Low-dimensional data embedding for scalable astronomical imaging in the SKA telescope era

THÈSE N^{O} 8353 (2018)

PRÉSENTÉE LE 1ER JUIN 2018
À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE TRAITEMENT DES SIGNAUX 5
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Vijay KARTIK

acceptée sur proposition du jury:

Prof. D. N. A. Van De Ville, président du jury
Prof. J.-Ph. Thiran, Prof. Y. Wiaux, directeurs de thèse
Prof. M. A. T. Figueiredo, rapporteur
Prof. O. M. Smirnov, rapporteur
Prof. J.-P. Kneib, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

Acknowledgements

Putting together a doctoral thesis is a many-person job, and I would like to acknowledge the contributions of several people in enabling me to reach this point.

I am grateful to my thesis co-directors, Prof. Jean-Philippe Thiran and Prof. Yves Wiaux, for giving me the opportunity to pursue this research direction and for their guidance and fruitful scientific advice over the past years. Thanks to Prof. Wiaux for his continued scientific supervision and availability for discussions, separations in time and space notwithstanding.

I would also like to thank the members of my thesis examination jury, Prof. Dimitri Van De Ville, Prof. Mário A. T. Figueiredo, Prof. Oleg Smirnov, and Prof. Jean-Paul Kneib for their careful reading of my thesis manuscript, their constructive feedback, and also for what I believe was a perfectly enjoyable thesis examination experience. I want to thank Dr Jean-Marc Vesin and his fun course which sparked off some of the early ideas that finally ended up being a significant portion of this thesis – perhaps daydreaming about (and sometimes during) lectures is not such a bad idea after all.

I am deeply grateful to Dr Rafael Carrillo for his invaluable help in all possible problems that I got myself into. Rafa patiently put up with my inane questions and taught me the ropes of my research topics early on, and always had sympathy and advice for my wailing about life as a doctoral student. I also want to thank Drs Alex Onose and Arwa Dabbech for many long and useful discussions, about radio interferometry as well as finishing a thesis. Thanks to Drs Arwa Dabbech, Audrey Repetti, and Pierre-Antoine Thouvenin for going through my thesis manuscript in great detail and providing very useful corrections – it was a big help, and the current version is all the better for it.

Thanks to Rosie for assisting with the labyrinthine administrative side of things at the lab and for encouraging me to do crazy things outside of it, like singing in a choir. Thanks also to (quasi-) neighbours from the corridor: Sasan, Sibylle, Baran – for long walks and talks about the world outside the corridor. A big thank you to old colleagues and friends from the lab who made lunches, dinners and the hours in between very enjoyable: Anna, Alessandra, Carlos, Christina, Didrik, Gabriel, Jelena, Mário, Murat, Ricardo, Tom, Valentina – each one long since gone from the lab but never from happy memories. Thanks also to the current members of the lab – Christophe, Damien, Elena, Mina, and others for the short-lived escapes away from the computer during long days at work. Thanks to Silvia for irregular but eagerly awaited yak sessions to convince each other that there would indeed be light at the end of the tunnel.

My heartfelt gratitude to Marina for tirelessly lending a strong shoulder to cry on, an understanding ear to sigh the day's woes into, a gentle hand to hold on to in times both happy and sad – and for the countless words and deeds of encouragement and help throughout.

I give my sincerest thanks to my parents and my sisters, without whom I wouldn't be writing these words. Their belief in me surpasses my own, and for their resolute support and understanding I am forever in debt.

Lausanne, December 2017 S. V. K.

Abstract

Astronomy is one of the oldest sciences known to humanity. We have been studying celestial objects for millennia, and continue to peer deeper into space in our thirst for knowledge about our origins and the universe that surrounds us. Radio astronomy – observing celestial objects at radio frequencies – has helped push the boundaries on the kind of objects we can study. Indeed, some of the most important discoveries about the structure of our universe, like the cosmic microwave background, and entire classes of objects like quasars and pulsars, were made using radio astronomy. Radio interferometers are telescopes made of multiple antennas spread over a distance. Signals detected at different antennas are combined to provide images with much higher resolution and sensitivity than with a traditional single-dish radio telescope. The Square Kilometre Array (SKA) is one such radio interferometer, with plans to have antennas separated by as much as 3000 km. In its quest for ever-higher resolution and ever-wider coverage of the sky, the SKA heralds a data explosion, with an expected acquisition rate of 5 terabits per second. The high data rate fed into the pipeline can be handled with a two-pronged approach – (i) scalable, parallel imaging algorithms that fully utilize the latest computing technologies like accelerators and distributed clusters, and (ii) dimensionality reduction methods that embed the high-dimensional telescope data to much smaller sizes without losing information and guaranteeing accurate recovery of the images, thereby enabling imaging methods to scale to big data sizes and alleviating heavy loads on pipeline buffers without compromising on the science goals of the SKA.

In this thesis we propose fast and robust dimensionality reduction methods that embed data to very low sizes while preserving information present in the original data. These methods are presented in the context of compressed sensing theory and related signal recovery techniques. The effectiveness of the reduction methods is illustrated by coupling them with advanced convex optimization algorithms to solve a sparse recovery problem. Images thus reconstructed from extremely low-sized embedded data are shown to have quality comparable to those obtained from full data without any reduction. Comparisons with other standard ‘data compression’ techniques in radio interferometry (like averaging) show a clear advantage in using our methods which provide higher quality images from much lower data sizes. We confirm these claims on both synthetic data simulating SKA data patterns as well as actual telescope data from a state-of-the-art radio interferometer. Additionally, imaging with reduced data is shown to have a lighter computational load – smaller memory footprint owing to the size and faster iterative image recovery owing to the fast embedding.

Extensions to the work presented in this thesis are already underway. We propose an ‘on-line’ version of our reduction methods that work on blocks of data and thus can be applied on-the-fly on data as they are being acquired by telescopes in real-time. This is of immediate interest to the SKA where large buffers in the data acquisition pipeline are very expensive and thus undesirable. Some directions to be probed in the immediate future are in transient imaging, and imaging hyperspectral data to test computational load while in a high resolution, multi-frequency setting.

Key words: big data, compressed sensing, convex optimization, dimensionality reduction, inverse problems, radio interferometry, Square Kilometre Array (SKA)

Résumé

L'astronomie est l'une des plus anciennes sciences. Nous étudions les objets célestes depuis des millénaires et nous continuons, aujourd'hui encore, à regarder plus profondément dans l'espace dans notre soif de connaissance de nos origines et celles de l'univers qui nous entoure. La radioastronomie – l'observation des objets célestes dans les fréquences radios – nous aide à repousser les limites du type d'objet que nous sommes capables d'étudier. Certaines des découvertes les plus importantes au sujet de la structure de notre univers, comme celle du fond diffus cosmologique (cosmic microwave background), ainsi que des classes entières d'objets célestes, comme les quasars et les pulsars, ont été faites grâce à la radioastronomie. Un interféromètre radio est un télescope qui se compose de plusieures antennes qui s'étalent sur de grandes distances. Les signaux détectés par chacune des antennes sont combinés afin de produire des images de plus haute résolution et avec une meilleure sensibilité que des images produites par des télescopes radio traditionels avec une seule parabole (ou que par des télescopes optiques). Le Square Kilometre Array (SKA) est un interféromètre radio qui sera construit dans les prochaines années et dont les antennes seront séparées par des distances considérables, jusqu'à 3000 km. Dans sa quête pour atteindre les plus hautes résolutions et les plus larges couvertures du ciel, le SKA présage d'une explosion de la quantité des données à traiter, avec un taux d'acquisition de données de 5 terabits par seconde. Avec cet imminent déluge d'information, la recherche en traitement des données est en pleine effervescence. Il est possible d'adresser les défis posés par ce taux élevé de production de données avec une stratégie reposant sur deux axes – (i) la parallélisation des algorithmes d'imagerie en utilisant les dernières avancées technologique en informatique telles que des accélérateurs ou des clusters distribués et (ii) la réduction de la dimensionnalité, qui permettra de représenter les données de manière très compacte sans perdre d'information, ce qui garantit la reconstruction correcte de l'image. Cette réduction de la dimensionnalité permet aux algorithmes de s'appliquer à des données de grande taille tout en atténuant les charges importantes sur les tampons du pipeline de traitement des données – sans faire de compromis sur les buts scientifiques du SKA.

Dans cette thèse, nous proposons des méthodes rapides et robustes pour réduire la dimensionnalité des données à des tailles très petites en gardant l'information contenue dans les données originales. Ces méthodes sont présentées dans le contexte de la théorie de l'acquisition comprimée et des techniques de récupération de signaux liées. L'efficacité de ces méthodes de réduction de la dimensionnalité est démontrée par les résultats obtenus par des algorithmes d'optimisation convexe qui les utilisent pour récupérer des images à partir des donnés réduites. La qualité de ces images est comparable à celle des images reconstruites à partir des données originales. Nous comparons nos méthodes avec d'autre méthodes de référence de réduction de la dimensionnalité (comme le calcul de la moyenne), et nous montrons qu'elle présentent un avantage évident. Nous confirmons ces assértions par des expériences sur des données simulées imitant les données du SKA ainsi que sur des données réelles acquises par un interféromètre radio de dernière génération. En outre, les données réduites entraînent une charge informatique moins importante – grâce á une réduction de taille dans la mémoire et à une accélération de la récupération des images.

Nous proposons également un certain nombre d'extensions pour les méthodes proposées dans cette thèse, notamment une version 'on-line' de nos méthodes qui fonctionne sur des blocs de données et qui peut donc être appliquée aux données pendant leur acquisition. Cela peut être particulièrement utile pour le SKA où il y a un risque d'avoir besoin de tampons d'une grande taille, ce qui pourrait entraîner des côuts importants. Quelques autres directions pour les prochaines étapes de recherche concernent les sujets d'imagerie des objets éphémères et d'imagerie hyperspectrale dans un cadre haute-résolution et multi-fréquence.

Mots clefs : big data (mégadonnées), acquisition comprimée, optimisation convexe, réduction de la dimensionnalité, problèmes inverses, interférométrie radio, Square Kilometre Array (SKA)

Contents

List of figures

List of acronyms

2-D	Two-dimensional
3-D	Three-dimensional
ADMM	Alternating direction method of multipliers
ASTRON	Netherlands institute for radio astronomy
BP	Basis pursuit
BPDN	Basis pursuit denoising
CASA	Common astronomy software applications
CS	Compressed sensing
DDE	Direction-dependent effect
DIE	Direction-independent effect
DR	Dynamic range
HPC	High performance computing
LOFAR	Low frequency array
MP	Matching pursuit
NRAO	National radio astronomy observatory
NSP	Null space property
PCA	Principal component analysis
PDFB	Primal-dual forward-backward
PSF	Point spread function
RI	Radio interferometry
RIME	Radio interferometry measurement equation

RIP	Restricted isometry property
SDMM	Simultaneous direction method of multipliers
SKA	Square kilometre array
SNR	Signal-to-noise ratio
SVD	Singular value decomposition
VLA	Karl G. Jansky Very large array

Introduction

Humankind has been curious about the skies above since time immemorial. From the very first attempts to chart different stars and planets, millennia of advancements have brought us to the present day where astronomy is a rich and diverse field of study, with discoveries of celestial objects covering practically every part of the electromagnetic spectrum.

The range of ‘radio frequencies’ of the spectrum are particularly interesting, because some of the most active and energetic objects in the universe emit radiation in these frequencies. ‘Radio’ astronomy, thus, provides an exciting window to study such astronomical objects. Indeed, radio astronomy has furthered collective human knowledge about the universe, and we have been able to draw conclusions about various celestial processes and objects that would otherwise have gone undetected with optical astronomy.

Of course, as in any field of science, the quest for higher precision and more detailed observations in radio astronomy has been the driving factor for many innovations. Chief among them is the development of radio *interferometers* – telescopes made by combining multiple radio antennas spread over large distances. Combining antennas in this manner effectively works like one huge telescope with very high precision – this is also seen by the fact that radio interferometers provide higher-resolution images than other telescopes.

To achieve ever increasing precision and detail, proposed next-generation radio interferometers are expected to map the sky with unprecedented resolution. One such ambitious project is the upcoming SKA telescope, which plans to produce gigapixel images of the sky with extremely fine detail and quality. This, however, is accompanied with an explosion of the data volume acquired by such telescopes. Extremely high data rates (of the order of terabits per second) are expected to swamp currently available signal acquisition pipelines, and a flurry of research activity is now focused on designing techniques to ingest this high data rate and process the incoming signals to provide images.

Motivation

The research work described in this thesis was undertaken primarily to address the issues arising from the imminent explosion of data from next-generation radio interferometers.

Currently employed imaging techniques are ill-equipped to perform effectively in a big data regime, and are consequently expected to struggle to scale with increasing data acquisition. Efforts at designing newer, scalable imaging algorithms are well underway. While these ongoing efforts address scalability in terms of parallel and distributed imaging methods, the research described here aims to complement these efforts by approaching the challenge of scalable imaging from an entirely different angle – that of reducing data dimensionality. The driving motivation behind this is the assumption that reducing data size *before* feeding the data to an imaging algorithm will tackle the root of the scalability issue. By reducing data size to 'manageable' quantities, current and future imaging techniques can be expected to perform efficiently and scale with increasing data size, since dimensionality reduction would render this increase invisible while preserving the information available in the data.

The final goal is, evidently, to be able to combine these new developments, in both scalable imaging techniques and dimensionality reduction techniques, and to forge an imaging solution that (i) continues to remain computationally efficient in a big data regime thanks to dimensionality reduction, and (ii) takes advantage of a scalable imaging algorithm, either through parallelization or faster design to exploit computing hardware resources.

Main contributions

The main contributions of the research presented in this thesis are briefly summarized here:

- Designing and developing dimensionality reduction methods that appropriately fit in the compressed sensing framework of signal reconstruction, while remaining computationally efficient and practical to implement [Kartik et al., 2017a,b].
- Proof-of-concept of a new model for dimensionality reduction, shown on data acquired by currently operating radio interferometers, including real-world image reconstruction results and comparisons with the prevalent methods in the field. Proposal of an 'on-line' dimensionality reduction method for real-time application during data acquisition [Kartik et al., 2017c].
- Designing and evaluating different dimensionality reduction methods, including spatial frequency thresholding and various standard random projection techniques. Evaluation included quality comparison of images reconstructed with reduced data with those from established image reconstruction methods on full data, and quantifying the computational benefits afforded by these methods for large-scale data [Kartik et al., 2015].
- Developing a scalable imaging algorithm for radio interferometry data. This involved an extensive study of a set of convex optimization algorithms for image reconstruction, particularly with respect to the effects of increasing data size on the computational efficiency of these methods [Carrillo et al., 2015].

Thesis outline

This thesis is structured as follows:

- The current chapter provides the context for the research work described in this thesis. We present a bird's eye view of the setting in which the current work is based, and outline the motivation for the undertaken research direction.
- Chapter 1 provides a self-contained description of radio interferometry and the big data challenges for imaging pipelines in next-generation interferometers. Chapter 2 describes radio-interferometric imaging, and also includes a survey of the state of the art. In chapter 3 we introduce the theoretical background of compressed sensing and convex optimization techniques, highlighting the relevance of the two topics in improving imaging techniques for big data in radio interferometry.
- Chapter 4 presents a detailed discussion of different dimensionality reduction methods, their use in analyzing large amounts of data, and their role in handling radio interferometry data for imaging algorithms. We move from general dimensionality reduction methods to specific dimension embeddings relevant to compressed sensing-based image reconstruction, and observe their applicability to radio interferometry data.
- Chapter 5 introduces a major contribution of the research work described in this thesis – a novel dimensionality reduction method which is shown to be a practical and fast way to handle large dimensional radio interferometry data. Experimental results on large-scale simulations show that the proposed method drastically reduces computational requirements for imaging algorithms, while maintaining the image reconstruction quality to a high degree.
- Chapter 6 describes the results of dimensionality reduction on large-scale data currently being acquired from radio interferometers in the United States and South Africa. The encouraging results on previously described simulations are confirmed on real observations, and provide validation of the proposed dimensionality reduction.
- Chapter 7 presents conclusions about the methods developed as described in chapters 5 and 6, highlighting their suitability in scalable imaging techniques for next-generation radio interferometers. We conclude by indicating avenues for future work in dimensionality reduction, especially by taking into account antenna characteristics and calibration effects. We outline concrete experiments to be undertaken in the immediate future on imaging problems with immediate astrophysical value – like imaging transient sources.

  The advances presented in the coming chapters, along with the indicated next steps for widening the scope of our method to handle more relevant, large-scale use cases will, we hope, provide convincing arguments for the inclusion of the reduction method proposed in this thesis into the data processing pipeline of next-generation radio interferometers like the SKA telescope.

1 Radio astronomy and the SKA

1.1 The role of radio astronomy

Radio astronomy is the study of celestial objects that emit radiation in ‘radio frequencies’, i.e., in the range of 3 kHz to 300 GHz. Fig. 1.1 shows the position of ‘radio’ frequencies in the electromagnetic spectrum, relative to sizes in terms of both wavelength and frequency of other common physical objects. Radio astronomy has several advantages over other techniques (e.g., optical astronomy) – especially when it comes to observing celestial objects that (i) have very fine angular detail, or need to be localized in space with high angular precision, or (ii) exhibit phenomena only in radio frequencies, while being relatively ‘silent’ in optical and other ranges of the electromagnetic spectrum. Radio astronomy techniques can produce angular accuracy of the order of 10^{-3} arcsec[1] for absolute positions of astronomical objects, as opposed to current optical measurements (from earth) which have a resolution of $\approx$ 0.05 arcsec. Advances in the technology available for telescopes in optical, infrared, X-ray and other frequencies continue to enable ever-finer resolutions, and are expected to eventually reach values comparable to those obtained in radio astronomy today. However, radio frequencies will continue to be important sources of astronomical information since they allow us to observe objects and processes that do not emit radiation in other parts of the electromagnetic spectrum, or are blocked en route to earth by galactic dust clouds [Thompson et al., 2001].

Radio emissions originating in the far reaches of the universe tend to manifest as very weak signals at receiving stations on earth. Most antennas identify a *signal* as voltage fluctuations; these signals are usually indistinguishable from Gaussian random noise. Signal characteristics are usually constant over the time scales of typical radio astronomy observations,[2] and are assumed to be stationary and ergodic. The power spectrum of the majority of the signal power (which is in the form of *continuum radiation*) varies very slowly with frequency, and may also be considered to be constant over the receiving bandwidth of the antenna (in most cases). The nature of the slow variation of the power spectrum can provide clues about the composition

---

[1] 1 degree = 3600 arcseconds.

[2] Observation durations are of the order of minutes or hours.

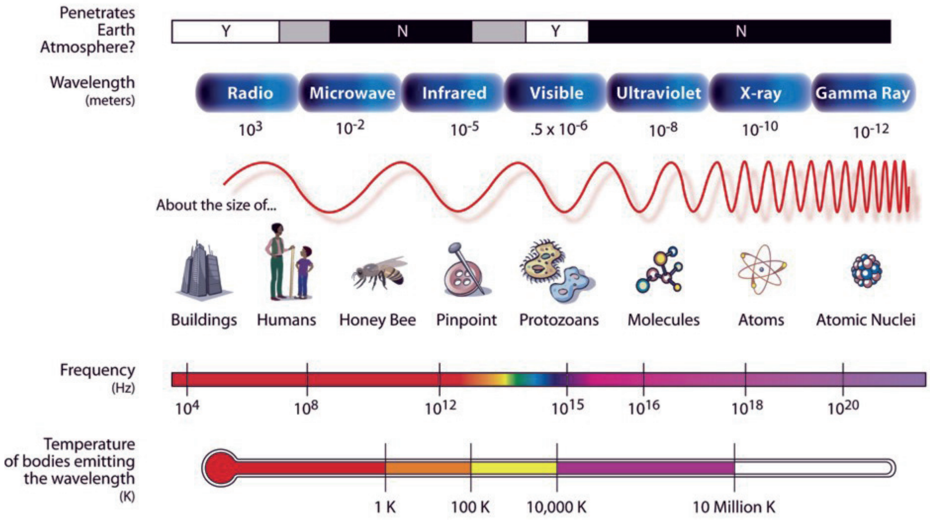


Figure 1.1 – The electromagnetic spectrum showing the relative position of 'radio' frequencies. [Source: NASA public domain image, CC-BY-SA 3.0]

of the underlying astronomical object, like electron densities and magnetic fields in certain radio galaxies, or electron-ion collisions in nebulae. In contrast, *spectral line radiation* shows a discontinuous power spectrum, with narrow peaks at specific frequencies corresponding to underlying atomic and molecular processes. A very well-known and well-studied example is the '21-cm spectral line' belonging to neutral atomic hydrogen. The 21-cm line and its Doppler-shifted variants are used to glean information about galactic structure and dynamics. It is widely studied since the signal reaches earth with minimal obstruction from galactic dust, which normally blocks radiation in optical frequencies, as mentioned earlier.

The power from continuum and spectral line radiation emitted by different astronomical objects is measured as a spectral flux density, expressed in watts per square meter per hertz. The unit is the jansky (Jy), named after the pioneering radio astronomer Karl G. Jansky.[3] $1\,\text{Jy} = 10^{-26}\,\text{Wm}^{-2}\text{Hz}^{-1}$. The spectral flux density received by a radio antenna per unit solid angle subtended by an astronomical object (or 'source') under observation gives us the *intensity* of the radiation emitted from that source, and it is this intensity that is used to generate images of the radio emission, from which further astrophysical conclusions may be drawn.

