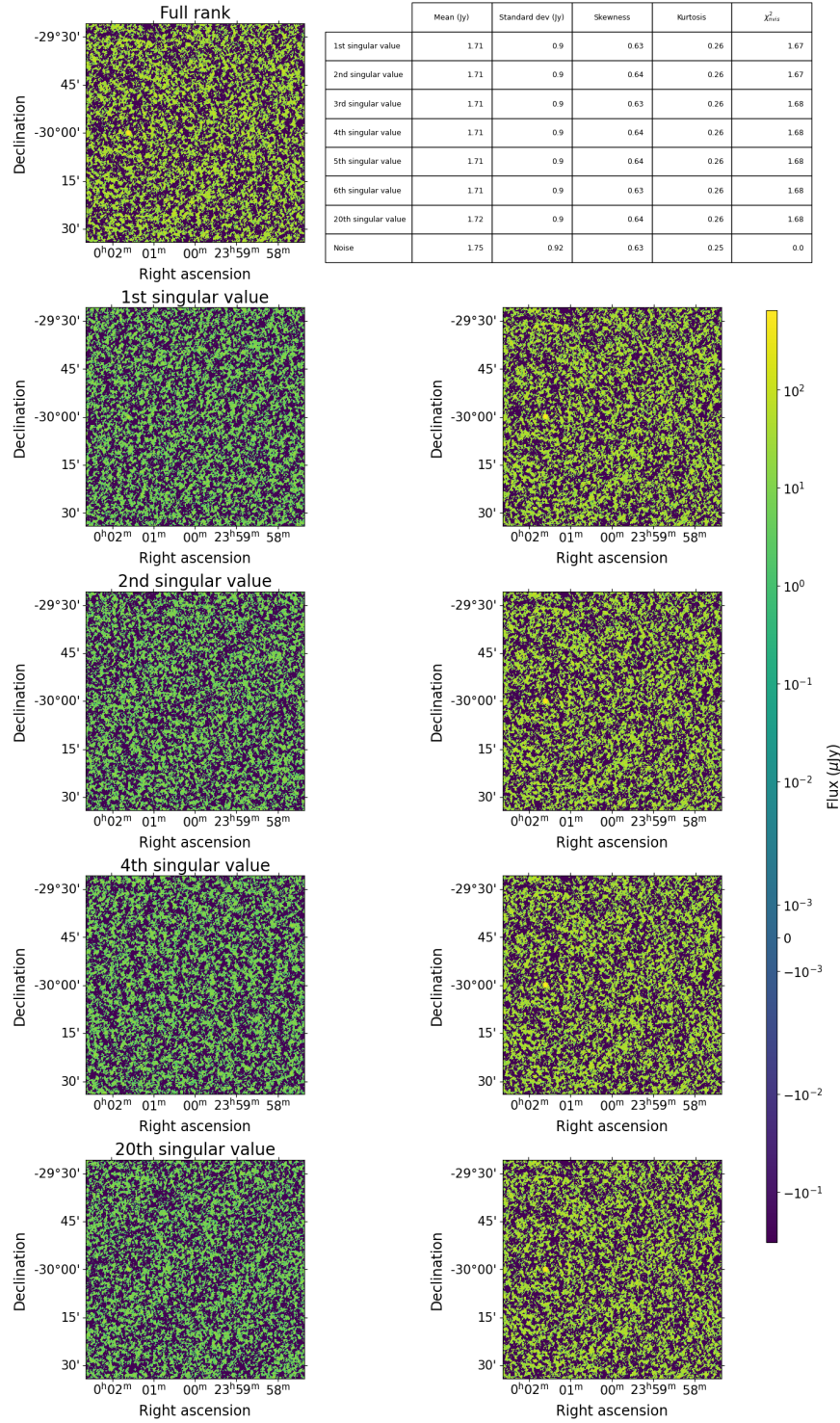


FIGURE 6.5: The left panel presents the original data along with the compressed data for the sky distribution with a point source at a distance of 0.4 deg with an SNR of 50. The right panel presents the residual statistics table along with the residual images.