The power received from radio sources is usually very small, with a correspondingly weak signal. Although single-dish telescopes are in use for many cases in radio astronomy, they need to be extremely large to collect enough emission to attain a reasonable signal-to-noise ratio (SNR). In addition, since angular resolution is proportional to the diameter of the dish (in units of the wavelength being observed), such radio telescope dishes would need to be orders

---

[3] Jansky birthed the field of radio astronomy by first discovering radio waves, which he correctly concluded to have an extraterrestrial origin (seemingly emanating from the centre of the Milky Way).

of magnitude larger than optical telescope dishes for comparable precision. These limitations were overcome by the development of *radio interferometry*, a sub-field of radio astronomy which involves observing radio emissions through an array of radio antennas trained at a portion of the sky.

1.2 Radio interferometry

Radio interferometers contain several antennas spread over a large area; individual recordings at each antenna are correlated with recordings at other antennas of the array to obtain a combined signal by *interfering* pairs of signals. Grossly speaking, this array of antennas functions as an equivalent single dish with a diameter equal to the separation between the two furthest antennas of the array. This allows observations of the sky with extremely high angular resolutions (increasing with the largest separation of antennas) – this is essential to measure the positions of sources with enough accuracy to identify them with corresponding measurements from optical and other frequencies – in their intensities, polarizations and spectra. A higher SNR is achieved by collecting data through multiple antennas; the achieved sensitivity increasing with the number of measurements and the total acquisition surface.

Interferometry as a way of observing astronomical sources was first proposed by Michelson and others in the early 20$^{\text{th}}$ century, and the first observation targets were stars in the optical frequencies. This was quickly adapted for observations in radio frequencies, and in 1946 Ryle and Vonberg successfully used radio interferometry to corroborate observations made years earlier by Jansky and other astronomers, thereby providing the proof-of-concept and the groundwork for the development of the field of radio interferometry [Ryle and Vonberg, 1946, 1948]. McCready et al. [1947] independently made the first radio interferometric observations of the sun in 1945, using a 'sea interferometer', by observing the sun directly and through its reflection in the sea. Ryle developed the technique of radio interferometry further, most notably with the introduction of phase switching in 1952 which enabled radio astronomers to detect very weak signals in the presence of instrument noise that was typically several orders of magnitude higher [Ryle, 1952]. Phase switching was then gradually replaced by what is known today as the *correlator*, which effectively performs the same function. Another equally important advance in radio astronomy was the development of 'aperture synthesis', put forward by Blythe [1957], Ryle and Hewish [1960]. Aperture synthesis forms the cornerstone of radio interferometric imaging, effectively emulating a large single-dish radio telescope by the simple process of moving around antennas in different configurations to cover the same area and then intelligently combining the measurements obtained at each configuration. A more detailed description of aperture synthesis is presented in section 2.1. Ryle and Hewish shared the 1974 Nobel Prize in Physics for this and other contributions to radio astronomy.

These developments provided a boost to the field and several large arrays were commissioned and developed using a combination of these new techniques with the latest advances in electronics and computing. A surge in new, miniaturized electronics allowed receiver equipment to be installed locally at each individual antenna, which consequently enabled antennas to spread further and further apart. All the radio interferometers built in the late 1960s and onwards exploit these advances, and have proven to perform really well and produce interesting science results. The One-Mile Radio Telescope in Cambridge, UK, operated by the Mullard Radio Astronomy Observatory (MRAO), produced the first detailed images of the structures of radio sources with very strong emissions – galaxies like Cygnus A and Cassiopeia A. Subsequently, other telescopes came into operation – like the Green Bank Telescope (GBT) in West Virginia, USA, and the Very Large Array (VLA) in New Mexico, USA, operated by the National Radio Astronomy Observatory (NRAO), the Westerbork Synthesis Radio Telescope (WSRT) in The Netherlands, and the Giant Metre-Wave Radio Telescope (GMRT) in India – bringing even higher resolution images and faster scanning of the sky at different frequencies. This trend continues today, with some of these telescopes receiving upgrades over the years; a good example is the VLA, which has been producing data since 1974, and was recently upgraded to have increased sensitivity over an extended observation frequency range of 1 to 50 GHz [Perley et al., 2009].[4]

1.3 The SKA era

The field of radio interferometry has advanced considerably – over the last six decades many novel contributions have been made, leading to the current high-fidelity imaging and high resolution mapping of the sky available to radio astronomers. The quest for ever-higher angular resolution in radio astronomy continues to fuel development in the design and use of bigger and more powerful radio interferometers. Most notably, in recent years, this has led to a unified effort towards constructing an unprecedented radio telescope – an array spread over the globe, centred mainly in South Africa and Australia. On completion, this telescope is expected to contain several thousand dishes (in Africa) and hundreds of thousands of individual antennas (in Australia), to give an effective collecting area of one square kilometre. This array is called, unsurprisingly, the Square Kilometre Array (SKA). Initial phases of the SKA design and development are well under way, and construction of the partial array has already begun. Fig. 1.2 shows an artist's depiction of a small portion of the SKA. Pathfinders and precursors to the SKA are currently being built and have already entered their data-taking phase. In South Africa, the MeerKAT telescope has been acquiring data since July 2016, and is acting as a technology demonstrator for the SKA, which will subsume MeerKAT antennas into its mid-frequency component in the first operation phase. Similarly, the Australian SKA Pathfinder (ASKAP) has been running its 'Early Science Program' since October 2016, with encouraging first results supporting the validity of SKA science projects for low-frequency observations. The ambitious scale of the SKA project means that, on completion, the SKA is expected to be the largest radio telescope in the world, observing the skies over a wide frequency range from 50 MHz to 25 GHz. A wide field of view combined with high sensitivity receivers is expected to allow the SKA to survey large parts of the sky at a time at a high rate.

---

[4] The upgraded VLA is now named the Karl G. Jansky Very Large Array.

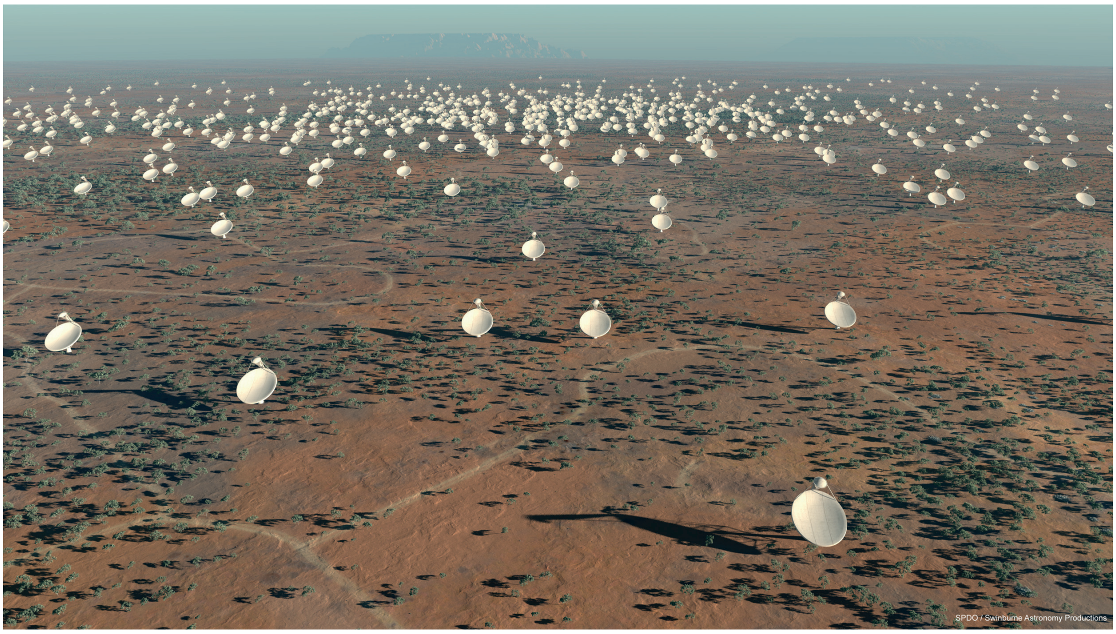


Figure 1.2 – SKA dishes in The Karoo, South Africa – an artist's depiction. [Source: SKA Project Development Office/Swinburne Astronomy Productions, CC-BY-SA 3.0]

1.3.1 SKA goals: science and technology

The SKA has very ambitious science goals. These science goals are divided into 'Key Science Projects' (KSPs) that touch practically every aspect of the universe, ranging from details of its origins and the birth of the very first galaxies (the 'Dark Ages' KSP), to the evolution of galaxies and dark energy, to studies of black holes and pulsars to test general relativity and theories of gravity, to planet formation and searching for extraterrestrial intelligence (the 'Cradle of Life' KSP), and many other unknowns, including the eventual death of our universe. The findings from the SKA will have far reaching consequences in fields as diverse as astronomy, cosmology, fundamental physics, and astrobiology.

Achieving these science goals means that the SKA will need to be a cutting edge instrument, with simultaneous abilities to have wide fields of view, high resolution and extreme sensitivity, acquiring hyperspectral data at an extremely high rate. To achieve such a high level of performance across these disparate figures of merit, the SKA will also need to push the envelope for engineering and technology advances. The scientific challenges will be addressed through appropriate application of cutting edge technology to enable that the final goals are met within budget constraints. The SKA will take advantage of the developments in high-speed data transmission with fibre optics, which would enable finer timing control over signals acquired over large distances. High speed digital signal processing solutions and dedicated integrated circuits and systems-on-chips are expected to be used for quick signal analysis, increasing survey speeds. Very recent improvements in computing, like high-frequency CPUs, memory chips and high-capacity storage solutions will be used to make possible an even larger search space for the SKA.

1.3.2 Data handling challenges

As seen in the previous section, the unprecedented science and technology objectives of next-generation radio telescopes can only be achieved by acquiring corresponding data at an explosive rate. The current estimate for the SKA is a data rate of up to 5 terabits per second [Broekema et al., 2015]. This high data rate is a direct consequence of the thousands of antennas collecting data over very wide fields of view across several gigahertz of bandwidth. The science data processing (SDP) for this deluge of data is a challenge, owing to the limitations on the computational resources, in terms of both raw numerical processing of the data, as well as providing large real-time buffers to enable unhindered data acquisition (i.e., CPU, RAM and storage). The cost of the data processing components is a major deciding factor in the choice of technical solutions for each science goal. The cost of adopting a particular technology can be the difference between rendering a particular data processing method infeasible for the SKA and it being a major contributor to achieving faster and better science data processing. Consequently, appropriate processing techniques need to be designed and evaluated in this context in order to be able to handle these imminent challenges. The main 'product' of the SKA is to be high-resolution, high-dynamic range images reconstructed from the raw data it acquires. So, a large proportion of the imminent data challenges falls squarely on the radio-interferometric imaging techniques that need to ingest a high data rate and produce high-quality output images.

2 Radio-interferometric imaging

In this chapter we discuss the general principles of recovering images from telescope data. A more detailed background of the acquisition of these data is warranted at this point to enable a deeper treatment of the imaging methods described later in the chapter. A common way of collecting cosmic signals as interferometer data is through aperture synthesis. The low- and mid-frequency collecting portions of the SKA are also aperture synthesis array designs. Therefore, our discussion on radio-interferometric imaging will be focussed on aperture synthesis imaging of radio interferometer data.

2.1 Aperture synthesis

Aperture synthesis was introduced by Ryle and Hewish [1960], building on the concept developed by Blythe [1957]. The basic principle states that we can effectively make observations of the sky as if we were using a single telescope with a large collecting area, or aperture – using much smaller antennas – either by placing stationary antennas spread over a large area, or by moving individual antennas to cover a given area over time. Radio interferometers lend themselves particularly readily to aperture synthesis because of the widespread use of electronics at individual receivers in antennas that handle the recording of signal timing information (like phase) and transport the digitized signal to a central location for correlation in software at a later stage. This is made possible because of the long wavelengths of the incoming signals handled by radio interferometers. As a counter example, aperture synthesis is less suited for optical wavelengths which cannot make use of standard electronics equipment to transport signals and correlate them in software. Performing optical interference and ensuring reliable transport of optical signals is an expensive endeavour, and affordable technology to make this feasible has only matured in recent years to the point where aperture synthesis became a possibility, whereas this has been used in radio interferometry since the 1960s.

At the time that aperture synthesis was proposed, the relationship between the intensity distribution of a source and the response of an interferometer observing that source was already well established. The Wiener-Khinchin relation states that, for weak-sense stationary signals,

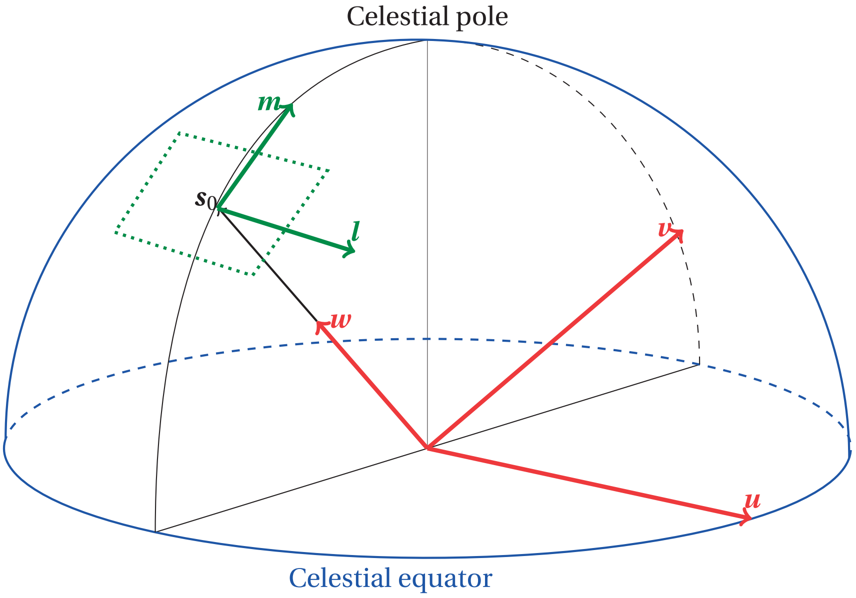


Figure 2.1 – Illustration explaining the coordinate system used in radio interferometry. The uv plane lies normal to the direction pointing at the source, $\boldsymbol{s}_0$. The corresponding lm plane lies tangent to the celestial sphere at the point where $\boldsymbol{s}_0$ meets the sphere. Note that the $\boldsymbol{w}$ axis is also taken to point at the source.

the power spectrum is the Fourier transform of the autocorrelation of the signal [Wiener, 1930]. We can already see how this statement may link to the acquisition of radio interferometer data – which are essentially cross correlation components of the intensity distribution of radio sources. This Fourier relationship did not go unnoticed, and was first used by McCready et al. [1947] in the analysis of their pioneering radio interferometric observations of the sun. In the following sections, we introduce the process of observing a source through a radio interferometer, in a general setting, and show that, with an appropriately chosen coordinate system, the output of the radio interferometer can be seen to be the Fourier transform of the power spectrum of the signal received from the source.

2.2 Basics of radio interferometer measurements

The antennas of an array point to the same portion of the sky, shown in Fig. 2.1 on the so-called celestial sphere.[1] A radio source, in general, may be spread out over a part of the sky, as shown by the outline on the celestial sphere. We assume that the source is so far away that the emissions received at the antennas may be assumed to be in the form of plane waves. In a typical observation, the antennas point towards the centre of the source, defined as the 'phase-tracking centre', given by a unit vector $\boldsymbol{s}_0 \in \mathbb{R}^3$. The receiving systems of the array measure the correlation between incoming plane waves recorded at each antenna pair. This measurement

---

[1] An imaginary sphere concentric to earth and with arbitrarily large radius, on which all celestial objects can be projected and consequently observed on the inner surface.

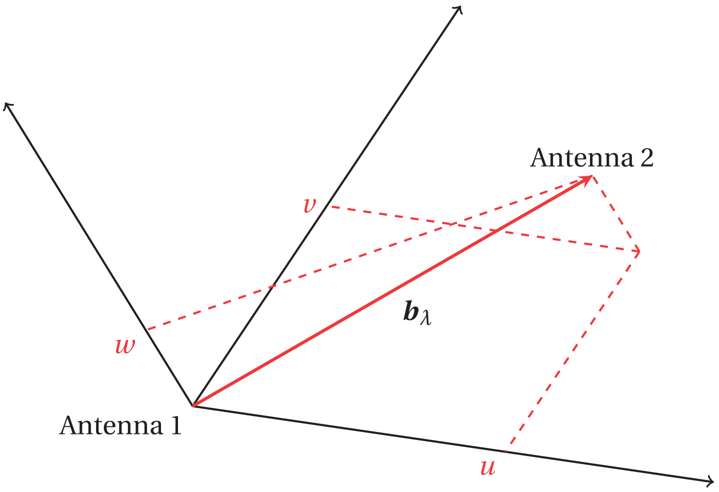


Figure 2.2 – Illustration of a baseline $\boldsymbol{b}_\lambda$ in 3-D space, and its components u, v, and w on the $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ axes, respectively. $|\boldsymbol{b}_\lambda|$, u, v, and w are all measured in multiples of the wavelength of the received radio emission.

can be seen as an interference fringe pattern, assumed to be due to a hypothetical point source at the phase-tracking centre, in the direction $\boldsymbol{s}_0$. This interference pattern varies in amplitude and phase as a function of the antenna separations, and is expressed by a complex *visibility* function. It is this visibility function that needs to be measured in order to recover the intensity distribution of the source. These complex measurements are commonly referred to simply as *visibilities*, and lie on a plane of observation normal to the direction of the source, $\boldsymbol{s}_0$. We define this plane with the coordinate axes $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. As marked in Fig. 2.1, $\boldsymbol{w}$ points in the direction of the source, i.e. in $\boldsymbol{s}_0$; $\boldsymbol{u}$ and $\boldsymbol{v}$ point to the 'east' and 'north', respectively, as seen from a plane through the origin, the phase-tracking centre and the pole. The antenna separations that result in complex visibilities are defined by displacement vectors referred to as *baselines*. Each baseline can be broken down into components along these coordinate axes $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. These components are then given by the values u, v, and w, as illustrated in Fig. 2.2. The observation plane thus containing the u and v components is called the 'uv plane', and the measured complex visibilities form a two-dimensional function on the uv plane. Baseline magnitudes are commonly measured in wavelength of the received radio emission,[2] λ, and thus baseline components u,v, and w are usually in multiples of λ. We note that to collect as much information as possible about the intensity distribution of a source, it is desirable to have the highest number of possible baselines. Each baseline is represented by a point in the uv plane. Thus the goal is to have measurements from a large number of points in the uv plane, or a comprehensive uv *coverage*. Fig. 2.3 illustrates a uv coverage from 254 antennas simulating a mid-frequency SKA array. Aperture synthesis attempts to achieve an improved uv coverage by generating multiple baselines, by placing antennas at different locations, either keeping them stationary or moving them over a given area. Even

---

[2] More precisely, the wavelength corresponding to the central frequency of the bandwidth of the receiving system.

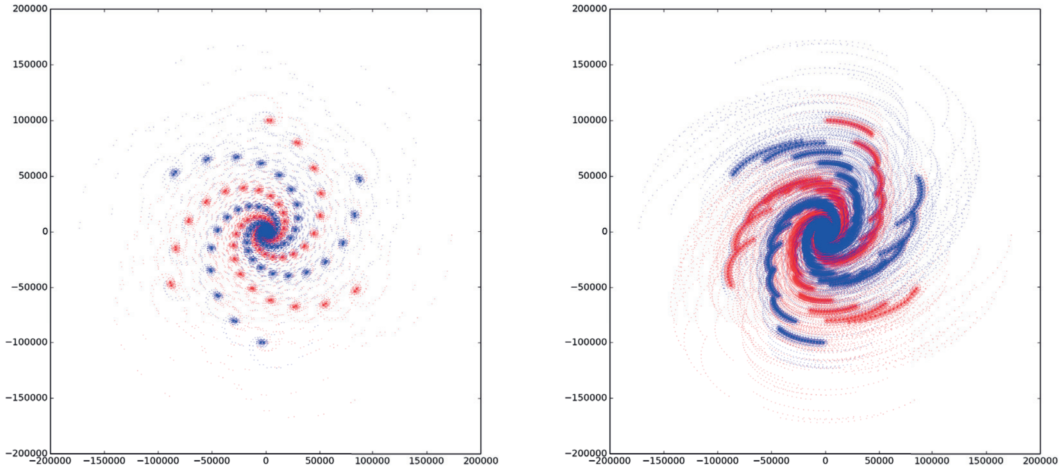


Figure 2.3 – Illustration of uv coverage with simulated SKA antenna configuration for mid-frequencies. Left: Short observation duration. Right: Longer observation duration, showing the effects of earth rotation synthesis. Note that each baseline shown in blue at position (u, v) has a corresponding conjugate baseline shown in red at $(-u, -v)$.

in the case where moving antennas was possible (if inconvenient), the number of possible antenna configurations was limited, which impacts the uv coverage. In order to collect more data through a more 'complete' coverage of the uv plane, Ryle and Hewish [1960] proposed to use the earth's rotation as a way to embellish any existing antenna configuration. The rotation of the earth causes variations in baseline positions relative to the source being observed, which manifest as distinct uv points. The method, which quickly gained popularity as a way to approximate a radio telescope with large aperture, is called earth rotation aperture synthesis. We can see the effect of the earth's rotation in Fig. 2.3, where both uv coverages are generated with the same antenna configuration, but with different observation times. The longer observation time is able to make use of the earth's rotation and populate a larger area of the uv plane, thus providing more data that is then used to recover the intensity distribution image of the source. We also see in Fig. 2.3 that the uv coverage with earth rotation aperture synthesis contains elliptical tracks. Each track is made of uv points that correspond to the same pair of antennas but at different different positions relative to the line of sight during the earth's rotation, resulting in different baselines over time.

Analogous to the uv plane for the measurements, we can also define an image plane that contains the underlying signal (i.e., the intensity distribution of the radio source being observed). The image plane is parallel to the uv plane, and lies tangent to the celestial sphere at the phase-tracking centre. The corresponding coordinate axes are labelled $(\boldsymbol{l}, \boldsymbol{m}, \boldsymbol{n})$, and are the direction cosines measured with respect to the $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{w}$ axes respectively. The phase-tracking centre initially given by $\boldsymbol{s}_0$ acts as the origin of the source intensity distribution, giving us the centre of the corresponding image.

The relation between the image on the lm plane and the corresponding visibility measurements from the uv plane is the basis for all further discussions of radio-interferometric imaging in this thesis, and is presented in the next section.

2.3 Measurement model

The aperture synthesis process of interfering signals received at different antennas relies on the van Cittert-Zernike theorem which established a connection between the interference pattern and the distribution of the source intensity [Zernike, 1938, Thompson et al., 2001]. Although synthesis imaging was being performed routinely since the 1960s, the importance of the van Cittert-Zernike theorem was only realized much later with developments in signal processing techniques that established a Fourier relationship between the correlator output and the underlying source intensity distribution.

The origin of this relationship, in the context of the different factors affecting the measurement of the source, including telescope characteristics, observation parameters, and physical constraints among others, can be discussed using the concept of a measurement equation [Hamaker et al., 1996, Sault et al., 1996, Smirnov, 2011a]. It is explained in this subsection along with the description of the solution for the equation. The solution is constructed as a convex optimization method, which lends itself to several computationally efficient algorithms developed in a more general sense.

2.3.1 Radio interferometry measurement equation (RIME)

We build our measurement model using the simplified single antenna pair observation set-up illustrated in Fig. 2.1. The argument can be extended over multiple such pairs that form an interferometer array.

A part of the radio source in the direction $\boldsymbol{s} = \boldsymbol{s}_0 + \boldsymbol{\sigma} \in \mathbb{R}^3$ subtends a solid angle $\mathrm{d}\Omega$. Recalling our definition of the spectral flux density in section 1.1, we can see that, over the observing bandwidth $\Delta\nu$ of the receiving system, this part of the source contributes $\frac{1}{2}\boldsymbol{A}(\boldsymbol{\sigma})\boldsymbol{x}(\boldsymbol{\sigma})\Delta\nu \mathrm{d}\Omega$ of power at each antenna. Here $\boldsymbol{A}(\boldsymbol{\sigma})$ denotes the effective area of an antenna and $\boldsymbol{x}(\boldsymbol{\sigma})$ is the unknown underlying source intensity distribution.

For an antenna separation given by the baseline $\boldsymbol{b}_\lambda$ (in wavelengths as mentioned in section 2.2), the extra distance (delay) for incoming emission to one antenna is given by $\boldsymbol{b}_\lambda \cdot \boldsymbol{s} = \boldsymbol{b}_\lambda \cdot (\boldsymbol{s}_0 + \boldsymbol{\sigma})$. The complex visibilities measured at the output of the receiving system are obtained by correlating the voltages generated at each antenna, producing an interference fringe pattern. This fringe pattern depends on the delay, and appears in the correlator output along with the received power, in the form

$$\boldsymbol{r}(\boldsymbol{b}_\lambda, \boldsymbol{s}_0) = \Delta\nu \int_{4\pi} \boldsymbol{A}(\boldsymbol{\sigma})\boldsymbol{x}(\boldsymbol{\sigma}) \cos[(2\pi \boldsymbol{b}_\lambda \cdot (\boldsymbol{s}_0 + \boldsymbol{\sigma})] \,\mathrm{d}\Omega. \tag{2.1}$$

The correlator output gives us our desired complex visibility for that baseline, which we define here as a complex value y with magnitude $|y|$ and phase ϕ, giving

$$y(\boldsymbol{b}_\lambda) = |y(\boldsymbol{b}_\lambda)|\,\mathrm{e}^{\mathrm{i}\phi_\lambda} \tag{2.2}$$

$$= \int_{4\pi} \boldsymbol{A}(\boldsymbol{\sigma})\boldsymbol{x}(\boldsymbol{\sigma})e^{-2\pi\mathrm{i}\,\boldsymbol{b}_\lambda\cdot\boldsymbol{\sigma}}\,\mathrm{d}\Omega. \tag{2.3}$$

We now use the coordinate system introduced in section 2.1 to reformulate this visibility computation. The baseline $\boldsymbol{b}_\lambda$ has components u, v, and w, and on the celestial sphere, the direction vectors $\boldsymbol{s}$, $\boldsymbol{s}_0$, and the distance $\boldsymbol{\sigma}$ can be expressed in terms of components l, m, and n such that

$$\boldsymbol{b}_\lambda\cdot\boldsymbol{\sigma} = \boldsymbol{b}_\lambda\cdot\boldsymbol{s} - \boldsymbol{b}_\lambda\cdot\boldsymbol{s}_0 \tag{2.4}$$

$$= (ul + vm + wn) - w, \tag{2.5}$$

and $\mathrm{d}\Omega = \frac{\mathrm{d}l\,\mathrm{d}m}{n}$, for $n^2 = 1 - l^2 - m^2$. Rewriting equation (2.3) using these coordinates, we get the complete *radio interferometry measurement equation* (RIME) –

$$y(u,v,w) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{n}\cdot\boldsymbol{A}(l,m)\boldsymbol{x}(l,m)\,\mathrm{e}^{-2\pi\mathrm{i}(ul+vm+w(n-1))}\,\mathrm{d}l\,\mathrm{d}m, \tag{2.6}$$

where the integral limits $(-\infty,+\infty)$ are made possible by the fact that $\boldsymbol{A}(l,m)\boldsymbol{x}(l,m)$ quickly decays to negligible values outside the field of interest which contains the source, due to the receiving characteristics of the antennas.

We can obtain a simplified form of equation (2.6) by making the assumption that the antenna array could lie entirely on the observation plane normal to the $\boldsymbol{w}$ direction, so that the w-component of the baselines would be zero. In this case, the integral would simplify to

$$y(u,v,w) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{n}\cdot\boldsymbol{A}(l,m)\boldsymbol{x}(l,m)\,\mathrm{e}^{-2\pi\mathrm{i}(ul+vm)}\,\mathrm{d}l\,\mathrm{d}m, \tag{2.7}$$

which we can immediately identify as the two-dimensional (2-D) Fourier transform relation. We can thus deduce from equation (2.7) that the complex visibilities $\boldsymbol{y}$ are indeed Fourier components of the unknown source intensity distribution $\boldsymbol{x}$. The fact that visibilities are complex can be explained here through the observation that the intensity distribution, while real, is not (usually) symmetric. Each spatial (l,m) point of the image plane thus has a corresponding (u,v) point in spatial frequency through the Fourier transform.

While the assumption that all baselines lie in a plane normal to the $\boldsymbol{w}$ direction rarely holds, we can continue to preserve the Fourier transform relation between the intensity distribution $\boldsymbol{x}$ and the measured visibilities $\boldsymbol{y}$ through a variety of workarounds. Firstly, if we consider small fields of view (i.e., where l,m are small), we then see that the term $w(n-1)$ in equation (2.6) becomes negligible, and we reach equation (2.7). This is actually quite common as many radio observations are performed on a limited portion of the sky. An approach to surveying large portions of the sky while maintaining the 2-D Fourier transform is to scan the sky as a mosaic of multiple independent observations, each limited to a separate, smaller field of view and then combining them together at a later stage with an appropriate change of coordinates. Finally, observations of sources directly overhead (or reasonably close to being overhead) also result in negligible w-components. Equation (2.7), which holds under some simplifying assumptions on l, m, and w among others, is an instance of the van Cittert-Zernike theorem.

The simplifications and assumptions made here introduce errors in the measurement model, and consequently in any signal recovered from these measurements. Some errors additionally arise due to inappropriate handling of direction-dependent effects (DDEs) of the receiving system. For example, the effective antenna collecting areas and response functions depend on the pointing direction. The heterogeneous makeup of the ionosphere also contributes in varying degrees depending on the direction. The non-negligible w-component of baselines is a DDE that needs to be accounted for during imaging. DDEs become more pronounced with wider fields of view and higher target sensitivities, and it is therefore crucial to correct for them in next-generation radio interferometers. Yet other errors are due to direction-independent effects (DIEs) from antenna electronics to propagation effects. DIEs can, in principle, be accounted for by correcting the visibilities measured with DIEs through *calibration* procedures [Smirnov, 2011b]. DDEs can also be (partially) handled by appropriately including them in the measurement model used for imaging [Smirnov, 2011c]. For example, as mentioned in section 2.3.3 that follows, general w-terms can be included in rows of the degridding matrix $\mathbf{G}$ as part of the measurement equation. Of course, DIEs and DDEs can only be calibrated and accounted for when they are detectable and quantifiable. The presence of unknown DIEs and DDEs, or failing to model or correct these effects through calibration, leads to errors in signal recovery. The recovered signals indicate the intensity distribution of the radio source, and provide information about underlying astrophysical phenomena. The signal thus provides a ‘map’ or image of the source, and so the signal recovery process is called mapping, or, as it is referred to in this thesis – radio-interferometric *imaging*.

2.3.2 Visibility sampling in the uv plane

The 2-D Fourier transform relation between the complex visibilities and the unknown image is a welcome boon to synthesis imaging methods. This is mainly due to the availability of the fast Fourier transform (FFT) algorithm that allows us to quickly compute the discrete Fourier and inverse Fourier components of a given function [Cooley and Tukey, 1965]. Indeed, in the years before the FFT algorithm became prevalent, 2-D Fourier transforms used to be calculated by hand in each dimension successively [Bracewell, 1956], and consequent imaging was – to put it mildly – an unenviable task.

While the advent of the FFT algorithm and increasingly powerful computing resources greatly helped the progress of imaging methods, the measurement model needed to be tweaked to fit the limitations of these faster numerical computing methods. The FFT computes *discrete*

Fourier transforms over a regular grid, whereas a typical uv coverage contains points at arbitrarily continuous locations in Fourier space. So, while great computational advantages can be had if the visibility data could be laid on a regular grid, each baseline of an interferometer array provides one sample of the visibility function at a location that may or may not conform to the grid. Interferometer arrays are indeed designed with antenna spacings meant to provide a desired sampling of the visibility function, but all (u, v) points cannot be made to lie on a regular grid[3] since baselines follow elliptical tracks on the uv plane. Therefore, the visibility sample value at each regular grid point is obtained by interpolating over the measured visibilities that fall in a pre-defined region/cell centred at the grid point[4] – this process is known as *gridding*. One simple way would be to average all visibilities in a cell and use the average value at the corresponding grid point. More sophisticated interpolation methods are usually employed, which apply a *weighting* on the visibilities before gridding. Different weighting schemes have been developed over the years, each with its own merits and demerits [Briggs, 1995, Boone, 2012]: 'natural' weighting uses inverse noise variances of the visibility measurements as weights thus maximizing sensitivity; 'uniform' weights, on the other hand, are inversely proportional to the number of measurements in a given cell, thus maximizing resolution but losing sensitivity; 'robust' weighting takes a middle path, guaranteeing reasonable levels of both resolution and sensitivity; some other weighting schemes have also been developed more recently, which propose a middle path through multi-step [Boone, 2012] and adaptive [Yatawatta, 2014] approaches. It is clear, then, that we should aim to have a reasonable number of measurements in a given cell, to enable a good guess at the visibility value at the corresponding grid point. Ideally, we would also like to have measurements in each cell over the entire uv plane, since cells with no measurements indicate an incomplete coverage of the Fourier space, and consequently, incomplete knowledge of the underlying visibility function. This is, however, not always possible, and it is this incomplete uv coverage that plays a major role in the measurement model, and consequently, the non-trivial process of recovering the underlying intensity distribution.

2.3.3 RIME in matrix form

The linear measurement model, as it appears through a Fourier transform relation in equation (2.7), is discretized for computational purposes through sampling methods mentioned in section 2.3.2. The original measurement model accounting for 'continuous' visibilities, however, can be approximated in matrix form as

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{n}, \tag{2.8}$$

where $\boldsymbol{x} \in \mathbb{C}^N$ is the (vectorized) image to be recovered and $\boldsymbol{y} \in \mathbb{C}^M$ the visibilities vector, $\boldsymbol{n} \in \mathbb{C}^M$ being the noise in the measurements. The 'measurement operator' $\boldsymbol{\Phi} \in \mathbb{C}^{M\times N}$ covers

---

[3] Unless an extremely fine grid is chosen, in which case the computational cost approaches that of a discrete-time Fourier transform anyway.

[4] This is usually done through a convolution and hence cannot be called an interpolation, strictly speaking.

the linear relation between the signal and the continuous visibilities, and is given by

$$\mathbf{\Phi} = \mathsf{G}\overline{\mathsf{F}}\mathsf{D}_{\mathrm{R}}\mathsf{Z}, \tag{2.9}$$

where Z is the zero-padding of the image needed to compute the 2-D discrete Fourier transform of $\boldsymbol{x}$ on a finer sampling grid in the Fourier domain and $\overline{\mathsf{F}}$ the 2-D discrete Fourier transform operator in the oversampled case. G is a convolution interpolation operator to map from the discrete frequency grid to the continuous uv plane – this is essentially the inverse of the gridding process mentioned in section 2.3.2. Each row of G contains an interpolation kernel of compact support [Fessler and Sutton, 2003]. Note that DDEs can be accounted for in this model by allowing general interpolation kernels in each row of G. D_{R}, the grid correction term, is a diagonal matrix to implement the reciprocal of the inverse Fourier transform of the interpolation kernel used in G, to undo the effects of the convolution by the interpolation kernel in the spatial frequency domain. For brevity, we also define the combined operator $\overline{\mathsf{Z}} = \mathsf{D}_{\mathrm{R}}\mathsf{Z}$. The measurement operator defined in equation (2.9) is then equivalently given by

$$\mathbf{\Phi} = \mathsf{G}\overline{\mathsf{F}}\overline{\mathsf{Z}} \quad \in \mathbb{C}^{M\times N}. \tag{2.10}$$

Equation (2.8) thus presents a *forward model*, capturing how visibilities are measured from an underlying intensity distribution image. Here we focus on the *inverse problem*, i.e., the problem of tracing back through the forward model to recover the underlying image from the measurements recorded at the observer's end. Solving the inverse problem is thus equivalent to obtaining the image $\boldsymbol{x}$ from the incomplete information available in the data $\boldsymbol{y}$ as given in equation (2.8). Holes in the uv coverage of the data and the presence of noise in the measurements provide an incomplete view of the visibility function, thus rendering the linear system ill-posed, even though there are typically many more measurements available compared to the image size, i.e., $M \gg N$. The incompleteness of the uv coverage is at the core of our inability to analytically arrive at the true underlying intensity distribution image. This ill-posed inverse problem is thus the cornerstone of the imaging challenge, and ultimately, the science objectives in radio interferometry.

2.4 Radio-interferometric imaging

Imaging is a key component of the processing pipeline in radio interferometers. The final science goals, from the astronomer's perspective, are achieved through the images reconstructed from the acquired data. Therefore it is crucial that the images recovered from the data have high fidelity, high dynamic range, high sensitivity, and high resolution. These features of the images help the astronomer study details of the astrophysical phenomenon captured in them. It is also essential, then, that the imaging process be faithful to the information contained in the underlying data. This means not only retaining all the features corresponding to astrophysical phenomena, but also ensuring that artefacts and fake 'features' are not introduced into the image by the imaging process – this would compromise the scientific conclusions drawn from the observation.

The dependence on highly accurate imaging methods is only expected to get stronger as more and more data are collected with next-generation interferometers. For instance, it is widely expected that the SKA would, in fact, not deliver raw data to astronomers, instead performing the imaging step at an early stage and directly providing high-fidelity images as the final 'product'. This would be done to address the issue of handling and storing the massive amounts of data expected to be acquired by the array. As more advanced interferometers and faster electronics enter regular use, the field of radio astronomy is expected to witness not only significant scientific progress but also a plethora of computational challenges.

2.4.1 State of the art: CLEAN and friends

Radio-interferometric imaging got a big boost forward by the development of a simple but effective imaging method called CLEAN . It was designed and proposed by Högbom [1974] to recover point-like radio sources, and quickly became the definitive imaging method for all science efforts in the field. CLEAN has been developed further by various contributions over the years [Schwab, 1984, Bhatnagar and Cornwell, 2004, Cornwell, 2008], each extending its use cases to handle more complicated data and imaging constraints, and to recover images of sources with more varied and complicated structure.

At its core, CLEAN is a greedy algorithm. This algorithm assumes that the measured visibilities, once filtered through an inverse Fourier transform, are the result of the convolution of the underlying original image with the point spread function (PSF) of the interferometer, corrupted by additive noise – this result is termed the 'dirty image' (and hence the term 'CLEAN' to obtain the original image from this dirty image). Methods that aim to recover the underlying original image from the dirty image by reversing the effects of the convolution of the PSF of the receiving system are hence referred to as deconvolution methods – this is not limited to CLEAN and its modified versions.

In its most basic form, CLEAN runs directly in the image plane as follows: it computes the 'dirty image' and 'dirty beam' by applying the inverse Fourier transform on the measured visibilities and uv coverage respectively, Then it identifies the pixel with maximum intensity in the dirty image and subtracts a scaled version of the PSF of the interferometer centred at that pixel. It notes the position of the pixel and intensity value of the removed component by adding a scaled discrete delta function in a 'model' – this model eventually results in the cleaned image in the end. CLEAN thus successively removes peaks from the dirty image until no significant structure remains. The resulting leftover image is called the 'residual image'. The model at this stage contains many discrete delta functions at locations where CLEAN claims to have detected point-like sources. The delta functions in the model are convolved with the 'CLEAN ' beam – a version of the PSF without any sidelobes. This is usually a Gaussian-like function. Finally, CLEAN adds the residual image to the convolved model, resulting in the final output, the so-called 'restored image'.

Seen through the lens of our matrix formulation of the various components of the measurement model, CLEAN attempts a non-linear deconvolution by iteratively performing the following steps: (i) predicting 'model' visibilities ($\boldsymbol{\Phi x}$), computing 'residual' visibilities by subtracting model from measured visibilities ($\boldsymbol{y} - \boldsymbol{\Phi x}$) and gridding them onto a discrete Fourier grid ($\mathbf{G}^\dagger(\boldsymbol{y} - \boldsymbol{\Phi x})$), (ii) forming the 'dirty' residual image by applying the inverse Fourier transform on the gridded visibilities ($\overline{\mathbf{Z}}^\dagger \overline{\mathbf{F}}^\dagger \mathbf{G}^\dagger(\boldsymbol{y} - \boldsymbol{\Phi x}) = \boldsymbol{\Phi}^\dagger(\boldsymbol{y} - \boldsymbol{\Phi x})$), and eventually (iii) selecting the peak of the residual image as a model component.

A notable assumption in Högbom-CLEAN is that the original image is a sum of isolated point sources whose amplitudes are constant across the observing bandwidth. Compact emissions can then be easily accounted for and subtracted from visibilities forming model components, since the energy is assumed to be concentrated at the isolated point sources: this is an implicit assumption of sparsity in the image domain.

Running CLEAN over an arbitrary but sufficiently high number of iterations usually results in satisfactory image reconstruction. CLEAN depends on other arbitrary parameters as well, to reach the desired solution, e.g., the factor by which PSF components are scaled before being subtracted from the dirty image. The choice of parameters heavily affects the running time and the effectiveness of Högbom-CLEAN, and many improvements have been proposed to alleviate this. A significant change was proposed by Clark [1980], who introduced major and minor cycles in the algorithm. Several minor cycles identify the PSF components to be removed from the dirty image, by only using a small patch of the PSF, followed by a major cycle in which the image is Fourier transformed to the uv plane where the actual subtraction with the full PSF takes place. This speeds up the algorithm because the later step of convolution in image space is replaced by a multiplication in Fourier space. Another popular variant, Cotton-Schwab-CLEAN (CS-CLEAN), performs the major cycle subtractions directly on the degridded, continuous visibilities, improving the final image quality by reducing aliasing effects [Schwab, 1984].

Even though CLEAN was introduced more than forty years ago, it remains the standard image reconstruction algorithm employed by radio astronomers today. Several other versions of CLEAN have been developed, each with specific modifications to account for more complex radio sources and more antenna-related effects or to extend the algorithm for more wide-band imaging. A 'multi-scale' method (MS-CLEAN) improves sparsity in a multi-scale decomposition, thus leading to a better reconstruction of images that contain extended sources in addition to point sources; though this performs better than Högbom-CLEAN, it is influenced by the choice of the basis profiles and scales [Cornwell, 2008]. Yet another modified CLEAN version, using 'adaptive scale pixel' decomposition, called ASP-CLEAN was developed to eliminate this influence by adaptively choosing scales [Bhatnagar and Cornwell, 2004]. MS-CLEAN has also been extended to wide-band imaging with the development of the MS-MFS-CLEAN (multi-scale multi-frequency CLEAN) algorithm. Bhatnagar et al. [2013] provide a summary of different developments in CLEAN-based deconvolution algorithms.