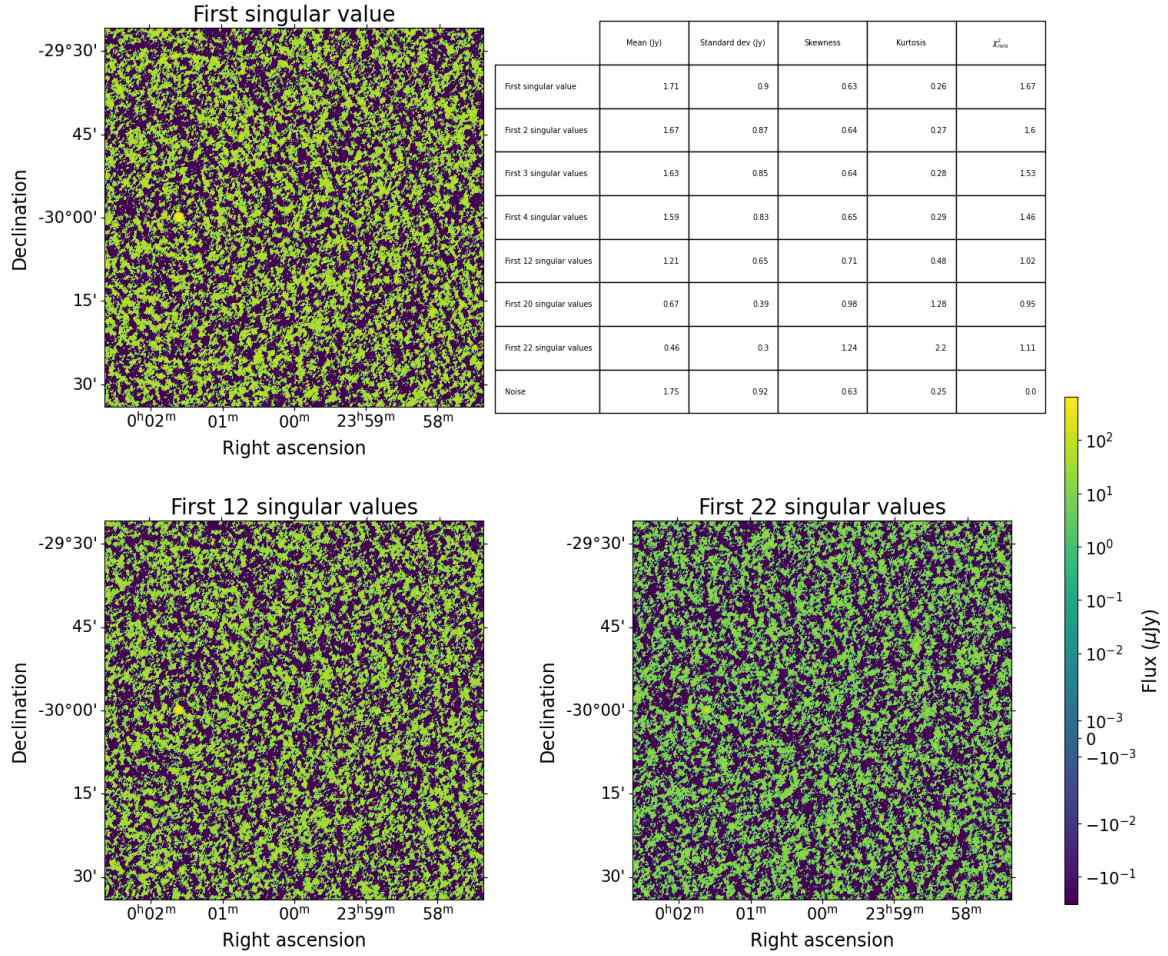


FIGURE 6.6: Residual statistics and images after reconstructing the data for the sky distribution with a point source at a distance of 0.4 deg with an SNR of 50 using the first  $n$  singular value.

### 6.0.3 Signal to Noise Ratio (SNR)-5

We consider a sky distribution with a point source located approximately 0.4 deg away from the phase centre with an SNR of 5 (flux =  $5\sigma_{\text{img}}$ ). Figure 6.7 demonstrates that all singular values have similar weights across all baselines. This indicates that data features are evenly distributed among all singular values. Consequently, the first singular value does not fully capture all data features and cannot solely represent the data (Figure 6.8 and 6.9).

For example, the first singular value recovers only about  $5.20468 \times 10^{-5} \pm 1.01201 \times 10^{-5}$  Jy, corresponding to approximately 14.9493% of the original peak brightness ( $3.48155 \times 10^{-4}$  Jy). Subsequent singular values also recover a significant amount

of the peak brightness. Therefore, all the singular values are required for the representation of the data. Table 6.3 summarizes the recovered peak brightness.

Therefore, BDSVD is not effective for compressing sky distributions consisting of faint sources.

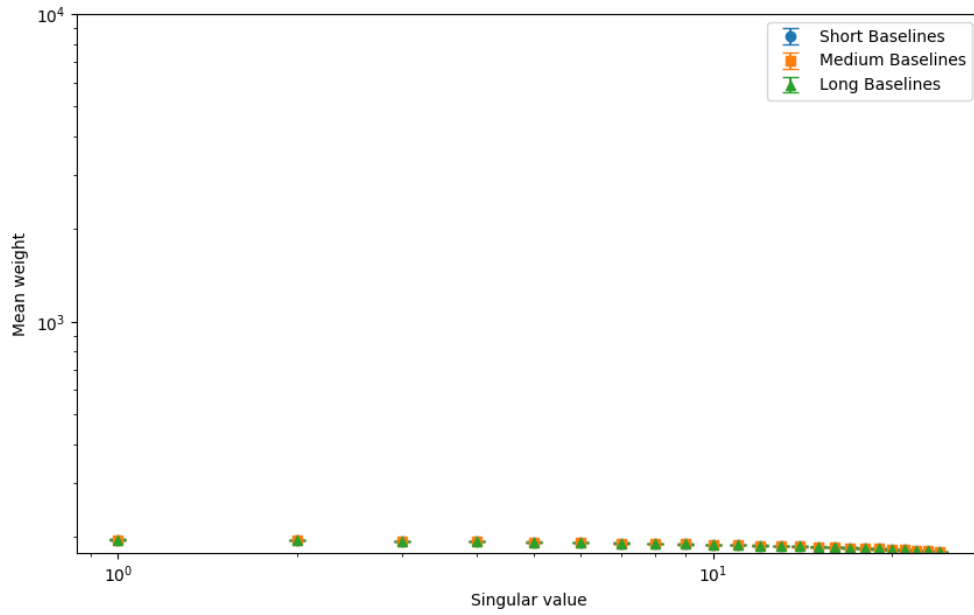


FIGURE 6.7: Singular values and their corresponding weights for the sky distribution with an off-centre point source with an SNR of 5

TABLE 6.3: Recovered peak brightness for a sky distribution with SNR of 5.

| Singular Values | Recovered Peak Brightness $\pm 1.01201 \times 10^{-5}$ (Jy) | Recovered Peak Brightness (%) |
|-----------------|-------------------------------------------------------------|-------------------------------|
| First           | $5.20468 \times 10^{-5}$                                    | 14.9493                       |
| First 2         | $6.92185 \times 10^{-5}$                                    | 19.8815                       |
| First 3         | $9.60653 \times 10^{-5}$                                    | 27.5927                       |
| First 4         | $1.06715 \times 10^{-4}$                                    | 30.6515                       |
| First 12        | $2.07539 \times 10^{-4}$                                    | 59.6111                       |
| First 18        | $2.65995 \times 10^{-4}$                                    | 76.4012                       |
| First 20        | $3.14035 \times 10^{-4}$                                    | 90.1996                       |
| First 22        | $3.27424 \times 10^{-4}$                                    | 94.0454                       |
| First 23        | $3.42218 \times 10^{-4}$                                    | 98.2947                       |
| All 24          | $3.48155 \times 10^{-4}$                                    | 100                           |