Most radio interferometers operating today use a CLEAN variant in their data processing pipeline. For example, the NRAO-run VLA uses CS-CLEAN and MS-CLEAN to produce images through their in-house software framework called CASA (Common Astronomy Software). Similiarly, the Low Frequency Array (LOFAR), run by the Netherlands Institute for Radio Astronomy (ASTRON), develops and uses the 'AWimager' which implements CLEAN with modifications and preprocessing for LOFAR-specific data. The major issues with the continued use of CLEAN, however, are that (i) it requires manual oversight for the different tunable parameters as well as defining patches in the image domain to guide the algorithm, (ii) it is unlikely to scale well to the big data regime for next-generation telescopes, and (iii) there are no guarantees on the stability of CLEAN as an algorithm, meaning that it may or may not converge to a meaningful solution – although it most often does provide a reasonable solution after an arbitrary but sufficiently high number of iterations.

In a bid to provide more robust, automated and reproducible deconvolution, several novel techniques have been proposed lately that claim to have image reconstruction performance similar to or better than that achieved by the classical CLEAN-based algorithms [Wiaux et al., 2009a, Li et al., 2011, Carrillo et al., 2012, 2014, Garsden et al., 2015, Dabbech et al., 2015, Ferrari et al., 2015, Onose et al., 2016]. The main draw of these new imaging methods is that they require little manual intervention and examination, and many of them provide proof of convergence, thus guaranteeing that a stable solution would be reached at the end of the deconvolution process. Additionally, they employ convex optimization algorithms that can potentially be implemented in a parallelizable fashion, thus rendering these imaging methods scalable to big data levels. The common thread across these newly proposed imaging methods is the notion of a *sparse representation* of the underlying signal, and the iterative reconstruction algorithms to exploit this *sparsity*. A discussion of these imaging methods would, therefore, benefit from a segue at this point into an overview of the concepts of sparsity and reconstruction algorithms, in particular, in light of a recent theory on acquiring and reconstructing such sparse signals – *compressed sensing*.

3 Compressed sensing and convex optimization

3.1 Compressed sensing

Compressed sensing (CS) theory has been developed under the assumption that many naturally occurring signals are sparse (or more generally, compressible) in some basis [Candès et al., 2006b, Donoho, 2006a]. Traditional sampling and data acquisition techniques use the Nyquist-Shannon theorem that requires the sampling rate to be at least twice the maximum frequency present in the signal. With the assumption of sparsity of the signal, CS theory claims that a very small number of measurements are enough for accurate reconstruction, condition to restrictions on how the measurements are taken. 'Accurate' reconstruction is taken to be equivalent to having an arbitrarily small probability for not fully recovering the signal [Candès and Wakin, 2008]. CS theory offers both acquisition *and* reconstruction techniques for such signals.

Data acquired by large radio interferometric arrays are typically many times larger than the size of any image one would reconstruct from the data, so radio interferometric imaging does not directly fall under the regime of CS theory. However, dimensionality reduction methods applied on acquired data can lead to low-sized data. Additionally, the signals underlying the data captured by radio interferometers can be naturally sparse in different bases, e.g. in the spatial domain, due to having discrete point sources in the sky, or in wavelet bases, among others. The measurement of these signals in a noisy environment can then be put in the familiar framework of CS by defining the inverse problem as given by equation (2.8), thus paving the way for CS-based reconstruction methods. Signal recovery in the CS framework can be achieved by solving global minimization problems, making use of the sparsity assumption as a regularizer in convex optimization algorithms.

Parts of this chapter have been published by Carrillo et al. [2015].

3.1.1 Theoretical background

We start by defining the concept of *sparsity.* A vector $\boldsymbol{\alpha} \in \mathbb{C}^d$ is k-sparse if it has at most k non-zero elements. $\boldsymbol{\alpha}$ is *compressible* if the remaining $d - k$ entries of $\boldsymbol{\alpha}$ are non-zero but negligible entries. A signal $\boldsymbol{x} \in \mathbb{C}^N$ has a k-sparse representation if there exists a basis (or dictionary) $\boldsymbol{\Psi} \in \mathbb{C}^{N\times d}$ so that $\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is k-sparse.

In the following sections, we shall also be using the concept of the 'ℓ_p-norm', denoted by $\|\cdot\|_p$. The ℓ_p-norm, for $1 \le p < \infty$, of a vector $\boldsymbol{v} \in \mathbb{C}^N$ is defined as

$$\|\boldsymbol{v}\|_p = \left(\sum_{\mathrm{i}=1}^{N} |v_\mathrm{i}|^p\right)^{1/p} \qquad \text{for } \; p \in [1,\infty), \tag{3.1}$$

where v_i are the elements of $\boldsymbol{v}$. The ℓ_p-norms we shall be using in this thesis are –

the ℓ_1-norm, also known as the taxicab norm – the sum of the magnitudes of v_i, and

the ℓ_2-norm, also known as the Euclidean norm – the 'usual' distance of $\boldsymbol{v}$ from the origin.

The sparsity of a vector can be expressed through a pseudo-norm, akin to the ℓ_p-norms. The ℓ_0-'norm' of the vector $\boldsymbol{v}$, given by $\|\boldsymbol{v}\|_0$, is simply the total number of its non-zero elements, $v_\mathrm{i} \neq 0$. The quotation marks around the word 'norm' indicate that $\|\cdot\|_0$ is, in fact, not a true ℓ_p-norm (e.g., $\|a\boldsymbol{v}\|_0 \neq |a|\|\boldsymbol{v}\|_0$ for a scalar a). We simply abuse the notation here for simplicity because $\lim\limits_{p\to 0} \|\boldsymbol{v}\|_p^p = \|\boldsymbol{v}\|_0$. So, a k-sparse vector $\boldsymbol{v}$ can be equivalently expressed as $\|\boldsymbol{v}\|_0 \le k$.

Now, given a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{M\times N}$ that senses/samples the signal $\boldsymbol{x} \in \mathbb{C}^N$ to give an M-dimensional measurement vector $\boldsymbol{y} \in \mathbb{C}^M$ (possibly affected by some additive noise $\boldsymbol{n} \in \mathbb{C}^M$), we probe the relevance of CS theory in the measurement model.

The theory of compressed sensing comes into play with this inverse problem in two ways. Firstly, it proposes the surprising possibility that, for this particular class of k-sparse signals, the Nyquist-Shannon sampling rate need not be reached in order to ensure accurate reconstruction of the signal. Secondly, it proposes techniques to perform this accurate signal reconstruction.

The first claim of needing far fewer measurements than that dictated by the Nyquist-Shannon sampling theorem implies that a value of $M \ll N$ is a sufficient number of non-adaptive, linear measurements [Candès et al., 2006b, Donoho, 2006a, Candès and Romberg, 2007, Tropp et al., 2010]. The fact that fewer measurements than unknowns are available defines the *ill-posed* nature of the inverse problem. It is clear that infinitely many solutions exist for the underdetermined linear system underlying the ill-posed inverse problem. CS theory urges the use of the sparsity of the signal representation to drastically limit the number of possible solutions. Although in principle this could be strictly applied by forcing solutions (given by $\hat{\boldsymbol{\alpha}}$) that satisfy $\|\boldsymbol{\alpha}\|_0 \le k$, it is more feasible to force the condition $\min\limits_{\boldsymbol{\alpha}\in\mathbb{C}^N} \|\boldsymbol{\alpha}\|_0$, since the exact value of k is rarely known beforehand. Applying these restrictions, we could then solve the inverse problem by finding a vector $\hat{\boldsymbol{\alpha}}$ that satisfies the following ℓ_0-*minimization problem*:

$$\min_{\boldsymbol{\alpha}\in\mathbb{C}^N} \|\boldsymbol{\alpha}\|_0 \qquad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{\Phi\Psi\alpha}. \tag{3.2}$$

This minimization problem, however, is NP-complete [Mallat and Zhang, 1993, Natarajan, 1995], and therefore computationally very demanding. One way to get around this is to use greedy algorithms to solve the problem. This family of algorithms includes matching pursuit (MP) and its derivatives, like orthogonal matching pursuit (OMP, Tropp and Gilbert [2007]), compressive sampling matching pursuit (CoSaMP, Needell and Tropp [2009]), preprojected OMP [Flinth and Kutyniok, 2017], gradient pursuit [Blumensath and Davies, 2008a] etc. Alternative methods that have been proposed to find a solution for equation (3.2) include iterative thresholding methods [Kingsbury, 2001, Herrity et al., 2006, Blumensath and Davies, 2008b]). More general descriptions of iterative thresholding have been proposed by Figueiredo and Nowak [2003], Figueiredo et al. [2007b] albeit not for the minimization problem as posed in equation (3.2). Each of these methods has its own merits and demerits, either in terms of its computational complexity, or its stability and convergence, or its robustness to perturbations in measurements. We focus on a different approach to equation (3.2), as described in detail in the next section.

3.1.2 ℓ_1-minimization

Another class of methods to handle the NP-completeness of ℓ_0-minimization is to use a convex relaxation of the problem, replacing the ℓ_0-norm by the ℓ_1-norm. This results in an ℓ_1-minimization problem (basis pursuit, BP) of the form $\min_{\boldsymbol{\alpha}\in\mathbb{C}^N} \|\boldsymbol{\alpha}\|_1$ subject to $\boldsymbol{y} = \boldsymbol{\Phi\Psi\alpha}$. In the presence of measurement noise, this becomes the basis pursuit denoising problem (BPDN), given by

$$\min_{\boldsymbol{\alpha}\in\mathbb{C}^N} \|\boldsymbol{\alpha}\|_1 \qquad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{\Phi\Psi\alpha}\|_2 \leq \varepsilon, \tag{3.3}$$

where $\varepsilon > 0$ is an upper bound on the ℓ_2-norm of the residual noise $\boldsymbol{y} - \boldsymbol{\Phi\Psi\alpha}$. The ℓ_1-minimization form is a *convex optimization* problem, which is computationally feasible and for which many standard algorithms exist [Daubechies et al., 2004, Combettes and Pesquet, 2007, Figueiredo et al., 2007b, Candès et al., 2008, Beck and Teboulle, 2009]. We limit ourselves to ℓ_1-minimization methods in further discussions in this thesis. The only caveat here is that the solution to the ℓ_1-minimization problem may not necessarily be the solution one would have obtained for the ℓ_0-minimization problem. We can, however, prove that ℓ_1-minimization favours sparse solutions in general. Fig. 3.1 presents a visual intuition on why this is the case. The unique solution to the ℓ_1-minimization problem can be seen to coincide with the sparsest solution in that space. Fornasier and Rauhut [2011] provide a more thorough mathematical treatment of this conjecture, with proofs. In this section we briefly analyze the utility of ℓ_1-minimization as a viable method to recover the sparsest solution, given the CS context for signal recovery.

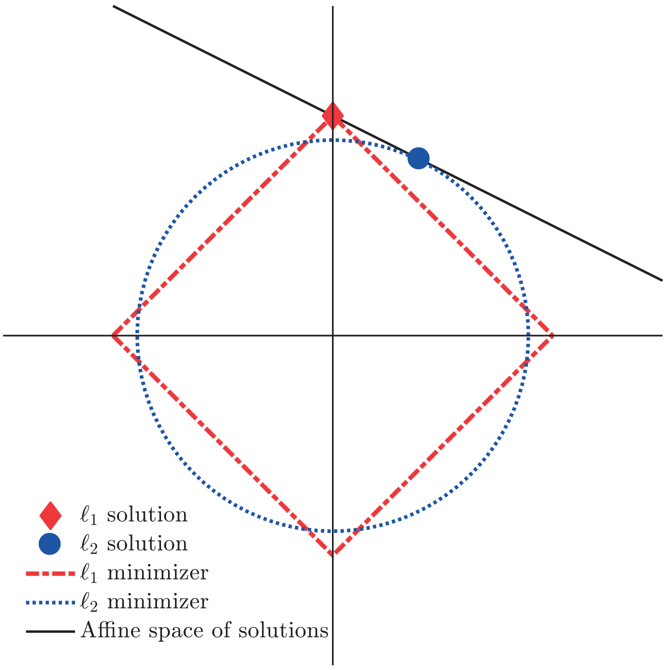


Figure 3.1 – ℓ_1-minimization favours the sparsest solution within the affine space of the solutions of $\boldsymbol{y} = \boldsymbol{\Phi x}$, as shown here for $\mathbb{R}^2$. Note that the ℓ_1- and ℓ_2-balls shown here are the smallest such balls that meet the affine space of the solutions, and the ℓ_2-minimizer can be seen to reach a non-sparse solution.

3.1.3 Mutual incoherence

Along with the requirement that the signal $\boldsymbol{x}$ have a sparse representation to be able to be recovered accurately through ℓ_1-minimization, CS theory relies on another relationship between the measurement matrix $\boldsymbol{\Phi}$ and the sparsity basis $\boldsymbol{\Psi}$, named *mutual incoherence*. Coherence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ is defined as follows:

$$\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) = \sqrt{N} \max_{1 \leq i,j \leq N} |\langle \boldsymbol{\Phi}_i, \boldsymbol{\Psi}_j \rangle|, \tag{3.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. The mutual coherence essentially measures the correlation between the measurement vectors and the sparsity basis vectors in $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ (i.e., $\boldsymbol{\Phi}_i$ are rows of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}_j$ are columns of $\boldsymbol{\Psi}$). It ranges from a value $\mu = 1$ for incoherent $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, to $\mu = \sqrt{N}$ for maximally correlated $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. Incoherence brings to light the idea that signals that have a sparse representation in $\boldsymbol{\Psi}$ must be 'spread out' in the measurement domain. This means that, contrary to the signal $\boldsymbol{x}$ itself, the sensing waveforms would ideally have a dense representation in $\boldsymbol{\Psi}$. An example can be the case where the signal is completely spread out in the time domain, but represented as a spike in the frequency domain. High mutual incoherence is necessary for CS theory because it implies that it is possible to exactly recover the underlying signal with overwhelming probability [Candès et al., 2006a, Candès and Romberg, 2007], provided a minimum number of random measurements are taken. This minimum number depends on the characteristics of the measurement matrix $\mathbf{\Phi}$. For instance, for uniformly random measurements, it suffices to take $M \geq C\mu^2(\mathbf{\Phi}, \mathbf{\Psi})K \log(N)$ measurements (for some positive constant C) to be able to recover the signal $\boldsymbol{x} \in \mathbb{C}^N$ with a K-sparse representation. The noteworthy feature here is that the measurements needed to accurately recover the signal are random and non-adaptive, i.e., we need no prior information on the structure of the sparsity. Although it seems surprising at first glance, this follows from the mutual incoherence between $\mathbf{\Phi}$ and $\mathbf{\Psi}$, which ensures that relevant information about the signal is spread out, which means that any M measurements would suffice. This set of randomly chosen M measurements will lead to no information loss, and in fact, the higher the mutual incoherence $\mu(\mathbf{\Phi}, \mathbf{\Psi})$, the fewer the samples needed to guarantee accurate signal recovery. This seeming indifference to the number of measurements can be hardened to also include robustness to small perturbations in the measurements, but this adds additional requirements on the measurement matrix $\mathbf{\Phi}$, which shall be briefly touched on in the following section.

3.1.4 The NSP and RIP

As mentioned at the beginning of section 3.1, the guarantees of accurate reconstruction provided by CS theory are condition to properties of (i) the signal to be reconstructed, and (ii) the sensing matrix used to probe the signal. As part of the second set of requirements, we discuss two properties of the sensing/measurement operator $\mathbf{\Phi}$, viz., the Null Space Property (NSP) and the Restricted Isometry Property (RIP). These properties, if satisfied by $\mathbf{\Phi}$, guarantee stable recovery of the signal $\boldsymbol{x}$ from the measurements $\boldsymbol{y}$ [Candès et al., 2006a].

The NSP is essential to guarantee exact signal recovery through ℓ_1-minimization; $\mathbf{\Phi}$ satisfies the NSP of order k with constant $\gamma \in (0, 1)$ if

$$\|\boldsymbol{\eta}_T\|_1 \leq \gamma \|\boldsymbol{\eta}_{T^c}\|_1 \tag{3.5}$$

for all sets $T \subset \{1, \ldots, N\}$, $|T| \leq k$ and for all $\boldsymbol{\eta} \in \text{Null}(\mathbf{\Phi})$. Here $\boldsymbol{\eta}_T$ is obtained by setting entries of $\boldsymbol{\eta}$ to zero for indices which are not in T. T^c is the complement of T. Put differently, the NSP means that no k-sparse signals are contained in the null space of the operator $\mathbf{\Phi}$. It can then be shown that for a k-sparse signal $\boldsymbol{x}$ the reconstruction achieved using ℓ_1-minimization is exact, and for a more general $\boldsymbol{x}$, the reconstruction error is bounded (Theorem 1 and its proof by Fornasier and Rauhut [2011, pp. 199]). Conversely, if ℓ_1-minimization can recover all k-sparse solutions, then $\mathbf{\Phi}$ can be shown to necessarily satisfy the NSP of order k [Cohen et al., 2009]. This shows the equivalence of the NSP and recovering sparse signals through ℓ_1-minimization.

The RIP, on the other hand, provides guarantees in the more general case – it characterizes stable signal recovery in the presence of noise, and ensures that two different k-sparse signals remain well separated even after the application of the measurement operator. $\mathbf{\Phi}$ satisfies the

RIP of order k with constant $\delta \in (0,1)$ if

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{\Phi x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2 \tag{3.6}$$

for all k-sparse signals $\boldsymbol{x}$. The RIP is a stronger condition than the NSP, and in fact implies the NSP (Lemma 2 and its proof by Fornasier and Rauhut [2011, pp. 200]).

In our case with the requirement for the measurement operator to satisfy the RIP of order $2k$, it is worth noting that an operator $\boldsymbol{\Phi}$ satisfying the RIP of order $2k$ is in fact the Johnson-Lindenstrauss embedding[1] for the case where $\boldsymbol{x}$ would be the difference between two k-sparse signals [Krahmer and Ward, 2011]. The fact that the RIP also implies robustness to measurement noise has been proven [Candès et al., 2006a], and bounds on the reconstruction error have also been deduced [Foucart and Rauhut, 2013]. These results have also been extended beyond exactly sparse signals to include noisy measurements of compressible signals [Candès et al., 2006b].

While it is difficult to verify using deterministic methods whether a given matrix satisfies the RIP [Rauhut, 2011, Bandeira et al., 2013], the major results in CS theory were proven using random matrices [Candès et al., 2006a,b]. Gaussian and Bernoulli random matrices satisfy the RIP with very high probability, provided a minimum number of measurements, $M \geq CK\log(N/K)$, are taken [Fornasier and Rauhut, 2011]. This shows that the minimum required measurements are linear in the sparsity K (up to a log factor), which has been proven to be optimal [Donoho, 2006b]. However, these completely random matrices are, as shown later in section 4.4.2, computationally inefficient. Some structured random matrices, like partial Fourier matrices, satisfy the RIP while remaining computationally feasible; preliminary studies on using random matrices are also presented in section 4.4.2.

3.1.5 Analysis-based problem

Now, given that our linear measurement operator $\boldsymbol{\Phi}$ satisfies the NSP – or the RIP – we can recover the signal $\boldsymbol{x}$ with the following non-linear reconstruction method: Limiting ourselves to only sparse representations of the signal $\boldsymbol{x}$ in $\boldsymbol{\Psi}$, while maintaining reasonable fidelity with the measurement vector $\boldsymbol{y}$, the solution to the inverse problem in equation (2.8) can be found by solving the so-called 'synthesis-based' problem given by

$$\min_{\boldsymbol{\alpha}\in\mathbb{C}^N} \|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y}-\boldsymbol{\Phi\Psi\alpha}\|_2 \leq \varepsilon. \tag{3.7}$$

The problem formulation here is identical to the BPDN formulation of equation (3.3). The signal $\hat{\boldsymbol{x}}$ can then be 'synthesized' from the solution $\hat{\boldsymbol{\alpha}}$ of equation (3.7) through applying $\hat{\boldsymbol{x}} = \boldsymbol{\Psi}\hat{\boldsymbol{\alpha}}$.

Alternatively, the signal $\hat{x}$ can be directly recovered by solving the 'analysis-based' prob-

---

[1] The Johnson-Lindenstrauss embedding is described in more detail in section 4.3.2.

lem [Elad et al., 2007] given by

$$\min_{\boldsymbol{x}\in\mathbb{C}^N} \|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2 \le \varepsilon. \tag{3.8}$$

We shall be using the analysis-based problem definition from here on, especially when referring to radio-interferometric imaging algorithms in the coming sections. Using equation (3.8) as the starting point has other advantages as well, e.g., when redundant dictionaries are used for the bases, the dimension of the problem does not increase as we solve for $\boldsymbol{x}$ and not $\boldsymbol{\alpha}$. The synthesis-based formulation has been used extensively with redundant dictionaries in the recent past, particularly in the image processing literature [Elad, 2010, Elad et al., 2010, Bruckstein et al., 2009]. In addition to the aforementioned advantage regarding dictionary sizes, analysis-based problems can more generally also exploit the idea of 'cosparsity' as a generative model [Nam et al., 2013, Giryes et al., 2014] and modifications in CS theory proposing an extended version of RIP called 'Dictionary-RIP' to enable the use of redundant dictionaries [Candès et al., 2011]. Convex optimization algorithms have been used successfully in solving problems such as equation (3.8) [Elad et al., 2010, Carrillo et al., 2014, Onose et al., 2016].