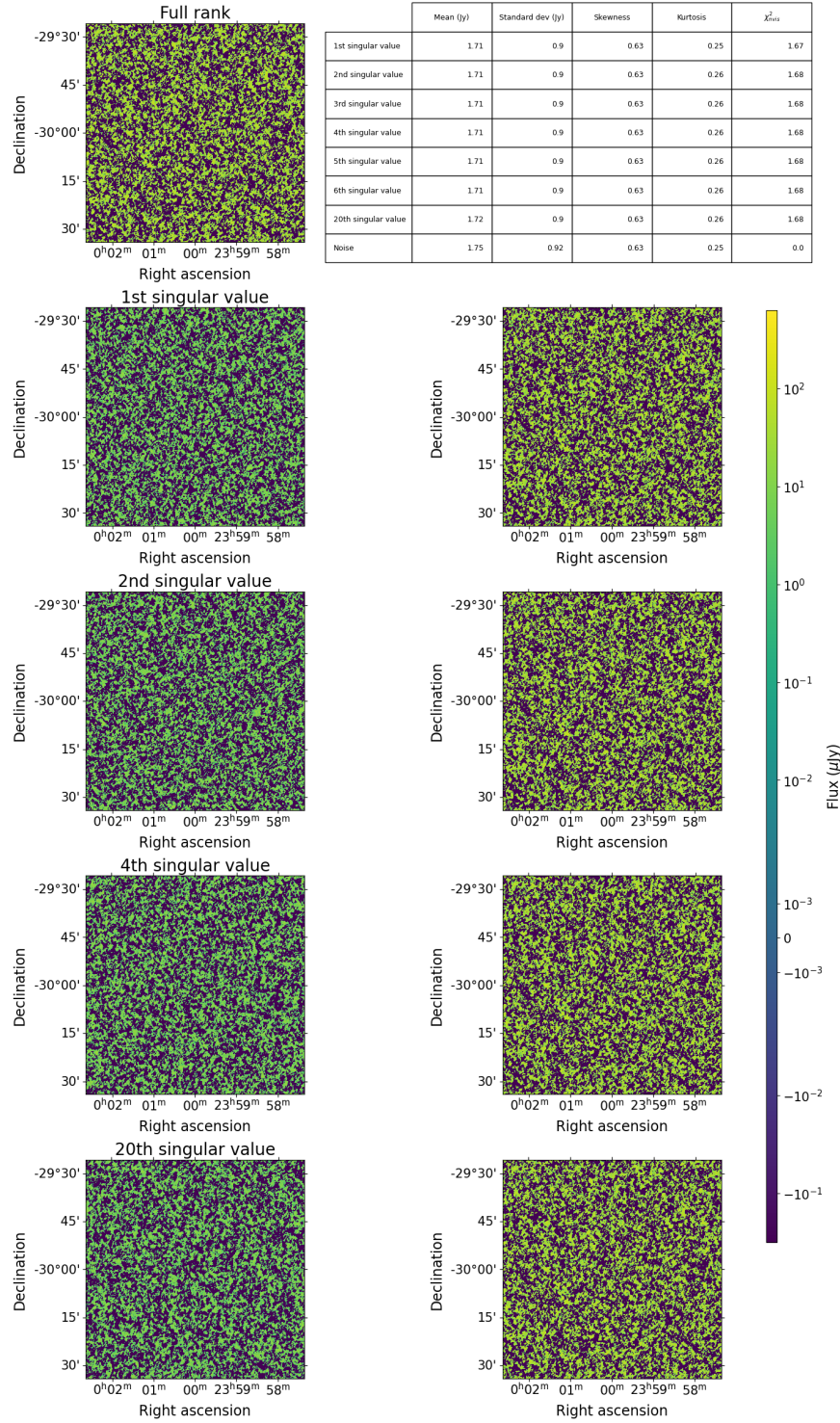


FIGURE 6.8: The left panel presents the original data along with the compressed data for the sky distribution with a point source at a distance of 0.4 deg with an SNR of 5. The right panel presents the residual statistics table along with the residual images.



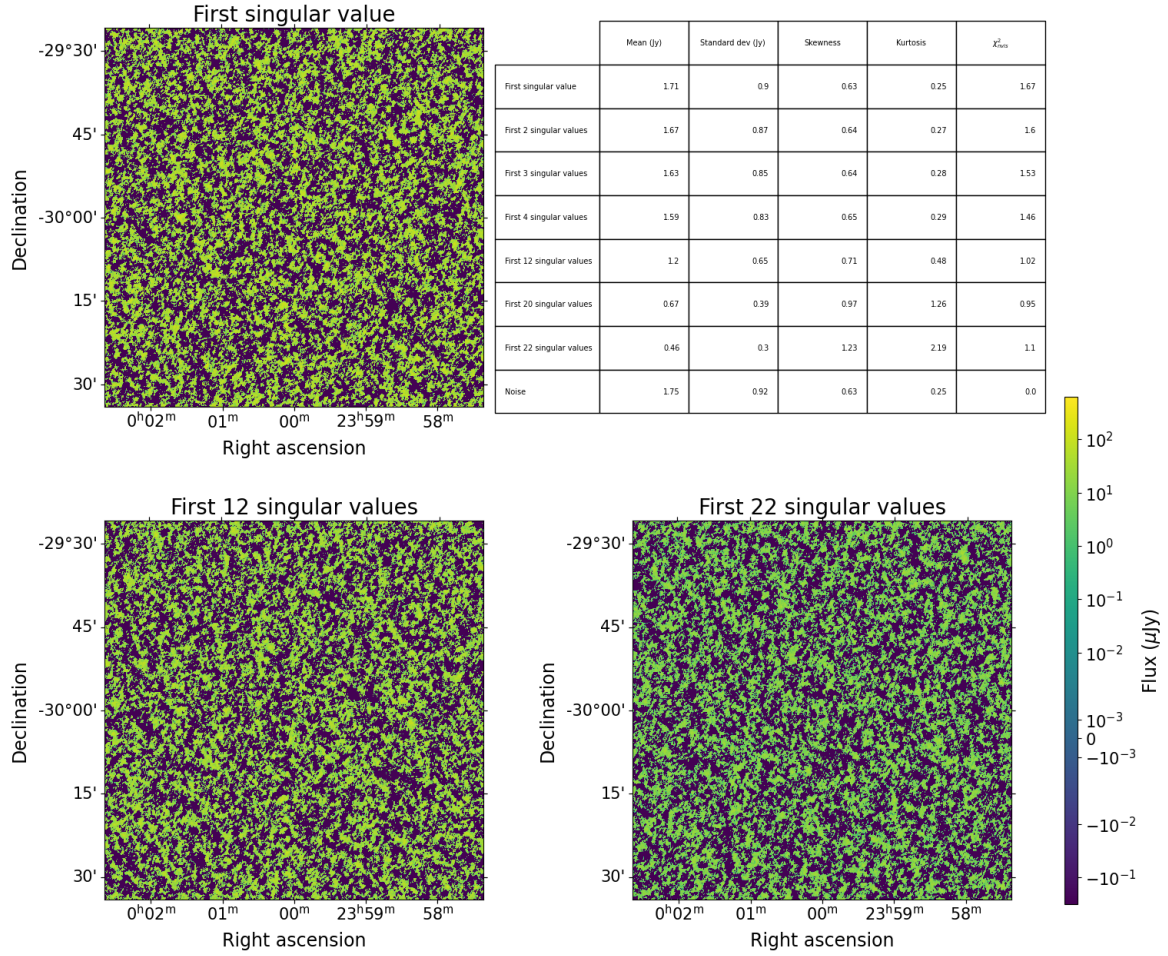


FIGURE 6.9: Residual statistics and images after reconstructing the data for the sky distribution with a point source at a distance of 0.4 deg with an SNR of 5 using the first  $n$  singular value.

#### 6.0.4 Source Direction - Declination of $-60$ deg

In this case, we focus on a 1 Jy point source located at a declination of  $-60$  deg. Figure 6.10 displays the mean weights of the singular values for different baselines. The singular value distribution is consistent with that shown in Figure 3.15, where the first singular value dominates, being significantly larger than subsequent singular values across all baseline lengths. This indicates that the first singular value effectively captures all data features, while higher-order singular values primarily capture noise features.



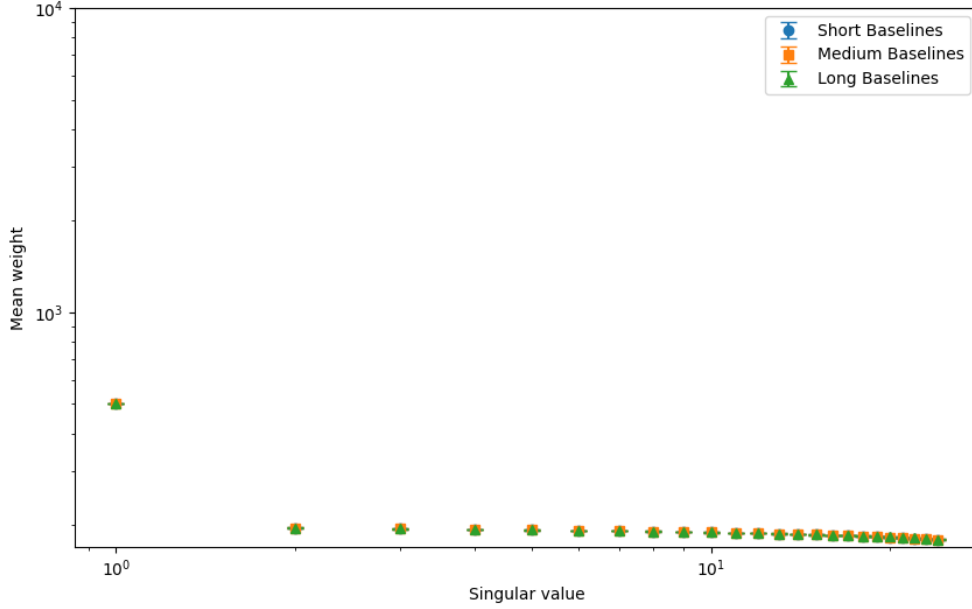


FIGURE 6.10: Singular values and their corresponding weights for the sky distribution with a point source 0.4 deg off-centre at a declination of  $-60$  deg.

Figure 6.11 confirms this observation. The first singular value successfully recovers  $9.99487 \times 10^{-1} \pm 2.54875 \times 10^{-6}$  Jy, corresponding to approximately 99.9877% of the original peak brightness ( $9.99609 \times 10^{-1}$  Jy). Subsequent singular values recover peak brightness below the RMS value of  $2.54875 \times 10^{-6}$  Jy, indicating that they primarily capture noise features.

Therefore, the first singular value is sufficient to effectively represent the data, as including additional singular values does not provide any significant benefit in terms of feature recovery or data fidelity.

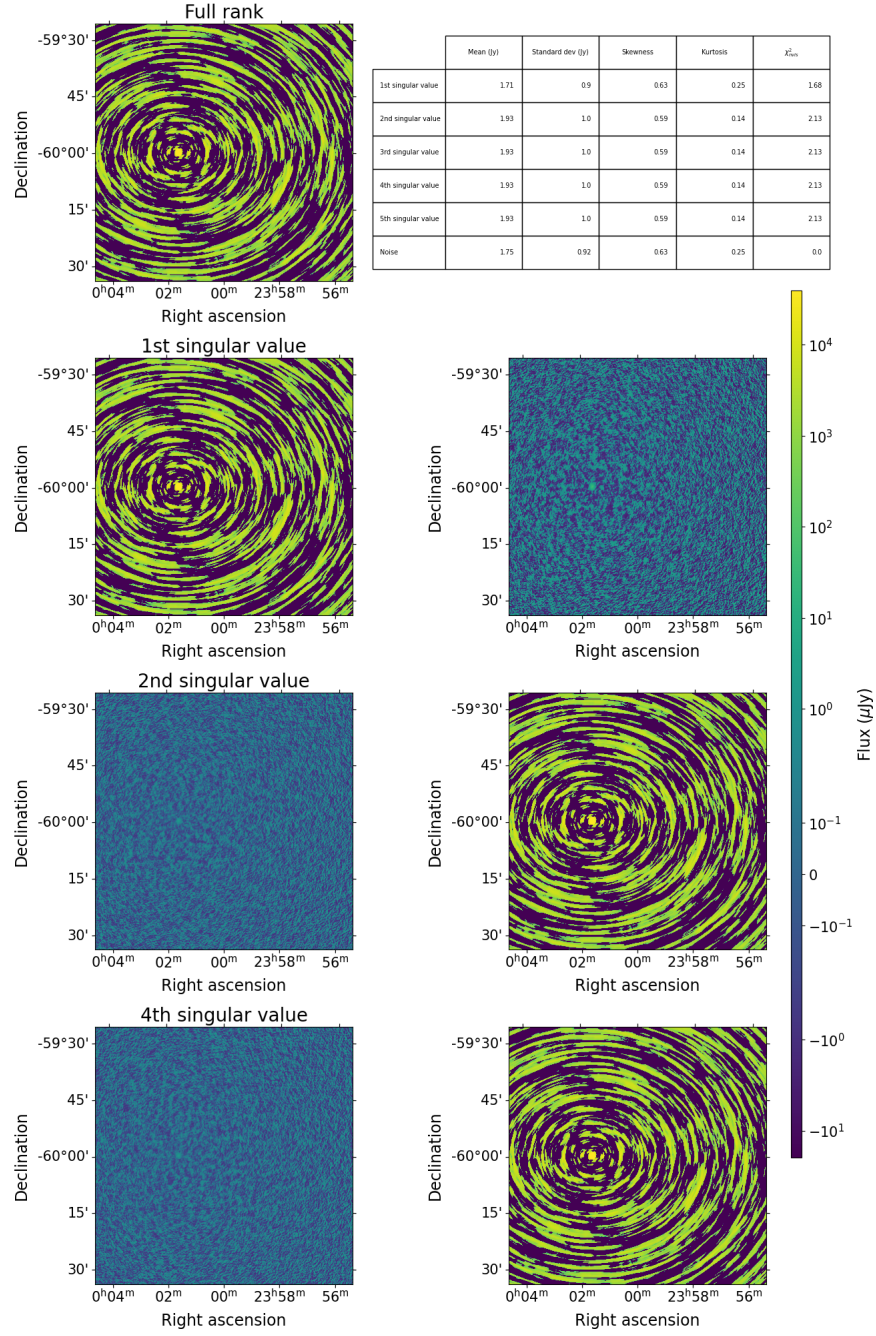


FIGURE 6.11: The left panel presents the original data along with the compressed data for the sky distribution with of a point source at a distance of 0.4 deg from the phase centre and a declination  $-60$  deg. The right panel presents the residual statistics table along with the residual images.