In equations (3.7) and (3.8), $\varepsilon > 0$ is the allowed tolerance on the disparity between the actual measurements and those generated by a guess at the underlying signal $\boldsymbol{x}$. In the general case this tolerance is actually a negative log-likelihood term given by $\boldsymbol{n}^\dagger \mathsf{C}_{\boldsymbol{n}}{}^{-1}\boldsymbol{n}$, where $\mathsf{C}_{\boldsymbol{n}}$ is the covariance matrix of the noise $\boldsymbol{n}$. Under the assumption that $\boldsymbol{n}$ has independent, identically distributed (i.i.d.) Gaussian random elements, this log-likelihood term simplifies to $\|\boldsymbol{n}\|_2^2/\sigma_n^2$. For reasonably high degrees of freedom – corresponding, in this setting, to high dimensional noise vectors – a χ^2 distribution exhibits concentration of measure, thus providing a sharp upper bound that is directly computable from the mean. In our simulations for image reconstruction we compute this upper bound as two standard deviations beyond the mean, which includes a large percentile of the distribution. If the assumption of i.i.d. Gaussian entries of the noise fails to hold true, then determining ε is not a direct analytical computation since the noise energy no longer follows a χ^2 distribution. Note that transforming visibilities to have i.i.d. Gaussian noise statistics is a typical effect of the so-called 'natural' weighting procedure that is commonly performed in state-of-the-art radio-interferometric imaging CLEAN-based approaches.

3.2 Convex optimization

Convex optimization (or convex minimization) deals with convex functions and constraints, and forms a class of computationally efficient, mathematically sound algorithms that are well-suited to solve convex problems. One major advantage of convex optimization problems is that any local minimum found by an algorithm must necessarily be *a* global minimum for the problem space. This follows from the definition of convexity of the functions and of the sets over which the functions are minimized. Algorithms to solve simple convex optimization problems have been designed since as early as the 1950s, and are, by now, a very familiar component in any numerical solver or toolkit. Examples include the class of subgradient methods [Shor and Zhurbenko, 1971, Lemaréchal et al., 1995] and interior-point methods [Karmarkar, 1984, Nesterov and Nemirovskii, 1994]. Here we briefly discuss another class of methods that are suitable for convex optimization in more complex cases, containing several concurrent conditions. These methods work by splitting the whole problem into simpler individual convex sub-problems, and are aptly named *proximal splitting* methods.

3.2.1 Proximal splitting methods

Proximal splitting methods solve optimization problems of the (unconstrained) form

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \ldots + f_n(\boldsymbol{x}), \tag{3.9}$$

where each $f_i(\boldsymbol{x}) : \mathbb{R}^N \to (-\infty, \infty)$ is a proper lower semi-continuous convex function. The constrained formulation of the problem can be re-expressed in an unconstrained form by introducing *indicator functions*. The indicator function χ_C of a convex constraint set C is a proper lower semi-continuous function defined as

$$\chi_C(\boldsymbol{x}) = \begin{cases} 0 & : \boldsymbol{x} \in C, \\ +\infty & : \boldsymbol{x} \notin C. \end{cases} \tag{3.10}$$

The re-expressed unconstrained formulation then simply uses $f_i(\boldsymbol{x}) = \chi_{C_i}(\boldsymbol{x})$ for some convex constraint set C_i.

Proximal splitting aims to process the whole minimization problem by handling different functions $f_i(\boldsymbol{x})$ individually, which can then be minimized separately. Each $f_i(\boldsymbol{x})$ is dealt with in the algorithm through its 'proximity' operator $\mathrm{prox}_{f_i}(\boldsymbol{x})$, defined as

$$\mathrm{prox}_{f_i}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}\in\mathbb{R}^N} f_i(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2. \tag{3.11}$$

This simplifies to a projection operator on C_i in the case $f_i(\boldsymbol{x}) = \chi_{C_i}(\boldsymbol{x})$. Indeed, the proximity operator can be seen as a general extension of the usual projection operator. Thus, the inverse problem noted in the forward model presented in equation (2.8) and reformulated as the analysis-based problem in equation (3.8) can then be solved with proximal splitting methods, using the indicator function of the convex constraint set as one of the functions $f_i(\boldsymbol{x})$.

Proximal splitting methods have been very successful in convex optimization due to their readily parallelizable structure and their handling of complex problems by splitting them into smaller, easier functions. Many of the most commonly used algorithms in convex optimization today – forward-backward, Douglas-Rachford, iterative thresholding, etc. – can be shown to be special cases derived from proximal algorithms. A detailed treatment of proximal splitting methods can be found in the comprehensive overview by Combettes and Pesquet [2011].

3.3 Relevant radio-interferometric imaging methods

Thanks to the prevalence of computationally feasible algorithms to solve the inverse problem in equation (3.8) through convex optimization methods, we are now at a point where recovering an underlying image from given measurements corresponding to an incomplete coverage of the Fourier plane is not only possible but reasonably robust. As mentioned in section 2.4.1, many sparse recovery methods have been proposed in recent years for radio-interferometric imaging. In this thesis, we shall limit ourselves to two methods in particular, namely the alternating direction method of multipliers (ADMM, Boyd [2010], Yang and Zhang [2011]), and the primal-dual forward-backward (PDFB, Onose et al. [2016]) method. ADMM and PDFB are coupled with the 'sparsity averaging re-weighted analysis' algorithm (SARA), first proposed by Carrillo et al. [2012] and further developed in subsequent work [Carrillo et al., 2013, 2014]. SARA proposes to use, as a regularizer in the optimization problem, the assumption of average signal sparsity over representations in *multiple* wavelet bases instead of just one basis (as is the case in the other sparse recovery algorithms in radio interferometry). It has been shown to outperform other existing reconstruction algorithms – a comparison for the case of discrete visibilities was performed by Carrillo et al. [2012]. SARA uses average sparsity as a structured sparsity prior for compressed sensing to deal with recovery of the signal, and stems partially from work that was originally published by Wiaux et al. [2009a,b] and McEwen and Wiaux [2011]; relevant work was also published by Wenger et al. [2010] and Hardy [2013].

3.3.1 The SARA algorithm

SARA is motivated by the fact that radio astronomical images often contain complex structures, ranging from multiple point sources or compact smooth structures, all the way to large extended sources covering significant portions of the image. It can be seen that these different image features are sparse in different domains, for instance in the image domain, in the gradient magnitude, and in certain wavelet bases. The concurrence of these features in a given image leads Carrillo et al. [2012] to claim that promoting a simultaneous, average sparsity over different bases would lead to a better-modelled optimization problem, and consequently, a more accurate recovery of the underlying image. To this effect, different orthonormal bases $\boldsymbol{\Psi}_i$ are concatenated to form a dictionary $\boldsymbol{\Psi} = \frac{1}{\sqrt{q}}[\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, ..., \boldsymbol{\Psi}_q] \in \mathbb{R}^{N\times D}$, where $D = qN$. The $\boldsymbol{\Psi}_i$s chosen for radio astronomical images contains nine different orthonormal bases: the Dirac basis and the first eight Daubechies wavelet bases. We note here that images can obviously not be sparse in multiple bases simultaneously, unless the bases are completely coherent (this has already been illustrated in section 3.1.3). The way to promote sparsity of different features in different bases in the same image is through a compromise – by computing the *average* sparsity over these bases and favouring representations that reach a high average value, which often implies a reasonable degree of sparsity in each basis $\boldsymbol{\Psi}_i$. The assumption is that for a given pair of bases, the average sparsity of a signal is expected to be above some lower bound, and that the SARA prior will promote solutions that satisfy/exceed such lower bounds for all such pairs of bases.

Re-weighted analysis in SARA refers back to the analysis-based problem formulation given in equation (3.8). The ℓ_1-norm substituting for a sparsity prior, $\|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1$, is modified to form a *weighted* ℓ_1 term $\|\mathbf{W}\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1$. The weights given by the diagonal matrix $\mathbf{W} \in \mathbf{R}^{D\times D}$ are the reciprocal values of the elements of $\hat{\boldsymbol{x}}$ (with minor adjustments to ensure stability for divide-by-zero errors). In iterative recovery algorithms, the weights are updated after a pre-defined, fixed number of iterations of the recovery algorithm, using the current guess of the solution. We can see that applying weights in this manner enables the re-weighted ℓ_1-norm to approximate the ℓ_0-'norm' more closely, making this a more appropriate sparsity prior. Indeed this was the motivation behind re-weighted ℓ_1-minimization as a potential recovery algorithm as proposed by Candès et al. [2008]. The SARA algorithm takes a similar approach, and solves a sequence of weighted ℓ_1-minimization problems to arrive at the final solution. The weights used in a given minimization problem are computed from the solution of the preceding problem. Both SARA and the reweighted ℓ_1-minimization by Candès et al. [2008] are iteratively reweighted algorithms that can be seen in the light of the general framework of Majorization-Minimization (MM) methods [Hunter and Lange, 2004]. An unconstrained formulation of the analysis-based problem has also been treated by Figueiredo et al. [2007a] using the MM framework, and the reweighting method from SARA can be shown to be equivalent to applying their MM algorithm using a nonconvex ℓ_p-'norm' regularizer with $0 < p < 1$. SARA also employs an additional constraint on the underlying image, that it contain only non-negative real values. This follows from the observation that the image represents an intensity distribution.

Image reconstruction from simulated discrete visibilities using SARA was reported to provide better results than several other sparse recovery algorithms [Carrillo et al., 2012]. This was further developed by Carrillo et al. [2014] to use simulated continuous visibilities; SARA in combination with proximal splitting methods, notably the simultaneous direction method of multipliers (SDMM, Setzer et al. [2010])[2] was then proposed as a general algorithm for radio-interferometric imaging, and was released as part of a software package called PURIFY .

3.3.2 ADMM and PDFB

We focus here on solving the inverse problem presented in equation (2.8) through the analysis-based formulation as in equation (3.8), using the sparsity prior as described in section 3.3.1 but without re-weighting. We also employ the additional constraint that the image contain only non-negative real values, as with SARA. Employing the proximal splitting method described in section 3.2.1 is very apt here, as it renders the optimization problem tractable, with easily solvable subproblems. The full minimization can thus be written as a sum of functions in the form of equation (3.9), giving

$$\min_{\boldsymbol{x}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{\Psi}^\dagger \boldsymbol{x}) + f_3(\boldsymbol{\Phi}\boldsymbol{x}), \tag{3.12}$$

---

[2] Afonso et al. [2010] independently proposed an equivalent algorithm, named 'split augmented Lagrangian shrinkage algorithm' (SALSA), developed further to solve constrained problem formulations (C-SALSA, Afonso et al. [2011]), both based on ADMM.

where the constraints appear here through their indicator functions. So, the applied constraints are in the following form:

$$\begin{aligned} f_1(\boldsymbol{v}) &= \chi_{C_1}(\boldsymbol{v}), \quad C_1 = \mathbb{R}_+^N, && \text{(real and non-negative)} \\ f_2(\boldsymbol{v}) &= \|\boldsymbol{v}\|_1, && \text{(sparse in } \boldsymbol{\Psi}\text{)} \\ f_3(\boldsymbol{v}) &= \chi_{C_3}(\boldsymbol{v}), \quad C_3 = \boldsymbol{v} \in \mathbb{C}^M : \|\boldsymbol{y} - \boldsymbol{v}\|_2 \leq \varepsilon. && \text{(data fidelity)} \end{aligned} \tag{3.13}$$

where $\boldsymbol{v}$ can be a vector in image, sparsity, or data space depending on the f_i in question.

In addition to SDMM that we use for preliminary tests and simulations on small-scale data, the two methods that we employ to solve this minimization problem are based on algorithms proposed by Carrillo et al. [2015] and Onose et al. [2016]. The first method is a proximal linear version of ADMM. It is applied to solve equation (3.12) by using a function $g = f_1 + f_2$ to get around the ADMM limitation of using two functions only. One advantage of using ADMM over the previously mentioned SDMM is that it avoids solving a linear system at each iteration of the algorithm, potentially saving expensive computations. ADMM can be implemented with a parallel structure, thus enabling scalability for big data. A detailed description of the proximal linear ADMM algorithm, along with image reconstruction results on simulated and real telescope data were presented by Carrillo et al. [2015], Onose et al. [2016].

The second method, proposed by Onose et al. [2016], is the 'primal-dual forward-backward' (PDFB) method. Primal-dual methods concurrently solve the original optimization problem at hand – the primal form – and an alternative, dual formulation of the same problem [Komodakis and Pesquet, 2015]. This brings many computational advantages, since full proximal splitting can be achieved (without resorting to bunching up functions together to achieve a partial splitting of the original problem as is the case with ADMM). PDFB continues with the fully split optimization problem with a forward-backward structure. The advantage of PDFB is the readily parallelizable algorithm structure, and the potential to extend it by using different subiterative schemes, e.g. stochastic gradients in the forward step, and the backward step using a proximity operator which approximates a projection. Implementation details and convergence proofs of PDFB are presented by Onose et al. [2016], along with its application to radio-interferometric imaging.

Several other methods have been proposed for radio-interferometric imaging in recent years, covering a wide variety of algorithmic and theoretical contributions that we shall not be discussing in detail here, like the undecimated multiscale method by Starck et al. [2007], Li et al. [2011], the distributed data transfer-based method by Ferrari et al. [2014], the Bayesian interference-based multi-frequency synthesis method of RESOLVE [Junklewitz et al., 2015], and the analysis-by-synthesis greedy algorithm of MORESANE [Dabbech et al., 2015].

3.3.3 CLEAN revisited

The inner working of CLEAN can now be seen through the perspective of convex optimization. In essence, CLEAN follows a 'matching pursuit' approach [Mallat and Zhang, 1993, Lannes et al., 1997]. The deconvolution process is non-linear, iteratively reconstructing the image one pixel at a time. Sparsity in the image domain is automatically promoted by CLEAN due to this pixelwise reconstruction and the underlying assumption of the sky being composed of a finite number of point sources. Alternatively, a gradient descent method can also explain the inner workings of CLEAN. The norm of residual visibilities ($\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2$) is minimized through a gradient descent regularized by an implicit sparsity prior on the recovered image.[3]

Flipping the interpretation around in the context of the proximal splitting-based convex optimization algorithms that we use in this work, ADMM and PDFB can be grossly seen as 'meta-CLEAN' image recovery methods, which run CLEAN-like iterations – in parallel – in multiple data, image, and sparsity spaces. These similar workings in multiple spaces can be seen in particular for PDFB, where we can identify specific portions of the algorithm which have clear correspondence with CLEAN subroutines described in section 2.4.1. A CLEAN major cycle, for instance, performs the same operations as the forward gradient step in PDFB which nudges the solution in the opposite direction of the gradient of the residual norm in data space. Similarly, the CLEAN minor cycles use a 'loop gain' to scale PSFs and choose significant components, whereas PDFB employs soft-thresholding in its proximal step to remove insignificant components across the image. Finally, the implicit sparsity constraint in CLEAN that is realized by placing delta functions in the model is introduced as an explicit prior in the minimization problem for PDFB. Onose et al. [2016] provide a thorough treatment of this idea, along with a more nuanced analogy of PDFB with CLEAN.

---

[3] This can also be seen as minimizing the residual norm using the Landweber algorithm [Landweber, 1951].

4 Big data and dimensionality reduction

The term ‘big data’ has entrenched itself into the everyday vocabulary of scientists, engineers, statisticians, policy planners, and most data professionals around the world today. It refers to very large, complex data sets that defy traditional data processing techniques, thereby necessitating new and innovative methods to perform analysis and information retrieval on the data. This is succinctly defined in the so-called ‘3Vs’ model, which defines big data as high-volume, high-velocity and high-variety information that requires specific technology and methods for analysis [De Mauro et al., 2016]. Big data has triggered a flurry of development in a wide variety of fields, from HPC services, storage and data handling to analytics, machine learning, and general purpose computing paradigms. These developments have touched a spectrum of research areas, including traditional computer science, but also physics, genomics, economics, healthcare, environmental science, and a host of other unrelated areas.

4.1 Motivation

Many massive scientific endeavours are underway in today’s big data era. In the field of particle physics, the largest particle accelerator in the world, the LHC (Large Hadron Collider), operated by CERN (the European organization for nuclear research), handles a staggering 30 petabytes of data annually, *after* discarding more than 99.9% of the collision events occurring in the LHC [Brumfiel, 2011]. Other fields have also seen sharp spikes in the scale of data being analyzed, like the complete documentation of the human genome with the genome project [Hood and Galas, 2003], and more recently, the human brain project; Radio interferometry is also experiencing a change of scale, especially in the context of next-generation telescopes like the VLA and the upcoming SKA. As mentioned in chapter 1, the SKA is expected to collect data at an extremely high rate of 5 terabits per second [Broekema et al., 2015]. In radio interferometry, the main motivation for collecting such large amounts of data is to achieve higher sensitivity, i.e., to be able to detect faint radio sources which may otherwise be overwhelmed in noisy measurements. This is why radio interferometers typically acquire data

---

Parts of this chapter have been published by Kartik et al. [2015] and Carrillo et al. [2015].

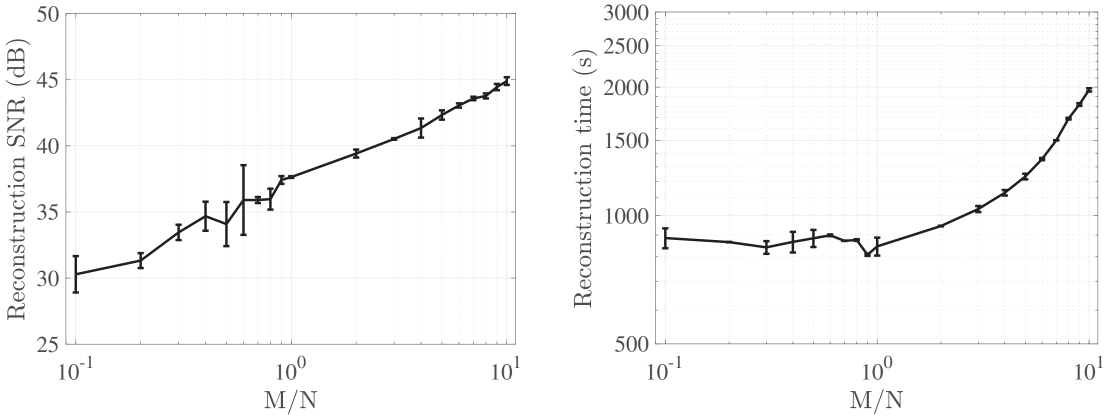


Figure 4.1 – The SNR and reconstruction time evolution of a modified version of ADMM for increasing data sizes using simulated data with 40 dB SNR. The x and y axes are in $\log_{10}$ scale. This modified ADMM algorithm [Carrillo et al., 2015] can be seen to scale well with data size.

of size hundreds of times larger than the desired image size.

We see a clear trend of large amounts of data playing a crucial role in achieving scientific goals, but this is also tightly coupled with reduction methods to intelligently select small portions of these data in order to realistically be able to store, study and process them. In radio interferometry, in addition to the limitations on the handling of large data, the motivation to devise dimensionality reduction methods comes from the observation that increasing data sizes leads to a disproportionately large increase in running times for many image reconstruction methods. In Fig. 4.1 we see the effects of increasing data size on both the image reconstruction quality and the running time of the ADMM algorithm. Starting from simulated data (with low noise levels), we reconstructed images using a non-parallelized version of ADMM. The results shown here were obtained from reconstructing images using data of varying sizes, ranging from 10% of the image size to 10 times the image size. On increasing the data size, we see a general increase in the quality of the corresponding reconstructed images. This is expected, as more data allows the reconstruction algorithm to capture better detail of the features in the underlying image. An interesting observation, however, is the effect of increasing data size on the algorithm running time. The time taken for image reconstruction increases with data size, which implies that indefinitely making more data available – while attractive as an option to gather more information – may not be feasible since it entails a huge cost in recovering this information. This cost may either be mitigated through more scalable image reconstruction algorithms [Carrillo et al., 2015, Onose et al., 2016, 2017] or through dimensionality reduction methods.

4.2 Handling big data

With the advent of the big data era, much effort has been put into research techniques to deliver us from the curse of dimensionality by reducing data size without losing information.

In general, these data handling techniques can be divided in two groups:

- Feature selection: re-representing the original data by a smaller set of 'features' or global characteristic vectors.
- Dimensionality reduction: recombining the original data set to exploit any redundancy present, and forming a new, smaller set of variables that contain the same information as the original data.

Dimensionality reduction has been a topic of study for a long time in different disciplines like statistics, bioinformatics, and meteorology, which deal with large number of samples over long time periods. In fact, the classical principle component analysis (PCA, also known as the Karhunen-Loève transform) used for data reduction dates back to 1901, developed by statisticians to handle large time series data [Pearson, 1901]. PCA can be understood as finding a set of new basis vectors for the original data so that they can be re-represented in this basis using much fewer data dimensions, while only incurring a very low error. In a mean-square error sense, PCA is the optimal linear dimensionality reduction method. However, though it is based on the sound assumption of maximizing the retention of the variance of the original data, the technique involves intensive computations involving covariance matrices of the same size as the data, and are therefore not suitable in today's big data regime with data sizes of the order of many million dimensions. There are several extensions to PCA that have been developed over the years (kernel PCA [Schölkopf et al., 1998], probabilistic PCA [Tipping and Bishop, 1999], etc.), but we have not studied these methods in the context of radio-interferometric imaging as the data dimensions we aim to handle preclude any possible extensive computational effort.