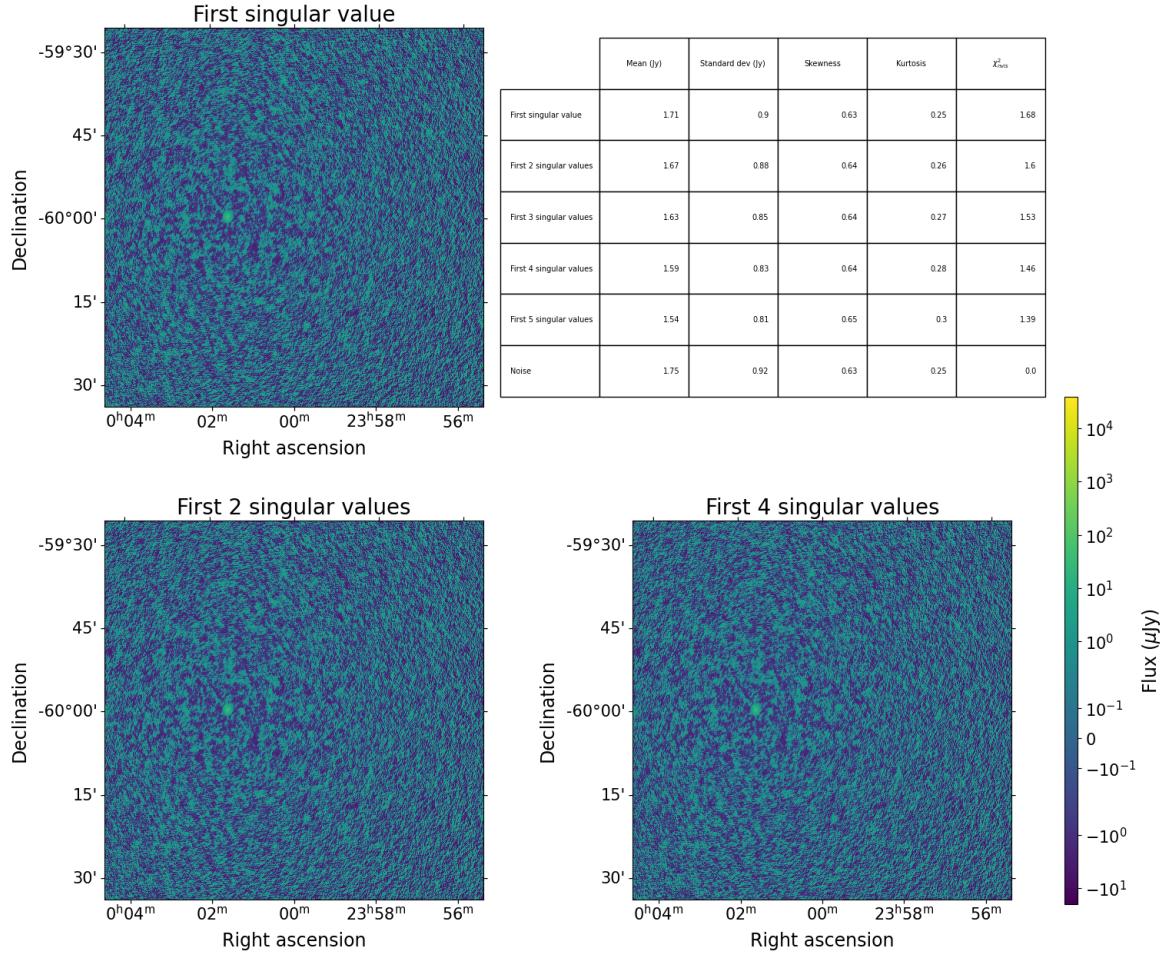


FIGURE 6.12: Residual statistics and images after reconstructing the data for the sky distribution with a point source at a distance of 0.4 deg from the phase centre and declination of  $-60^\circ$  using the first  $n$  singular value.

### 6.0.5 Source Direction - Declination of $0^\circ$

In this case, we focus on a 1 Jy point source at a declination of  $0^\circ$ . Figure 6.13 displays the mean weights of the singular values for different baselines. The first singular value dominates, being significantly larger than subsequent singular values across all baseline lengths. This indicates that the first singular value effectively captures all data features, while higher-order singular values primarily capture noise features.

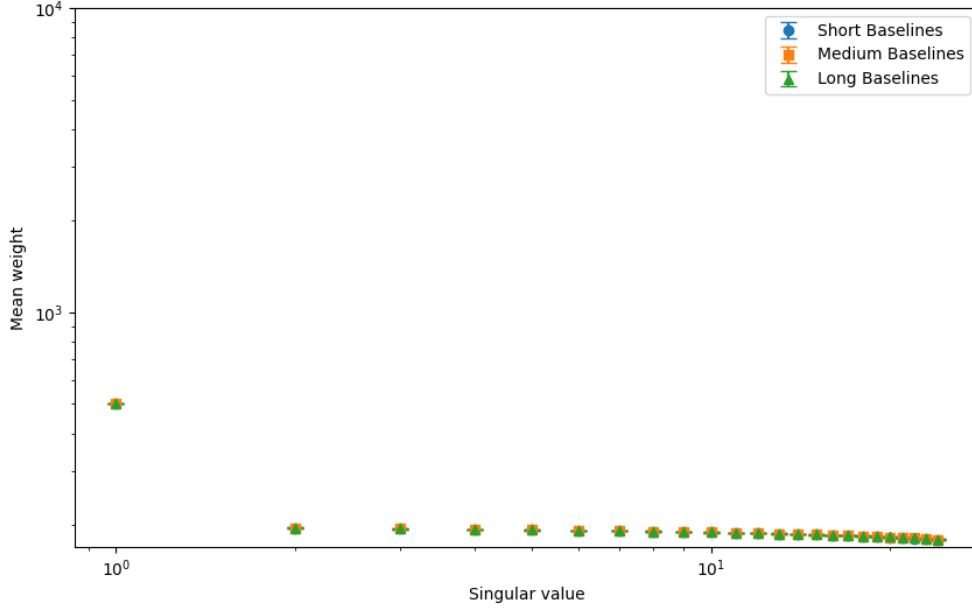


FIGURE 6.13: Singular values and their corresponding weights for the sky distribution with a point source 0.4 deg off-centre at a declination of 0 deg.

Figure 6.14 confirms this observation. The first singular value successfully recovers  $9.998\,03 \times 10^{-1} \pm 3.676\,42 \times 10^{-6}$  Jy, corresponding to approximately 99.9797% of the original peak brightness (1.000 01 Jy). Subsequent singular values recover peak brightness below the RMS value of  $2.548\,75 \times 10^{-6}$  Jy, indicating that they primarily capture noise features.

Therefore, the first singular value is sufficient to represent the data, as additional singular values do not provide any significant benefit in terms of feature recovery or data fidelity. This demonstrates that the source or field direction does not impact the BDSVD compression.

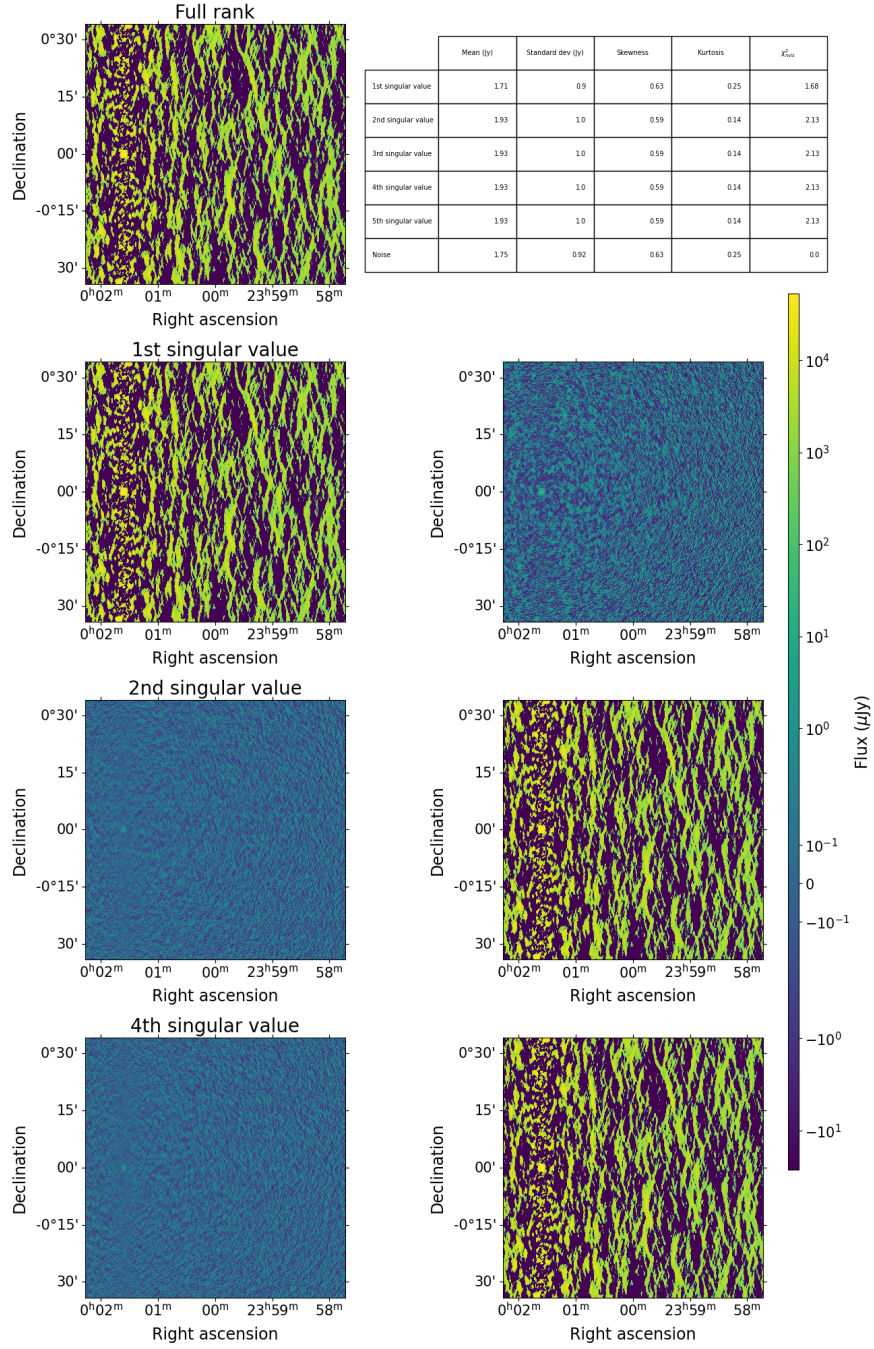


FIGURE 6.14: The left panel presents the original data along with the compressed data for the sky distribution with of a point source at a distance of 0.4 deg from the phase centre and a declination 0 deg. The right panel presents the residual statistics table along with the residual images.