The computational challenge with radio-interferometric imaging is demonstrated in Fig. 4.2, which illustrates two separate effects related to big data sizes and image reconstruction quality. Overall, with increasing data size we see an increase in the time taken for an iterative image reconstruction algorithm to converge to a solution – this increase is directly proportional to the data size used for reconstruction (we use a non-parallelized version of SDMM for the experiment here, in contrast to the case shown in Fig. 4.1). We performed image reconstruction from sets of simulated data with varying noise levels and, interestingly, we notice that the increase in the time taken for image reconstruction is much more drastic for data with very low noise levels, as compared to noisier data. This implies that under a high-sensitivity, high-accuracy regime, large data sets incur a more significant time penalty due to the more stringent convergence criteria for iterative imaging algorithms. This effect of increasing image reconstruction time was confirmed as part of preliminary studies for dimensionality reduction techniques undertaken by Kartik et al. [2015].

Dimensionality reduction would play a critical role in such low-noise situations where collecting more data is essential for science objectives, but where, at the same time, we would like to avoid impractical image reconstruction due to the behaviour shown in Fig. 4.2. In section 4.4,

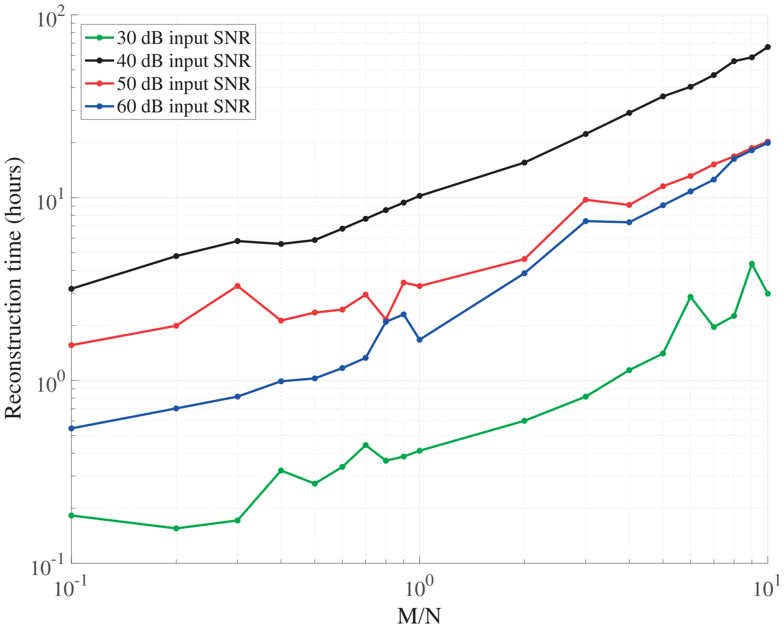


Figure 4.2 – Time taken by an SDMM-based image reconstruction algorithm, shown here for increasing data sizes on independent simulated data with different noise levels. The x and y axes are in $\log_{10}$ scale.

we describe studies on different dimensionality reduction techniques that were undertaken in the context of radio-interferometric imaging with big data.

4.3 Dimensionality reduction approaches

Data dimensionality reduction is a non-trivial problem because we need to balance a few antagonistic qualities of the reduction process. On the one hand, we would like to preserve all information available in the original full-sized data set. On the other hand, we would like to minimize the number of data points, in order to have reasonable timeframes for image reconstruction. Fewer data points also imply a lower computational load in terms of memory and storage, which is one of the main goals of data dimensionality reduction. The challenge is to find a good balance between these two opposing ‘requirements’, while staying competitive in terms of implementation and application.

We have tried to devise several dimensionality reduction methods of varying complexity within the context of radio-interferometric imaging. This was done partly to comprehensively cover different approaches to this non-trivial problem and partly as sanity checks for the correctness of the solutions that subsequent imaging algorithms would recover. In the following sections of this chapter, we shall briefly describe several of these attempts, and place them in the more general context of data reduction as used in a variety of computing and signal processing tasks.

4.3.1 Random projections and 'sketching'

Projecting higher-dimensional data into a lower-dimensional subspace is a classical linear reduction method. The popular PCA is essentially a projection into a lower-dimensional subspace spanned by the new basis vectors we choose, as mentioned in section 4.2. Another projection method, independent component analysis (ICA, Roberts and Everson [2001]), can be seen as a more general version of PCA, replacing the PCA requirement of uncorrelated data dimensions by independent dimensions. Fodor [2002] presents an overview of standard dimensionality reduction techniques in statistics, signal processing and machine learning, including PCA, ICA and other methods like factor analysis and projection pursuit. A common thread in these standard reduction methods is that the projection is performed through deterministic basis vectors, chosen on a case-by-case basis. While this type of projection method works well, in principle, for the particular situation and a given fixed data set, it is not well-suited for very large data dimensions due to computational limitations. Moreover, finding a new set of basis vectors every time a different projection needs to be performed is also computationally cumbersome.

Projections using random matrices, or 'sketches', address most of these issues and, surprisingly, continue to guarantee high levels of information retention, at least in terms of signal recovery errors. Random projection methods have been deployed very successfully to reduce data dimensionality in a wide variety of cases, and their success is mainly owed to the ease of construction of the projection matrices and the computational ease of applying the projection. Random projections were introduced by Kaski [1998] as a pre-processing step before a clustering algorithm on text documents. It was proposed as a more suitable alternative to PCA due to the 'large' data dimension (6000) that PCA could not handle.

Random projections use a matrix with column entries pulled from an independent and identically distributed (i.i.d.) random variable. The surprising observation that low-dimensional data obtained after random projections are able to reasonably represent the underlying signal can be explained as follows: If we denote the random projection matrix by $\mathbf{R}_\mathrm{p} \in \mathbb{C}^{M_L \times M}$, we can see that it reduces the dimension of a data vector $\boldsymbol{y} \in \mathbb{C}^M$ to a lower value M_L through a simple application $\mathbf{R}_\mathrm{p}\boldsymbol{y}$. This reduced-dimensional data can be rewritten as $\sum_i y_i \mathbf{R}_{\mathrm{p}_i}$, where y_i and $\mathbf{R}_{\mathrm{p}_i}$ are the i^{th} element of $\boldsymbol{y}$ and i^{th} column of $\mathbf{R}$ respectively. We can then see that while the basis vectors of the original data $\boldsymbol{y}$ were obviously the orthonormal unit vectors of $\mathbb{C}^M$, each of those unit dimensions are replaced in the reduced data by the columns of $\mathbf{R}$. This means that the closer to orthonormal the columns of $\mathbf{R}$, the more similar the data vectors before and after dimensionality reduction. As $\mathbf{R}_\mathrm{p}$ contains random elements, if the original dimension M is high enough, randomly choosing directions would result in vectors that are reasonably close to being orthonormal, and therefore act as a basis (up to a good approximation). It is this combination of a high enough starting dimension coupled with random projection that allows similarity between vectors to be preserved through a dimensionality reduction. A more detailed description of this intuition, along with related mathematical background, is presented by Kaski [1998]. This concept of preserving 'similarities' is central to other dimensionality reduction methods as well, and we shall revisit it in a more formal sense in the following section as part of the discussion on the Johnson-Lindenstrauss lemma.

4.3.2 Johnson-Lindenstrauss embedding

Recent work in the literature drawing parallels between dimensionality reduction and compressed sensing [Baraniuk et al., 2008, Krahmer and Ward, 2011] can be traced back to the seminal work on the Johnson-Lindenstrauss (JL) lemma [Johnson and Lindenstrauss, 1984]. The JL lemma claims that we can find an embedding of any n points in high-dimensional Euclidean space to a low-dimensional subspace with $\mathcal{O}(\epsilon^2 \log n)$ dimensions. The JL lemma has been widely used as a tool for dimensionality reduction in many fields, owing to its applicability with most Gaussian or orthogonal random projections. It essentially provides conditions that the distance, or degree of 'similarity', between two points that are in a high-dimensional Euclidean space continues to be respected in the low-dimensional subspace under a JL embedding. So, given points $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{C}^N$, the JL lemma states that an embedding $e : \mathbb{C}^N \to \mathbb{C}^{M_L}$ exists that maps them to a lower-dimensional subspace such that

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \leq \|e(\boldsymbol{x}_i) - e(\boldsymbol{x}_j)\|_2^2 \leq (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2. \tag{4.1}$$

Considering the set of differences $D = \boldsymbol{x}_i - \boldsymbol{x}_j$ and a linear embedding e that is equivalent to applying a matrix $\mathsf{E} \in \mathbb{C}^{M_L \times N}$, equation (4.1) can be re-expressed as

$$(1-\epsilon)\|\boldsymbol{d}\|_2^2 \leq \|\mathsf{E}\boldsymbol{d}\|_2^2 \leq (1+\epsilon)\|\boldsymbol{d}\|_2^2 \qquad \forall \boldsymbol{d} \in D. \tag{4.2}$$

We cannot help but note the similarity between the formulation of the JL lemma as given here in equation (4.2) and the definition of the RIP in equation (3.6). Indeed, it has been shown that matrices satisfying equation (4.2) also satisfy the RIP (of the appropriate order). Baraniuk et al. [2008] provide a proof of this equivalence for random matrices. These results, therefore, indicate that the JL lemma implies the RIP. The connection between the JL lemma and compressed sensing is established through this equivalence, and provides a perspective for dimensionality reduction methods that would satisfy the CS theory requirements.

4.4 Dimensionality reduction of radio-interferometric data

The convex optimization algorithms mentioned in section 3.2 that are employed for radio-interferometric imaging amount to a non-linear iterative reconstruction involving repeated application of the measurement basis $\boldsymbol{\Phi}$, the sparsity basis $\boldsymbol{\Psi}$, and their adjoint operators (Yang and Zhang [2011], Carrillo et al. [2014] provide more detailed overviews of these algorithms). This means that the recovery methods that lead to a solution of equation (3.8) heavily rely on fast implementations of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. In the context of our imaging techniques for the analysis-based approach with a predefined concatenation of bases $\boldsymbol{\Psi}$ whose dimensionality is proportional to the image size, we can see that as the data size increases, so does the memory and computing requirement to manipulate and perform operations with $\mathbf{\Phi}$, whose dimensionality is proportional to data size.

As mentioned in section 3.1.2, a fast and scalable implementation of the measurement operator $\mathbf{\Phi}$ is critical for the viability of convex optimization-based image recovery methods. The complexity of applying $\mathbf{\Phi}$ $(= \mathsf{G}\overline{\mathsf{FZ}})$ in the optimization algorithm can increase rapidly with increasing data size. Given an image of N pixels, and M visibilities obtained with a non-uniform Fourier transform involving an interpolation kernel of size $k \times k$, the asymptotic complexity of applying $\mathbf{\Phi}$ is seen to be $\mathcal{O}(Mk^2 + N\log N)$, since it is the complexity of matrix operations involving the matrix G with M rows of k^2 non-zero entries, added with the complexity of an N-sized FFT. As seen from the complexity, this image recovery solution is very demanding for large values of M, both in terms of computing time and memory.

Dimensionality reduction of radio-interferometric data is therefore particularly essential for enabling these imaging methods to continue working in a big data regime. The desirability of dimensionality reduction springs from its ability to (i) reduce the data *size*, thereby decreasing memory requirements, (ii) keep the measurement operator *fast*, thus reducing the computing time, (iii) preserve compressed sensing properties (notably the NSP, RIP) of the measurement operator (to the extent that these properties are satisfied for $\mathbf{\Phi}$ as given in equation (2.10)) to guarantee accurate signal reconstruction, and (iv) preserve the i.i.d. Gaussian properties of the original measurement noise, in order to facilitate an easy computation of the data fidelity in our convex optimization algorithms through an ℓ_2-norm. In the setting of the discretized radio-interferometric measurement equation as given by equation (2.8), we understand dimensionality reduction as the process of linearly mapping the higher-dimensional visibilities vector $\boldsymbol{y} \in \mathbb{C}^M$ to a lower-dimensional vector $\boldsymbol{y}' \in \mathbb{C}^{M_L}$ such that $M_L \ll M$. As touched upon in section 4.3.1, this is typically achieved by applying an 'embedding' operator $\mathsf{R} \in \mathbb{C}^{M_L \times M}$, so that $\boldsymbol{y}' = \mathsf{R}\boldsymbol{y}$ is of dimension M_L. Applying such an embedding operator to equations (2.10) and (2.8), we obtain the full measurement operator

$$\mathbf{\Phi}' = \mathsf{R}\mathbf{\Phi}, \tag{4.3}$$

and the reduced inverse problem

$$\boldsymbol{y}' = \mathsf{R}\mathbf{\Phi}\boldsymbol{x} + \mathsf{R}\boldsymbol{n}. \tag{4.4}$$

The choice of R is critical as it affects not just the distortion of the visibilities $\boldsymbol{y}$ but also the properties of $\mathbf{\Phi}$ that originally led to guaranteed image recovery through compressed sensing-based reconstruction methods. Additionally, R modifies the original noise vector $\boldsymbol{n}$. In our setting for radio-interferometric imaging, $\boldsymbol{n}$ is assumed to be uncorrelated, having i.i.d. zero-mean Gaussian components with a variance given by σ_n. This assumption brings some advantages to the convex optimization problem formulation, especially related to the ease of maximizing data fidelity as discussed in section 3.1.5. After applying a general embedding operator R, the ‘embedded’ noise $\boldsymbol{n}' = \mathsf{R}\boldsymbol{n}$ has a covariance matrix

$$\mathsf{C}_{\boldsymbol{n}'} = \mathbb{E}\left[\boldsymbol{n}'(\boldsymbol{n}')^{\dagger}\right] = \sigma_n^2 \mathsf{R}\mathsf{R}^{\dagger}, \tag{4.5}$$

where $\mathbb{E}$ denotes the expected value. $\mathsf{C}_{\boldsymbol{n}'}$ is not necessarily diagonal; i.e., the embedded noise $\boldsymbol{n}'$ is, in general, correlated.

4.4.1 State of the art: visibility averaging

The standard dimensionality reduction method used in radio interferometry today is visibility averaging. Averaging is performed either over time, approximating several measurements on an elliptical track of a baseline by one data point, or over spectral channels at a given time for a given baseline, or both. It is a very simple and effective approach to reduce data size, as illustrated here in its matrix form as

$$\mathsf{R}_{\text{avg}} = \frac{1}{T}\begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}, \tag{4.6}$$

where T is the number of measurements over which the averaging is performed.

Visibility averaging is widely used thanks to its scalability and ease of use. However, averaging also leads to several issues in the imaging methods that follow [Bridle and Schwab, 1999]. The first issue is the ‘misuse’ of dimensionality reduction as it is performed currently. Radio interferometry data are averaged with the goal of reducing their size, but once the averaged data are obtained, subsequent imaging methods use the original measurement operator $\boldsymbol{\Phi}$ and not the appropriate, combined measurement operator $\boldsymbol{\Phi}' = \mathsf{R}_{\text{avg}}\boldsymbol{\Phi}$, thus completely ignoring the presence of R_{avg} in the measurement model. This mismodelling effectively treats the averaged data as if they were the obtained measurements, and this has a detrimental effect on the quality of the reconstructed image. On the other hand, in the case of appropriately handling R_{avg} in the measurement model, averaging over a small number of visibilities can be seen as a convolution in uv space with a boxcar function of appropriate width (and then sampling at the mid-point), which corresponds, in image space, to multiplication of the image with sinc-like tapering window. This tapering window is different for different baselines, and results in artefacts in the image recovered from such averaged data. An immediately apparent effect of averaging is also the disappearance of the highest frequency component of the data. There are proposals found in the literature to mitigate the adverse effects of averaging [Offringa et al., 2012, Parsons et al., 2016, Atemkeng et al., 2016], but the proposals have tradeoffs, some resulting in lower dynamic range, and others resulting in mismatched processing with calibration procedures. Additionally, averaging has limits on how low the reduced data vector size can be, as it depends on the conditions of data acquisition, namely, the total number of points available either on a track in the uv plane, or across spectral channels. We discuss the dimensionality reduction performance of R_{avg} in greater detail in chapter 6 which addresses image reconstruction using real telescope data.

Noting the various side-effects linked with the state of the art averaging method, we aim to find a more appropriate dimensionality reduction method that would minimize artefacts in the recovered image, while providing a robust, computationally efficient implementation. In the next sections and chapters, we focus on our search of alternate dimensionality reduction methods and evaluate their effectiveness using simulated and real data.

4.4.2 Preliminary studies on dimensionality reduction

The attractive nature of random projections [Achlioptas, 2003] leads us to attempt dimensionality reduction of interferometric data with specific constructions of matrices based on different random distributions. In this section we show results for images reconstructed from simulated data of varying sizes, with their dimensionality reduced using different random projection methods: (i) a Gaussian random projection, (ii) a subsampled Hadamard transform, and (iii) a very sparse modified Bernoulli random projection.

The theoretically ideal random projection – from a compressed sensing perspective – would be provided by a Gaussian random matrix. For a Gaussian random matrix R_G, the full measurement operator $\boldsymbol{\Phi}' = \mathsf{R}_G\boldsymbol{\Phi}$ becomes a Gaussian operator, and would thus continue to retain the original NSP satisfied by $\boldsymbol{\Phi}$, thus guaranteeing accurate signal recovery (see section 3.1). Also thanks to R_G, elements of the embedded noise $\boldsymbol{n}'$ would, on average over realizations in $\mathbb{C}$, follow an i.i.d. Gaussian distribution even if the original noise were not i.i.d. Gaussian [Kartik et al., 2015]. As described in section 3.1.5, a decorrelated i.i.d. Gaussian noise vector greatly simplifies the convex optimization problem formulation by allowing an analytically computable upper bound on the data fidelity term based on (embedded) noise statistics.

We indeed see the advantages of such an embedding on simulations performed using SDMM for image reconstruction. We simulated data using a 128×128-pixel test image of the HII region of the M31 galaxy (Fig. 4.3). Complex visibilities were generated using synthetic, incomplete uv coverages generated with random variable density sampling profiles. These profiles attempt to mimic telescope sampling patterns by imposing denser sampling at lower spatial frequencies (corresponding to shorter baselines) and sparser sampling at higher spatial frequencies (from fewer, longer baselines) – this is also illustrated in Fig. 4.3. We used a Gaussian profile for the sampling pattern in our experiments, and removed the (u, v) point $(0, 0)$ to follow standard practice in radio interferometry.[1] We generated visibility sets of varying sizes, ranging from image size to 10 times larger. Each set of visibilities was corrupted by adding Gaussian noise,

---

[1] Data at this location are a result of auto-correlation and are dominated by noise. Ignoring this (u, v) point leads to absence of total flux information, which is why radio astronomers separately use single-dish telescopes to measure this component.

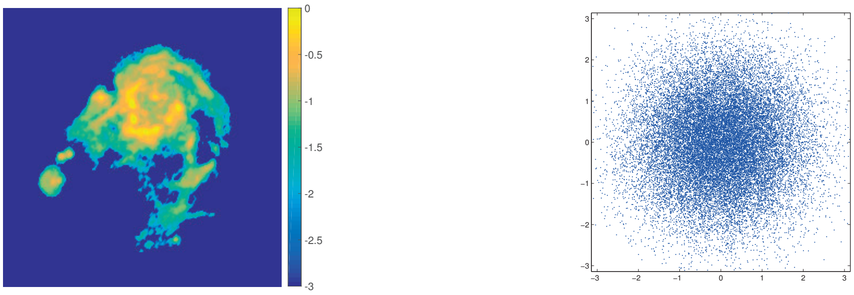


Figure 4.3 – Left: M31 test image ($\log_{10}$ scale) used in simulations. A 128×128-pixel image is used in the simulations here. Higher resolution images are used for experiments described in chapter 5. Right: Test uv coverage with a variable density sampling profile, showing sparser sampling with increasing spatial frequency.

giving a signal-to-noise ratio (SNR) of 40 dB for each visibility set. We define the 'input' SNR as

$$\mathrm{SNR_i} = 20 \log_{10} \frac{\|\boldsymbol{y}_0\|_2}{\|\boldsymbol{n}\|_2} \tag{4.7}$$

with $\boldsymbol{y}_0 = \boldsymbol{\Phi x}$ being the true visibilities vector before adding the noise vector $\boldsymbol{n}$. To test Gaussian random projections, data of different sizes were embedded to image size using different Gaussian random matrices, and reconstruction was then performed with SDMM using these image-sized data. For comparison without any dimensionality reduction, the original data was also used directly to reconstruct images using SDMM. Image reconstruction quality was also measured with an 'output' SNR value,

$$\mathrm{SNR}_{\hat{\boldsymbol{x}}} = 20 \log_{10} \frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2}, \tag{4.8}$$

where $\hat{\boldsymbol{x}}$ is the reconstructed image while $\boldsymbol{x}$ is the 'ground truth' image available to us.

We see from Fig. 4.4 that images reconstructed from lower-dimensional data obtained using an $N \times M$-sized Gaussian random matrix (red curve) follow the SNR values of images reconstructed from the full data set without any size reduction (black curve). This is especially true when starting from high-dimensional visibility vectors (e.g., $M = 10N$), where the random projection has more initial information to use. The discrepancy between the reconstruction qualities is wider for lower starting data sizes because of unavailability of more information for the random projection to be effective. We see that, in general, starting with more data in hand leads to higher output SNR values, reaching the point where we can claim denoising effects even after reducing the data dimensionality down by a factor 10 – the output SNR reaches ≈45 dB for an input SNR of 40 dB.

While Gaussian random matrices are effective as a dimensionality reduction method in terms of maintaining high SNR for reconstructed images, this quality comes at a computational cost [Kartik et al., 2015]. The main issue is that we lose the fast application of the original measurement operator $\boldsymbol{\Phi}$. The combined measurement operator $\boldsymbol{\Phi}'$ is no longer sparse and

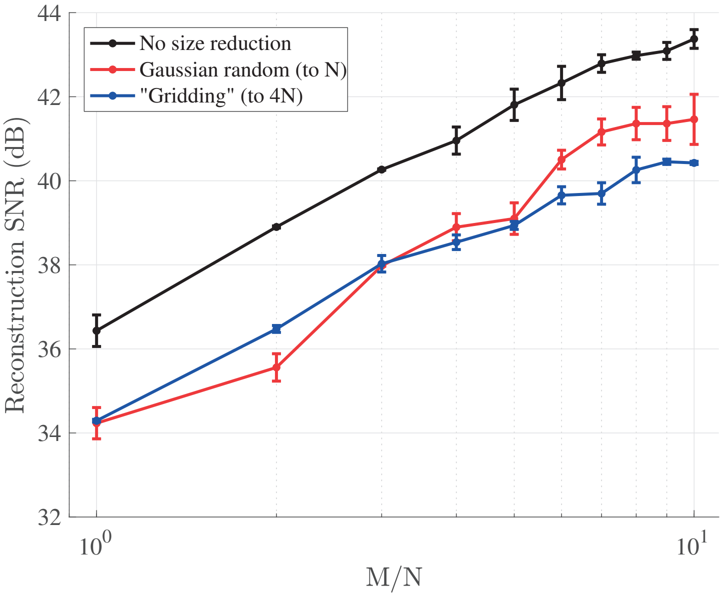


Figure 4.4 – SNR of reconstructed images obtained from different dimensionality reduction methods on the same simulated data. The initial data size ranges from $M = N$ to $M = 10N$. Gaussian random projections always resulted in N-sized data; 'Gridding' always resulted in $4N$-sized data, from which images were reconstructed using SDMM. The x and y axes are in $\log_{10}$ scale.

cannot be applied as a sequence of fast submodules. The asymptotic complexity of applying $\mathbf{\Phi}'$ can be seen to be $\mathcal{O}(M_L N)$ since applying the Gaussian random matrix leads to a fully dense matrix of size $M_L \times N$. As M_L is typically some proportion p of the image size N, the asymptotic complexity for the Gaussian random matrix is then $\mathcal{O}(pN^2)$. In a big-data regime, next-generation radio interferometers like the SKA will produce gigapixel images, giving us values of $N \approx 10^9$, whereas data sizes can range up to $M \approx 1000N \approx 10^{12}$. In this context, the asymptotic complexity $\mathcal{O}(pN^2)$ for Gaussian random matrices is clearly much worse than the original asymptotic value which was dominated by $\mathcal{O}(Mk^2)$, since $M \ll N^2$. This takes a heavy toll on the image reconstruction algorithms which repeatedly apply $\mathbf{\Phi}'$ and its adjoint, and the time taken for algorithm convergence is prohibitively high, rendering Gaussian random projections impractical.

The alternative in this case would be to apply the Gaussian random matrix separately, without creating a combined measurement operator $\mathbf{\Phi}'$. However, this defeats the purpose of dimensionality reduction, since the subsequent imaging algorithm would then involve repeated computations in the higher dimension M, which is precisely what we would like to avoid. Thus, the Gaussian random matrix approach is computationally infeasible independent of how we approach its application.

Several other random projections have been shown in the signal processing literature to be effective at preserving distances in a lower-dimensional space. For instance, a 'spread-spectrum' like operator $\mathsf{R} = \mathsf{M}\bar{\mathsf{F}}\mathsf{D}_{\pm 1}$, where $\mathsf{D}_{\pm 1}$ is a diagonal random sign matrix with entries ± 1, $\bar{\mathsf{F}}$ the discrete Fourier transform operator, and M an $M_L \times N$ random selection matrix which embeds the data vector to the final size M_L. The spread-spectrum technique was shown by Wiaux et al. [2009b] to be optimal in a compressed sensing context, as it makes the sensing and sparsity basis maximally incoherent, leading to guaranteed signal recovery. Carrillo et al. [2013] used the SARA prior on simulated data with random Gaussian and spread-spectrum acquisition schemes to show improved signal recovery compared to other priors. We note also that the spread-spectrum technique is similar to the subsampled randomized Hadamard transform (SRHT, Krahmer and Ward [2011], Tropp [2011]) and the random convolution technique for CS proposed by Romberg [2009]. The SRHT is proposed as the embedding $\mathsf{R} = \mathsf{S}\mathsf{H}_{\mathsf{W}}\mathsf{D}_{\pm 1}$, where $\mathsf{H}_{\mathsf{W}} \in \mathbb{R}^{N\times N}$ is the Walsh-Hadamard transform and $\mathsf{S} \in \mathbb{R}^{M_L\times N}$ is a uniformly random subsampling operator to reduce vectors to the final size M_L.

As in the case of Gaussian random projections, these embeddings suffer from the same crippling effect of leading to a dense matrix resulting in prohibitively high image reconstruction time, or needing to perform a multi-step measurement operation which involves repeated computations in the high dimension, which is undesirable. We therefore turned our attention to very sparse random projections involving Bernoulli ensembles and related embeddings. Section 4.3.1 provides the intuition behind why sparse random projections would continue to provide good approximations of higher-dimensional subspaces (on *expectation*). In addition, sparse projections have a great advantage in low computing times for the application of the operators and, consequently, practical image reconstruction algorithms. However, when looked at through a compressed sensing perspective, these combined measurement operators $\boldsymbol{\Phi}' = \mathsf{R}\boldsymbol{\Phi}$ lack the properties required by CS theory to provide accurate signal recovery. Moreover, it is non-trivial to directly prove if $\boldsymbol{\Phi}'$ in these cases indeed satisfies the RIP [Bandeira et al., 2013], therefore throwing into doubt the feasibility of using the CS-based image reconstruction methods outlined in section 3.3.1 and 3.3.2.

We ran simulations with settings identical to the tests described earlier in this section for Gaussian random projections, which confirmed these suspicions. We implemented the specific 'very sparse random projection' proposed by Li et al. [2006], which uses a random matrix with entries in $\{-1, 0, 1\}$ and claims to provide a speed-up over the classical sparse random projection proposed by Achlioptas [2003]. The reconstruction quality using very sparse random projections is far worse than any other dimensionality reduction method. This may be attributed to the fact that the specific random matrices used in our tests did not fit in the CS regime, namely the RIP/NSP, and consequently, ℓ_1-minimization algorithms were ill-suited to recover the images from data embedded through these matrices. Output SNR values of images reconstructed from low dimensional data obtained using very sparse random projections remained below 25 dB for all initial data sizes, and have been omitted in Fig. 4.4 to enable us to clearly compare the distinct performance profiles of the other dimensionality reduction methods.

We also compared these 'new' data embedding techniques with a standard method of reducing dimensionality in radio interferometry: ‘gridding’ continuous visibilities to the discrete Fourier grid. Our version of gridding was based on the classical gridding operation performed as part of the standard CLEAN imaging process, and we implemented an efficient application of the gridding operator by employing a precomputed matrix for the embedding, as described by Sullivan et al. [2012]. As Fig. 4.4 shows, gridding (blue curve) works very well when compared to both the Gaussian random projection method as well as to images reconstructed without any dimensionality reduction. We continued to investigate gridding as a viable embedding, and developed our own extended version of gridding which includes weighting and subsampling steps, with very encouraging performance. More details on our implementation and related results are described in chapter 5.

5 A Fourier model for dimensionality reduction

5.1 CS-friendly dimensionality reduction

Our attempts at approaching dimensionality reduction from a compressed sensing perspective lead us to consider the NSP and the RIP of the full measurement operator. With the exception of random matrices, constructing fast matrices satisfying the RIP is known to be non-trivial – although there have been recent attempts towards addressing this challenge [Nelson et al., 2014]. Also, verifying the RIP for deterministic matrices is NP-hard, as shown by Bandeira et al. [2013]. So the idea is to devise an embedding operator that reduces the dimensionality of the measurements while preserving the NSP of the original measurement operator, thus maintaining the same compressed sensing-based guarantees on recovering the image.

We note that many state-of-the-art imaging techniques in radio interferometry include a 'gridding'-like subroutine, mapping continuous visibilities to the discrete Fourier grid with an operator similar to $\mathbf{R} = \boldsymbol{\Phi}^\dagger$ [Dabbech et al., 2015, Li et al., 2011]. Typically, this is a gridding to the discrete Fourier grid through $\mathbf{G}^\dagger$, or a mapping back to image space through $\boldsymbol{\Phi}^\dagger$. Gridding has been studied and developed further by Sullivan et al. [2012] with the introduction of the 'Fast Holographic Deconvolution' technique; this technique leverages the lossless information property [Tegmark, 1997] that is being used to reduce Cosmic Microwave Background data, and introduces the *Holographic Mapping* function $\mathbf{H} = \mathbf{G}^\dagger\mathbf{G}$. $\mathbf{H}$ models the mapping between a continuous visibility and the corresponding equivalent in the gridded, 'holographic' map without having to go through separate interpolation and gridding steps – which are the most time consuming parts of standard imaging techniques. A pre-computed holographic matrix $\mathbf{H}$ is stored before image reconstruction starts, and therefore provides a way to quicken the imaging process. Additionally, the compact support of the interpolation kernel present in each row of the matrix $\mathbf{G}$ ensures that $\mathbf{H}$ remains sparse, so its repeated application is also not a hindrance to the imaging technique.

We build on these developments and propose a modified gridding-based dimensionality

---

This chapter contains work published by Kartik et al. [2017a,b].

reduction method that closely follows the state-of-the-art data preprocessing, while resulting in reasonable image reconstruction quality from gridded data. In this chapter we see that gridding is an appropriate technique to reduce data dimensionality while maintaining information content, and with the use of a holographic mapping it can be incorporated in imaging techniques without incurring a large cost in terms of image reconstruction time. Another major contribution presented here takes an entirely different approach, resulting in a post-gridding reduction technique that introduces an additional step to achieve reducing dimensionality to much lower sizes than with our gridding-based method. This is presented in the next section.

5.1.1 Singular vector space embedding

Optimal dimensionality reduction model

An ideal dimensionality reduction method would result in a final data dimension as small as possible while simultaneously guaranteeing accurate image reconstruction by retaining the NSP of the original measurement operator $\boldsymbol{\Phi}$. The null space of $\boldsymbol{\Phi}$ arises from the incomplete Fourier coverage that forms all visibilities. Taking inspiration from PCA-based representations (which are known to retain most of the information in the original data, as mentioned in section 4.2), we observe the singular value decomposition (SVD) of $\boldsymbol{\Phi}$ given by

$$\boldsymbol{\Phi} = \mathsf{U}\boldsymbol{\Sigma}\mathsf{V}^{\dagger}, \tag{5.1}$$

where $\mathsf{U} \in \mathbb{C}^{M\times M}$, $\mathsf{V} \in \mathbb{C}^{N\times N}$ are unitary matrices and $\boldsymbol{\Sigma} \in \mathbb{C}^{M\times N}$ is a rectangular diagonal matrix containing the singular values of $\boldsymbol{\Phi}$. Here we note that the existence of the null space of $\boldsymbol{\Phi}$ implies that some singular values are necessarily zero. In fact, the singular values $\boldsymbol{\Sigma}_i$ occupy a continuous spectrum of values, with large values corresponding to Fourier grid points with contribution to multiple interpolation kernels present in G, and gradually decreasing to the minimum value of zero corresponding to Fourier grid points with no such contribution, thus leading to an incomplete uv coverage. We can see that retaining the non-zero singular values of $\boldsymbol{\Phi}$ effectively retains the orthogonal complement of the null space of $\boldsymbol{\Phi}$. Following this, we rewrite the SVD as

$$\boldsymbol{\Phi} = \mathsf{U}_0\boldsymbol{\Sigma}_0{\mathsf{V}_0}^{\dagger}, \tag{5.2}$$

where $\mathsf{U}_0 \in \mathbb{C}^{M\times N_0}$, $\boldsymbol{\Sigma}_0 \in \mathbb{C}^{N_0\times N_0}$ and $\mathsf{V}_0 \in \mathbb{C}^{N\times N_0}$ are truncated versions of U, $\boldsymbol{\Sigma}$ by only retaining columns (rows for V) corresponding to non-zero singular values of $\boldsymbol{\Phi}$. Clearly, the number of non-zero singular values is $N_0 \leq \min(N, M)$ since $\boldsymbol{\Sigma} \in \mathbb{C}^{M\times N}$.

An optimal dimensionality reduction operator to be applied on $\boldsymbol{\Phi}$ would then be an embedding on its left singular vectors that correspond to non-zero singular values, since the column space of $\boldsymbol{\Phi}$ is retained through these left singular vectors and thus no information is lost. This embedding is given by:

$$\begin{aligned}\mathbf{R}_{\text{sing}-\text{o}} &= \mathbf{U}_0{}^{\dagger} \\ &= \mathbf{\Sigma}_0^{-1}\mathbf{V}_0{}^{\dagger}\mathbf{\Phi}^{\dagger}.\end{aligned} \tag{5.3}$$

We note that this is a mapping onto the range space of $\mathbf{\Phi}$. The full measurement operator therefore reads as a weighted subsampling in the right singular vector basis:

$$\mathbf{\Phi}'_{\text{sing}-\text{o}} = \mathbf{\Sigma}_0\mathbf{V}_0{}^{\dagger}. \tag{5.4}$$

The corresponding 'embedded' noise $\mathbf{U}_0{}^{\dagger}\boldsymbol{n}$ has a covariance matrix

$$\begin{aligned}\mathbf{C}_{n'} &= \sigma_n^2\mathbf{U}_0{}^{\dagger}\mathbf{U}_0 \\ &= \sigma_n^2\mathbf{I}.\end{aligned} \tag{5.5}$$

This follows from equation (4.5), since columns of $\mathbf{U}$ are orthonormal by definition. The noise thus remains fully decorrelated after dimensionality reduction, which allows us to continue using an ℓ_2-norm as the data fidelity term in the minimization algorithm, as explained in section 3.1.5.

Put differently, the ideal dimensionality reduction involves a gridding-like operation performed in radio interferometry to obtain the dirty image (shown here by $\mathbf{\Phi}^{\dagger}$), followed by an embedding on to the right singular vectors of $\mathbf{\Phi}$ corresponding to non-zero singular values, and finally followed by a weighting operation with the inverse of the non-zero singular values of $\mathbf{\Phi}$.

In theory, therefore, the ideal dimensionality reduction $\mathbf{R}_{\text{sing}-\text{o}}$ reduces data to a dimension $N_0 \leq N$. It ensures that the full measurement operator $\mathbf{\Phi}'_{\text{sing}-\text{o}}$ preserves the null space of $\mathbf{\Phi}$, therefore retaining any original NSP of $\mathbf{\Phi}$. It also induces a decorrelated noise in the reduced dimension, thus enabling the minimization algorithm to use an ℓ_2-norm of the noise for the data fidelity term. In reality, however, this operator $\mathbf{R}_{\text{sing}-\text{o}}$ is difficult to implement since the SVD is computationally expensive, with an asymptotic complexity of $\mathscr{O}(N^3)$ [Golub and van Loan, 1996]. Additionally, since $\mathbf{R}_{\text{sing}-\text{o}}$ may not have a guaranteed fast implementation, applying it iteratively in our minimization algorithms would also be prohibitively expensive. This renders the optimal reduction method impractical. We propose to get around this limitation by building an approximate version $\mathbf{R}_{\text{sing}}$ of $\mathbf{R}_{\text{sing}-\text{o}}$ that can be readily computed and applied.

Approximate Fourier reduction model

In finding a valid approximation of the ideal dimensionality reduction given by equation (5.3), we attempt to approximate the unitary matrix $\mathbf{V}$. We can note that $\mathbf{V}$ is in fact the eigenbasis of $\mathbf{\Phi}^{\dagger}\mathbf{\Phi}$, since it contains the right singular vectors of $\mathbf{\Phi}$ as defined in Section 5.1.1. To understand the eigenbasis of $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$, we probe its structure, expanding it to its constituent operators as defined in equation (2.10). This gives us

$$\begin{aligned}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} &= (\mathbf{G}\overline{\mathbf{F}}\overline{\mathbf{Z}})^\dagger(\mathbf{G}\overline{\mathbf{F}}\overline{\mathbf{Z}}) \\ &= (\overline{\mathbf{Z}}^\dagger\overline{\mathbf{F}}^\dagger)(\mathbf{G}^\dagger\mathbf{G})(\overline{\mathbf{F}}\overline{\mathbf{Z}}). \qquad (5.6)\end{aligned}$$

The central term in equation (5.6) is the holographic map $\mathbf{H} = \mathbf{G}^\dagger\mathbf{G}$, comprising individual elements $(\mathbf{G}^\dagger\mathbf{G})_{ij}$ that denote the simultaneous contributions of different interpolation kernels that would map continuous visibilities onto the pixel (i, j) on the discrete Fourier grid. Since each visibility is obtained by integrating a small region of the uv plane, the interpolation kernels have compact support (e.g., the 8×8 Kaiser-Bessel kernels used to calculate the non-uniform Fourier transform in our simulations). Thus, the simultaneous contributions of different interpolation kernels are largely limited to small areas of overlapping support, and consequently limited to having significant contribution only for pixels (i, j) where i, j are of similar value. In other words, the largest values of $\mathbf{G}^\dagger\mathbf{G}$ are on and immediately around its diagonal. This is also seen through numerical results as shown in Fig. 5.1, where the illustrations for $\mathbf{G}^\dagger\mathbf{G}$ can be seen to be extremely close to a diagonal matrix. It should be noted that here we implicitly assume that there are no DDEs and that the antenna array is coplanar ($w = 0$). If these assumptions become invalid, the interpolation kernels present as rows of $\mathbf{G}$ can no longer be simply represented with compact support, and $\mathbf{G}^\dagger\mathbf{G}$ no longer remains overwhelmingly diagonal.

Now we prepend and append $\mathbf{F}^\dagger\mathbf{F}$ to equation (5.6) – the crucial observation being that $\mathbf{F}$ is an image-sized Fourier transform as opposed to the oversampled Fourier transform $\overline{\mathbf{F}}$. $\mathbf{F}$ can then be expressed as $\mathbf{F} = \mathbf{Z}^\dagger\overline{\mathbf{F}}\overline{\mathbf{Z}}$. equation (5.6) can then be rewritten as

$$\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} = \mathbf{F}^\dagger\left[(\mathbf{F}\overline{\mathbf{Z}}^\dagger\overline{\mathbf{F}}^\dagger)(\mathbf{G}^\dagger\mathbf{G})(\overline{\mathbf{F}}\overline{\mathbf{Z}}\mathbf{F}^\dagger)\right]\mathbf{F}. \qquad (5.7)$$

We note that the term $\mathbf{F}\overline{\mathbf{Z}}^\dagger\overline{\mathbf{F}}^\dagger = \mathbf{Z}^\dagger\overline{\mathbf{F}}\overline{\mathbf{Z}}\overline{\mathbf{Z}}^\dagger\overline{\mathbf{F}}^\dagger$ as a whole performs a convolution with the inverse Fourier transform of $\overline{\mathbf{Z}}\overline{\mathbf{Z}}^\dagger$. Since $\overline{\mathbf{Z}}\overline{\mathbf{Z}}^\dagger$ is – within limits of the scaling introduced in $\overline{\mathbf{Z}}$ by $\mathbf{D}_\mathrm{R}$ (see Section 2.3) – a partially distorted version of a two-dimensional pulse function of width equal to half of the field of view of the observations, its inverse Fourier transform is given by a sinc function with non-zero values at integer-indices, and a two-pixel wide main lobe. The convolution with such a sinc function, when performed on $\mathbf{G}^\dagger\mathbf{G}$, results in a 'smearing' of its diagonal character, with more non-zero values now appearing at off-diagonal locations. This smearing effect is compounded since $\mathbf{F}\overline{\mathbf{Z}}^\dagger\overline{\mathbf{F}}^\dagger$ occurs as a pre- and post-operation on $\mathbf{G}^\dagger\mathbf{G}$. However, the smearing does not radically affect the diagonal character since the main lobe of the sinc function has a small width. Numerical results seen in Fig. 5.1 show that $\mathbf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathbf{F}^\dagger$ indeed regains much of the diagonal character from $\mathbf{G}^\dagger\mathbf{G}$, remaining close to a fully diagonal matrix. The ringing effect observed around the diagonal may be attributed to the side lobes of the sinc function.