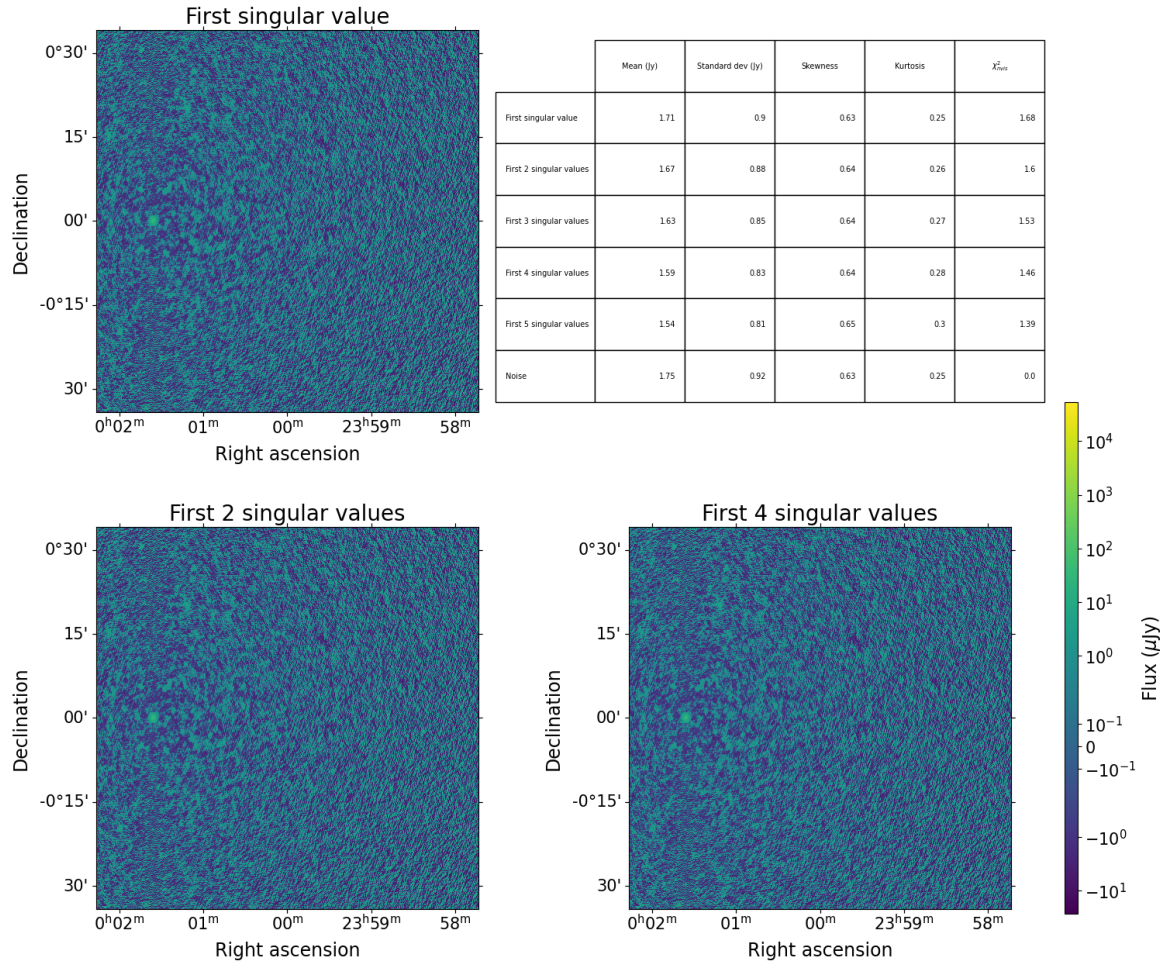


FIGURE 6.15: Residual statistics and images after reconstructing the data for the sky distribution with a point source at a distance of 0.4 deg from the phase centre and declination of 0 deg using the first  $n$  singular value.

# Bibliography

- [1] R. S. Booth et al. *MeerKAT Key Project Science, Specifications, and Proposals*. 2009.
- [2] S. Johnston et al. "Science with the Australian Square Kilometre Array Pathfinder". In: 24.4 (Dec. 2007), pp. 174–188.
- [3] D. R. DeBoer et al. "Australian SKA Pathfinder: A High-Dynamic Range Wide-Field of View Survey Telescope". In: *IEEE Proceedings* 97.8 (Aug. 2009), pp. 1507–1521.
- [4] M. P. van Haarlem et al. "LOFAR: The LOw-Frequency ARray". In: 556, A2 (Aug. 2013), A2.
- [5] C.J. Lonsdale et al. "The Murchison Widefield Array: Design Overview". In: *Proceedings of the IEEE* 97.8 (Aug. 2009), 1497–1506.
- [6] S. J. Tingay et al. "The Murchison Widefield Array: The Square Kilometre Array Precursor at Low Radio Frequencies". In: *Publications of the Astronomical Society of Australia* 30 (2013).
- [7] Richard Schilizzi et al. *The Square Kilometre Array: A Science Mega-Project in the Making, 1990-2012*. Springer, Jan. 2024. ISBN: 978-3-031-51373-2. DOI: [10.1007/978-3-031-51374-9](https://doi.org/10.1007/978-3-031-51374-9).
- [8] M Atemkeng et al. *Lossy Compression of Large-Scale Radio Interferometric Data*. 2023.
- [9] K.G. Jansky. "Directional Studies of Atmospherics at High Frequencies". In: *Proceedings of the Institute of Radio Engineers* 20.12 (1932), pp. 1920–1932.
- [10] Karl G. Jansky. "Radio Waves from Outside the Solar System". In: 132.3323 (July 1933), p. 66.
- [11] Grote Reber. "Cosmic Static." In: 100 (Nov. 1944), p. 279.
- [12] Grote Reber. "Cosmic Static". In: *Proceedings of the IRE* 28 (Feb. 1940), pp. 68–70.

- [13] Leonid Belov et al. *Handbook of RF, Microwave, and Millimeter-Wave Components*. Artech, 2012, pp. 26–39.
- [14] T.L Wilson. *Tools of Radio Astronomy*. Springer Berlin Heidelberg, 2009.
- [15] Jesse L. Greenstein et al. “The Quasi-Stellar Radio Sources 3C 48 and 3C 273.” In: 140 (July 1964), p. 1.
- [16] Gregory A. Shields. “A Brief History of Active Galactic Nuclei”. In: *Publications of the Astronomical Society of the Pacific* 111.760 (June 1999), 661–678.
- [17] A. Hewish et al. “Observation of a Rapidly Pulsating Radio Source”. In: 217.5130 (Feb. 1968), pp. 709–713.
- [18] A. A. Penzias et al. “A Measurement of Excess Antenna Temperature at 4080 Mc/s.” In: 142 (July 1965), pp. 419–421.
- [19] National Radio Astronomy Observatory. *Introduction to the Green Bank Telescope (GBT)*. <https://science.nrao.edu/facilities/gbt/gbt-basics/introduction-to-the-gbt>. Accessed: 2024-07-29.
- [20] Editors of Encyclopaedia Britannica. *Arecibo Observatory*. <https://www.britannica.com/topic/Arecibo-Observatory>. Accessed: 2024-07-30. 2024.
- [21] Peng Jiang et al. “Commissioning progress of the FAST”. In: *Science China Physics, Mechanics Astronomy* 62 (May 2019).
- [22] Special Astrophysical Observatory. *Cold Universe Part II: The Cold ISM*. <https://www.sao.ru/hq/CG/cold/part2.htm>. Accessed: 2024-07-29.
- [23] Albert A. Michelson. “I. On the application of interference methods to astronomical measurements”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 30.182 (1890), pp. 1–21.
- [24] A. A. Michelson et al. “Measurement of the Diameter of  $\alpha$  Orionis with the Interferometer.” In: 53 (May 1921), pp. 249–259.
- [25] VONBERG D. D RYLE M. “Solar Radiation on 175 Mc./s”. In: *Nature* 158 (1946), 339–34.
- [26] L. L. McCready et al. “Solar Radiation at Radio Frequencies and Its Relation to Sunspots”. In: *Proceedings of the Royal Society of London Series A* 190.1022 (Aug. 1947), pp. 357–375.



- [27] III Sullivan W. T. "Some highlights of Interferometry in early Radio Astronomy". In: *IAU Colloq. 131: Radio Interferometry. Theory, Techniques, and Applications*. Ed. by T. J. Cornwell et al. Vol. 19. Astronomical Society of the Pacific Conference Series. Jan. 1991, p. 132.
- [28] National Radio Astronomy Observatory. *Very Large Array (VLA)*. <https://public.nrao.edu/telescopes/vla/>. Accessed: 2024-07-29.
- [29] P. J. Napier et al. "The Very Large Array: design and performance of a modern synthesis radio telescope." In: *IEEE Proceedings* 71 (Nov. 1983), pp. 1295–1320.
- [30] T. P. Krichbaum et al. *SKA in VLBI: Impact on Studies of Small Scale Structures in Active Galactic Nuclei*. 1999.
- [31] T. P. Krichbaum et al. *SKA in VLBI: Impact on Studies of Small Scale Structures in Active Galactic Nuclei*. 1999.
- [32] Marcello Felli et al. *Very Long Baseline Interferometry: Techniques and Applications*. 1989th. Springer, 1989.
- [33] J. A. Zensus et al. *European VLBI Network: Present and Future*. 2015.
- [34] Anita Loots. "The African VLBI network project". In: *American Astronomical Society Meeting Abstracts #225*. Vol. 225. American Astronomical Society Meeting Abstracts. Jan. 2015, 126.02, p. 126.02.
- [35] L. I. Gurvits. *Space Radio Astronomy in the next 1000001 (binary) years*. 2012.
- [36] Leonid I. Gurvits. *Radio Interferometers Larger than Earth: Lessons Learned and Forward Look of Space VLBI*. 2018.
- [37] N. S. Kardashev et al. "'RadioAstron'-A telescope with a size of 300 000 km: Main parameters and first observational results". In: *Astronomy Reports* 57.3 (Mar. 2013), 153–194.
- [38] Thomas Young. "On the theory of light and colours". In: *Philosophical Transactions of the Royal Society of London* 92 (1801).
- [39] Max Born et al. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. 7th ed. Cambridge University Press, 1999.
- [40] A. Richard Thompson et al. *Interferometry and Synthesis in Radio Astronomy, 3rd Edition*. Springer, 2017.

- [41] T. D. Carozzi et al. "A generalized measurement equation and van Cittert-Zernike theorem for wide-field radio astronomical interferometry". In: *Monthly Notices of the Royal Astronomical Society* 395.3 (May 2009), 1558–1568.
- [42] JP Hamaker et al. "Understanding radio polarimetry. I. Mathematical foundations". In: *Astronomy and Astrophysics Supplement Series* 117.1 (1996).
- [43] O. M. Smirnov. "Revisiting the radio interferometer measurement equation". In: *Astronomy and Astrophysics* 527 (2011), A106.
- [44] O. M. Smirnov. "Revisiting the radio interferometer measurement equation". In: *Astronomy and Astrophysics* 527 (2011), A107.
- [45] O. M. Smirnov. "Revisiting the radio interferometer measurement equation". In: *Astronomy and Astrophysics* 527 (2011), A108.
- [46] A. H. Bridle et al. "Bandwidth and Time-Average Smearing". In: *Synthesis Imaging in Radio Astronomy II*. Ed. by G. B. Taylor et al. Vol. 180. Astronomical Society of the Pacific Conference Series. Jan. 1999, p. 371.
- [47] Marcellin Atemkeng. "Data compression, field of interest shaping and fast algorithms for direction-dependent deconvolution in radio interferometry". PhD thesis. Rhodes University, 2017, p. 8.
- [48] A.J. Kemball et al. *MeasurementSet definition version 2.0*. 2000.
- [49] Shou-Lin Wei et al. "Fine-grained distributed averaging for large-scale radio interferometric measurement sets". In: *Research in Astronomy and Astrophysics* 21.4 (May 2021), p. 080.
- [50] W. D. Cotton. "Special Problems in Imaging". In: *Synthesis Imaging in Radio Astronomy II*. Ed. by G. B. Taylor et al. Vol. 180. Astronomical Society of the Pacific Conference Series. Jan. 1999, p. 357.
- [51] W. D. Cotton. *Effects of Baseline Dependent Time Averaging of UV Data*. 2009.
- [52] S J Wijnholds et al. "Baseline-dependent averaging in radio interferometry". In: *Monthly Notices of the Royal Astronomical Society* 476.2 (2018), pp. 2029–2039.
- [53] Qing-Wen Deng et al. "Performance Evaluation of Baseline-dependent Averaging Based on Full-scale SKA1-LOW Simulation". In: *Research in Astronomy and Astrophysics* 22.4 (Mar. 2022), p. 045014.

- [54] M. T. Atemkeng et al. "Using baseline-dependent window functions for data compression and field-of-interest shaping in radio interferometry". In: *Monthly Notices of the Royal Astronomical Society* 462.3 (July 2016), 2542–2558.
- [55] M Atemkeng et al. "Baseline-dependent sampling and windowing for radio interferometry: data compression, field-of-interest shaping, and outer field suppression". In: *Monthly Notices of the Royal Astronomical Society* 477.4 (Mar. 2018), 4511–4523.
- [56] Karthik Rajeswaran et al. "Lossless Compression of SKA Data Sets". In: *Communications and Network* 05 (Jan. 2013), pp. 369–378.
- [57] K. Masui et al. "A compression scheme for radio data in high performance computing". In: *Astronomy and Computing* 12 (Sept. 2015), 181–190.
- [58] A. R. Offringa. "Compression of interferometric radio-astronomical data". In: *Astronomy and Astrophysics* 595 (Nov. 2016), A99.
- [59] William Press et al. *Numerical Recipes: The Art of Scientific Computing*. Third Edition. Cambridge University Press, 2007.
- [60] Gilbert Strang. *Linear algebra and its applications*. Fourth Edition. Thomson, Brooks/Cole, 2000, p. 367.
- [61] K.M. Aishwarya et al. "Lossy image compression using SVD coding algorithm". In: *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. 2016, pp. 1384–1389.
- [62] A. R. Offringa et al. "Post-correlation radio frequency interference classification methods". In: 405.1 (June 2010), pp. 155–167.
- [63] Daniel Shenon Briggs. "High fidelity deconvolution of moderately resolved sources". PhD thesis. New Mexico Institute of Mining and Technology, Jan. 1995.
- [64] S. Makhathini. "Advanced radio interferometric simulation and data reduction techniques". PhD thesis. Rhodes University, Apr. 2018.
- [65] RATT Team. *simms: A Tool for Simulating Measurement Sets*. Accessed: 2024. 2024. URL: <https://github.com/ratt-ru/simms>.

- [66] The CASA Team et al. “CASA, the Common Astronomy Software Applications for Radio Astronomy”. In: *Publications of the Astronomical Society of the Pacific* 134 (2022), p. 114501.
- [67] J. E. Noordam et al. “The MeqTrees software system and its use for third-generation calibration of radio interferometers”. In: 524, A61 (Dec. 2010), A61.
- [68] A. R. Offringa et al. “wsclean: an implementation of a fast, generic wide-field imager for radio astronomy”. In: *Monthly Notices of the Royal Astronomical Society* 444.1 (Aug. 2014), 606–619.
- [69] G. B. Taylor et al. *Synthesis Imaging in Radio Astronomy II*. Astronomical Society of the Pacific, 1999, pp. 171–186.
- [70] Mukundi Ramanyimi. *Visco: A tool for compressing radio interferometric data using Singular Value Decomposition (SVD) techniques*. GitHub repository. 2025. URL: <https://github.com/MukundiRams/visco>.