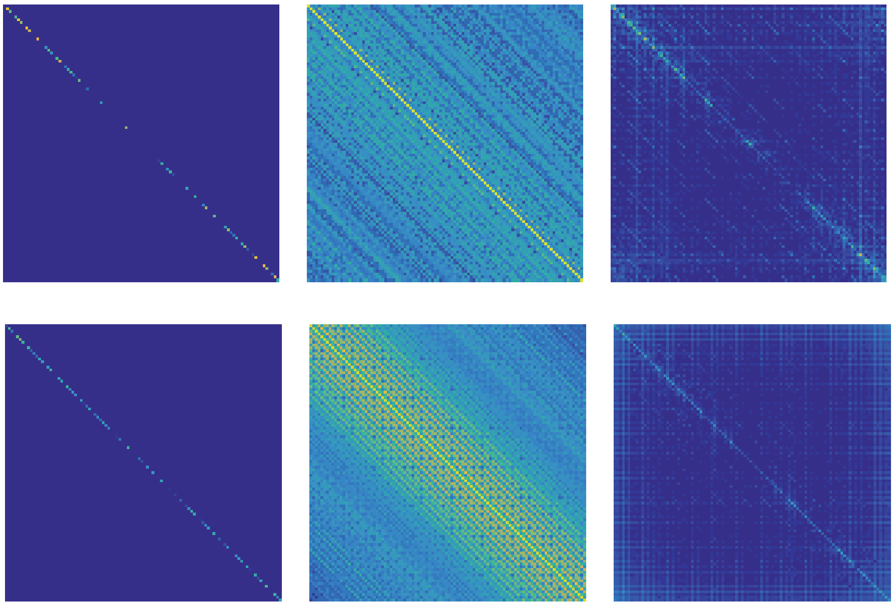


Figure 5.1 – Numerical results (shown here in $\log_{10}$ scale) illustrating correlation of noise in the reduced dimension through varying degrees of 'diagonality' of the initial noise covariance matrix. Left: $\mathbf{G}^\dagger\mathbf{G}$, for reduction through $\mathbf{R}_{\text{grid}}$; Centre: $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$, for reduction through $\mathbf{R}_{\text{dirt}}$; Right: $\mathbf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathbf{F}^\dagger$, for reduction through $\mathbf{R}_{\text{sing}}$. Top row: Gaussian random coverage. Bottom row: SKA-like coverage.

As a quantitative measure of the diagonal character of the matrices shown in Fig. 5.1, we define, for a matrix $\mathbf{C}$, the ratio $\beta_{\mathbf{C}} = \|\,\text{Diag}_{\text{band}}(\mathbf{C})\|_F / \|\mathbf{C}\|_F$, where $\text{Diag}_{\text{band}}(\mathbf{C})$ is a band diagonal matrix formed from a thin band around the main diagonal of $\mathbf{C}$ given by $\text{Diag}(\mathbf{C})$, and $\|\cdot\|_F$ is the Frobenius norm. Numerical results using test uv coverages included in this work show the following typical values: $\beta_{\mathbf{G}^\dagger\mathbf{G}} \approx 0.95$, $\beta_{\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}} \approx 0.50$, and $\beta_{\mathbf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathbf{F}^\dagger} \approx 0.90$, illustrating that the overwhelming majority of significant elements of the matrix $\mathbf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathbf{F}^\dagger$ are on and around the diagonal – in a thin band corresponding to 2% of the matrix size.

We thus see that $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} \approx \mathbf{F}^\dagger[\text{Diag}(\mathbf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathbf{F}^\dagger)]\mathbf{F}$, which is very close to the eigendecomposition of $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$ given by $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\dagger$. This motivates the approximation of

$$\mathbf{V}^\dagger \approx \mathbf{F} \quad \text{and} \quad \boldsymbol{\Sigma}^2 \approx \text{Diag}(\mathbf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathbf{F}^\dagger). \tag{5.8}$$

The approximation of the eigenbasis of $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$ by the orthogonal columns of the discrete Fourier transform operator immediately renders our ideal dimensionality reduction operator feasible. The computationally expensive SVD of $\boldsymbol{\Phi}$ no longer needs to be explicitly calculated, as the discrete Fourier transform operator is known without any knowledge of $\boldsymbol{\Phi}$. Moreover, fast implementations of $\mathbf{F}$ exist in the form of the FFT algorithm. $\mathbf{V}_0$ is then given by $\mathbf{SF}$, $\mathbf{S}$ being a subsampling matrix to select the dimensions corresponding to the $N_0 \le N$ non-zero singular values present in $\boldsymbol{\Sigma}$, thus producing a dimensionality reduction below image size. The diagonal matrix $\boldsymbol{\Sigma}$ is obtained by simply computing the square root of $\text{Diag}(\mathbf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathbf{F}^\dagger)$. A similar selection of N_0 dimensions leads to $\boldsymbol{\Sigma}_0 = \mathsf{S}\boldsymbol{\Sigma}$, thus giving the approximation

$$\mathsf{V}_0{}^\dagger \approx \mathsf{SF} \quad \text{and} \quad \boldsymbol{\Sigma}_0^2 \approx \mathsf{S}\,\mathrm{Diag}(\mathsf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathsf{F}^\dagger). \tag{5.9}$$

This leads us to propose a Fourier model of the dimensionality reduction operator consisting of mapping gridded visibilities back to image space, i.e., computing a dirty image, and then performing a weighted subsampled discrete Fourier transform, given by

$$\mathsf{R}_{\text{sing}} = \boldsymbol{\Sigma}_0^{-1}\mathsf{SF}\boldsymbol{\Phi}^\dagger \quad \in \mathbb{C}^{N_0\times M}. \tag{5.10}$$

The full measurement operator is then given by

$$\boldsymbol{\Phi}'_{\text{sing}} = \boldsymbol{\Sigma}_0^{-1}\mathsf{SF}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} \quad \in \mathbb{C}^{N_0\times N}, \tag{5.11}$$

where $\boldsymbol{\Phi}$ is given by equation (2.10). Finally, the full measurement operator is tangible, and is suitably fast for repeated application in minimization algorithms.

Here we review the properties of our Fourier dimensionality reduction. $\boldsymbol{\Phi}'_{\text{sing}}$ being an approximation of $\boldsymbol{\Phi}'_{\text{sing}-\text{o}}$, we assume that it approximately preserves the null space of $\boldsymbol{\Phi}$. We have also seen that $\mathsf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathsf{F}^\dagger$ is largely diagonal – this diagonal character is maintained as long as the interpolation/de-gridding kernels used to compute the continuous visibilities have compact support. This in turn implies that the covariance matrix $\sigma_n^2\mathsf{R}_{\text{sing}}\mathsf{R}_{\text{sing}}{}^\dagger$ of the embedded noise is largely diagonal as well. The weighting by $\boldsymbol{\Sigma}_0^{-1}$ explicitly normalizes all the diagonal values of the noise covariance matrix to the original noise variance σ_n^2. $\boldsymbol{\Phi}'_{\text{sing}}$ also achieves the same dimensionality reduction to N_0 as $\boldsymbol{\Phi}'_{\text{sing}-\text{o}}$. Crucially, it exhibits a fast implementation since its constituent operators are diagonal, sparse and Fourier matrices. A summary of properties of $\boldsymbol{\Phi}'_{\text{sing}-\text{o}}$ and $\boldsymbol{\Phi}'_{\text{sing}}$ is shown in Table 5.1.

We extend the idea of approximating the initial noise covariance matrix to a further degree by assuming $\mathsf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathsf{F}^\dagger \propto \mathsf{I}$ without explicit computation of $\mathsf{F}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\mathsf{F}^\dagger$, thus leading to $\mathsf{R} = \mathsf{F}\boldsymbol{\Phi}^\dagger$. This variant of our approach is also investigated in our simulations.

5.1.2 Gridding-based dimensionality reductions

Embedding visibilities to the dirty image

Embedding visibilities to the dirty image is a standard way to reduce dimensionality, and is performed in many image reconstruction methods in radio interferometry, essentially through an image-based deconvolution with appropriate weighting. Setting it in the terminology presented here, it amounts to using an embedding operator $\mathsf{R} = \boldsymbol{\Phi}^\dagger$. The corresponding noise is highly correlated, and this can be seen from the covariance matrix $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$ shown in Fig. 5.1, which contains significant off-diagonal elements.

In order to be able to use this dimensionality reduction in our minimization problem formulation (equation (3.1.5)), the embedded noise would need to have i.i.d. Gaussian entries. This is achieved by assuming, as done previously, an approximation of the noise covariance matrix by its diagonal

$$\mathbf{W}^2 = \mathrm{Diag}(\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}). \tag{5.12}$$

$\mathbf{W}$ is invertible since $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$ (applied to a point source image) is the dirty beam and hence typically non-zero along its main diagonal – this also implies that there is no potential for further subsampling based on zero values. We subsequently apply a weighting $\mathbf{W}^{-1}$ to obtain the dimension embedding operator

$$\mathbf{R}_{\mathrm{dirt}} = \mathbf{W}^{-1}\boldsymbol{\Phi}^\dagger \quad \in \mathbb{C}^{N\times M}. \tag{5.13}$$

The full measurement operator is therefore given as

$$\boldsymbol{\Phi}'_{\mathrm{dirt}} = \mathbf{W}^{-1}\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} \quad \in \mathbb{C}^{N\times N}. \tag{5.14}$$

It preserves the null space of $\boldsymbol{\Phi}$, thus retaining any original NSP. Indeed the SVD or eigendecomposition of $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$ reads as $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\dagger$. Applying $\mathbf{R}_{\mathrm{dirt}}$ is also fast as the individual suboperators in $\boldsymbol{\Phi}$ have fast implementations. However, the embedded noise covariance matrix $\sigma_n^2\mathbf{R}_{\mathrm{dirt}}\mathbf{R}_{\mathrm{dirt}}{}^\dagger$ is far from diagonal, as seen in Fig. 5.1, though with diagonal entries all equal to the original noise variance σ_n^2. For completeness in our comparison of results, we extend our approximation of the initial noise covariance matrix by nevertheless assuming $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi} \propto \mathbf{I}$, resulting in the variant of $\mathbf{R}_{\mathrm{dirt}}$ given by $\mathbf{R} = \boldsymbol{\Phi}^\dagger$.

Gridding visibilities

Our proposed gridding-based dimensionality reduction relies on the routinely performed step of gridding continuous visibilities to discrete Fourier points to reduce the data dimension to the size of the oversampled discrete Fourier grid. Gridding can be seen as applying the embedding $\mathbf{R} = \mathbf{G}^\dagger$ to continuous visibilities. As noted earlier in Section 5.1.1, in the general case with DDEs and non-coplanar antenna arrays, the corresponding noise covariance matrix $\sigma_n^2\mathbf{G}^\dagger\mathbf{G}$ is non-diagonal, but under our initial assumptions of calibration error-free data and $w = 0$, we note that the noise covariance matrix is largely diagonal. This is also seen in Fig. 5.1 in the form of a highly diagonal structure of $\mathbf{G}^\dagger\mathbf{G}$ computed on simulated data.

The first step of our proposed modification appears as follows: an approximate i.i.d. Gaussian nature of the embedded noise is ensured by weighting the embedding operator with $\overline{\mathbf{W}}^{-1}$, where

$$\overline{\mathbf{W}}^2 = \mathrm{Diag}(\mathbf{G}^\dagger\mathbf{G}). \tag{5.15}$$

More precisely, this weighting $\overline{\mathbf{W}}^{-1}$ simply normalizes all the diagonal values of the noise covariance matrix to the original noise variance σ_n^2.

Here we note a natural further dimensionality reduction by discarding those discrete Fourier

Table 5.1 – Summary of the different dimensionality reduction methods with their advantages and disadvantages with respect to compressed sensing-based imaging.

Properties of $\boldsymbol{\Phi}' = \mathbf{R}\boldsymbol{\Phi}$ (Full meas. operator)	$\mathbf{R} = \mathbf{R}_{\text{sing}-\text{o}}$	$\mathbf{R} = \mathbf{R}_{\text{sing}}$	$\mathbf{R} = \mathbf{R}_{\text{dirt}}$	$\mathbf{R} = \mathbf{R}_{\text{grid}}$
Approximate null space preservation	Yes	Yes	Yes	Yes
Fast implementation	No	Yes	Yes	Yes
Largely diagonal noise covariance matrix	Yes	Yes	No	Yes
Final dimension	$N_0 \leq N$	$N_0 \leq N$	N	$\overline{N} \leq 4N$

grid points that are not covered by any interpolation kernel support over the uv plane. As contributions of the different interpolation kernels over a given discrete Fourier grid point correspond to individual columns of the matrix $\mathbf{G}$, discrete grid points that are not thus covered manifest as all-zero columns of $\mathbf{G}$ and consequently, zeros on the diagonal of $\mathbf{G}^\dagger\mathbf{G}$. A subsampling operator $\overline{\mathbf{S}}$ can then be applied to the embedding operator to only select dimensions corresponding to non-zero diagonal values of $\mathbf{G}^\dagger\mathbf{G}$. A similar selection of dimensions on $\overline{\mathbf{W}}$ gives $\overline{\mathbf{W}}_0$ of size $\overline{N}$ below the dimension of the oversampled discrete Fourier grid, thus leading to the dimensionality reduction operator

$$\mathbf{R}_{\text{grid}} = \overline{\mathbf{W}}_0^{-1}\overline{\mathbf{S}}\mathbf{G}^\dagger \quad \in \mathbb{C}^{\overline{N}\times M}. \tag{5.16}$$

We see that the full measurement operator

$$\boldsymbol{\Phi}'_{\text{grid}} = \overline{\mathbf{W}}_0^{-1}\overline{\mathbf{S}}\mathbf{G}^\dagger\boldsymbol{\Phi} \quad \in \mathbb{C}^{\overline{N}\times N} \tag{5.17}$$

preserves the null space of $\boldsymbol{\Phi}$, following from $\text{Null}(\boldsymbol{\Phi}) \subseteq \text{Null}(\mathbf{G}^\dagger\boldsymbol{\Phi}) \subseteq \text{Null}(\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi})$, thus retaining any original NSP. The diagonal dominated nature of $\sigma_n^2\mathbf{R}_{\text{grid}}\mathbf{R}_{\text{grid}}^\dagger$ has already been shown, which leads to an appropriate modelling of the noise. Also, applying $\mathbf{R}_{\text{grid}}$ is fast owing to the sparsity of $\mathbf{G}^\dagger$.

Note that in the context of this dimensionality reduction with $\mathbf{R}_{\text{grid}}$, the weighting matrix $\overline{\mathbf{W}}_0^{-1}$, in fact designed for optimal weighting of the embedded visibilities, also operates as uniform weighting. Indeed, the diagonal values of $\mathbf{G}^\dagger\mathbf{G}$ are a measure of the density of continuous visibilities at each discrete grid point.

As with $\mathbf{R}_{\text{sing}}$, we also introduce and test a variant of our proposed reduction method $\mathbf{R}_{\text{grid}}$. We approximate the initial noise covariance matrix in this case by assuming $\mathbf{G}^\dagger\mathbf{G} \propto \mathbf{I}$, leading to the dimensionality reduction $\mathbf{R} = \mathbf{G}^\dagger$. This variant is included in comparisons of reconstruction quality.

5.1.3 Feature comparison

A comparison of the different proposed dimensionality reduction methods R_{sing}, R_{dirt} and R_{grid} is shown in Table 5.1. We note in the listing that R_{sing} provides a good combination of the desired final dimension, the guarantees for compressed sensing-based imaging to reconstruct images, and a largely diagonal noise covariance matrix that enables us to embed this technique in the convex optimization algorithm we employ for imaging. R_{dirt} embeds the data to image size while maintaining any original NSP of $\boldsymbol{\Phi}$, and has a fast implementation; however, it fails to appropriately model the noise and is therefore less suitable for the minimization problem which requires i.i.d. Gaussian embedded noise to enable a simple ℓ_2-norm data fidelity term. R_{grid} continues to maintain the NSP of $\boldsymbol{\Phi}$ and the i.i.d. Gaussian properties of the noise, which are essential for image reconstruction using our convex optimization algorithms. However, the data reduction is limited to a size $\overline{N} \leq o^2 N$ for an oversampling factor of o in each dimension.

As mentioned in Sections 5.1.1 and 5.1.2, for each of R_{sing}, R_{dirt} and R_{grid}, an attempt is also made to further approximate the initial noise covariance matrix by the identity matrix in order to render the application of the respective dimensionality reduction methods even faster. equations (5.10), (5.13) and (5.16) are then simplified to $\mathsf{R} = \mathsf{F}\boldsymbol{\Phi}^\dagger$, $\mathsf{R} = \boldsymbol{\Phi}^\dagger$ and $\mathsf{R} = \mathsf{G}^\dagger$ respectively. However, this approximation is seen to be inappropriate, leading to poorer modelling of the noise and consequently lowering image reconstruction quality. R_{sing}, R_{dirt}, R_{grid} and their respective variants are used to reduce dimensionality before performing image reconstruction in different settings and with varying data sizes.

Another feature for comparing the different reduction methods was introduced in section 4.4.2 in the context of random projections like the Gaussian random matrix. It is related to the fast implementation of the dimensionality reduction, which is already included in Table 5.1, and focuses on the speed of any preprocessing steps needed to arrive at $\boldsymbol{\Phi}'$. In the case of our proposed reduction methods, the sub-operators of $\boldsymbol{\Phi}'$ are all seen to be fast, and any computations in the higher dimension M are avoided. There is, however, a one-time cost in the form of a pre-computation in M dimensions to obtain the holographic matrix H. We propose to also eliminate this one-time cost in the preprocessing stage by leveraging the block-separable structure of the measurement operator $\boldsymbol{\Phi}'$. The full M-dimensional data vector may be divided in smaller-sized blocks, and each block could then undergo dimensionality reduction separately, with the associated pre-computations also performed in the lower dimension. As this proposed modification is more readily apparent in the case of real data acquisition where incoming data can be split in blocks as and when they are received, we describe our method in greater detail in section 6.1, along with its mathematical underpinnings.

5.1.4 Further reduction by thresholding

As a conservative dimensionality reduction method, R_{sing} would embed to a final data size $N_0 \approx N$ under the assumption that there are very few zero-valued singular values of $\boldsymbol{\Phi}$, within limits of numerical precision, and all corresponding singular vectors are thus necessary to retain the information content of $\mathbf{\Phi}$. Similarly, R_{grid} embeds to a size $\overline{N} \approx o^2 N$ under the assumption of having contributions to the continuous visibilities from most discrete Fourier grid points. However, further dimensionality reduction may be obtained in both cases by a thresholding strategy.

We first consider the Fourier dimensionality reduction model based on R_{sing}. The approach described below to further reduce the final embedding dimension consists of discarding the data dimensions associated with singular values $\mathbf{\Sigma}_i$ below a threshold, rather than conservatively discarding those equal to zero only – this is made possible due to the fact that the singular values occupy a range of values going down to zero, as discussed in Section 5.1.1. In other words, through such a thresholding operation, we will attempt a low-rank approximation of the original singular value matrix $\mathbf{\Sigma}$, and consequently reduce our final data dimension to the corresponding low-rank. For the sake of the simplicity of this argument, we consider R_{sing} to be equal to $\mathsf{R}_{\text{sing}-\text{o}}$. From equation (5.4), we see that the full measurement operator $\mathbf{\Phi}'_{\text{sing}-\text{o}}$ reads as a weighted subsampling in the orthonormal basis $\mathsf{V}^\dagger$, with the weights given by the singular values of $\mathbf{\Phi}$. From equation (5.5), the noise covariance matrix reads as $\sigma_n^2\mathsf{I}$. In order to safely discard a given singular value $\mathbf{\Sigma}_i$ for the dimension i without losing information, its effect on the corresponding embedded visibility $\boldsymbol{y}'_i$ would need to be negligible relative to the embedded noise level σ_n:

$$|\boldsymbol{y}'_i| < \gamma\sigma_n, \quad \text{with } \gamma = \mathcal{O}(1). \tag{5.18}$$

In general, one has $|\boldsymbol{y}'_i| \leq \Sigma_i \|\boldsymbol{x}\|_2$, which is saturated only in the case where $\boldsymbol{x}$ is fully aligned with the right singular vector V_i. This means that the condition

$$\mathbf{\Sigma}_i \|\boldsymbol{x}\|_2 < \gamma\sigma_n \tag{5.19}$$

is sufficient to ensure the requirement imposed by equation (5.18) to disregard dimension i. In other words, the data dimension i can be discarded with no adverse effect on signal reconstruction if the corresponding singular value, computed as given in equation (5.9), is below a noise-based threshold:

$$\mathbf{\Sigma}_i < \frac{\gamma\sigma_n}{\|\boldsymbol{x}\|_2}. \tag{5.20}$$

Secondly, we consider the gridding-based dimensionality reduction R_{grid}. The approach to further reducing the final embedding dimension will again consist of discarding the data dimensions associated with the weights $\overline{\mathsf{W}}_i$ below a threshold, rather than conservatively discarding those equal to zero only. A bound similar to equation (5.20) can be deduced as follows for thresholding out data dimensions. Again, for the sake of this very argument only, the full measurement operator $\mathbf{\Phi}'_{\text{grid}}$ in equation (5.17) can be approximated as weighted subsampling in the Fourier basis, with weights $\overline{\mathsf{W}}$ computed from equation (5.15). The noise covariance matrix exhibits diagonal values all equal to the original noise variance σ_n^2. The same reasoning as for R_{sing} now applies and the data dimension i can be discarded with no adverse effect on signal reconstruction if the corresponding weight is below the following noise-based threshold:

$$\overline{\mathbf{W}}_i < \frac{\gamma \sigma_n}{\|\boldsymbol{x}\|_2}. \tag{5.21}$$

The threshold computation in equations (5.20) and (5.21) needs knowledge of $\|\boldsymbol{x}\|_2$, which is a priori not available from interferometric data. One would naturally want to estimate $\|\boldsymbol{x}\|_2$ from the dirty image. This is supported by recent work showing that $\boldsymbol{x}$ can be bounded by the dirty image in the ℓ_2 sense [Wijnholds and van der Veen, 2011, Sardarabadi et al., 2016].

5.1.5 Reduced computational requirements

Current radio-interferometric imaging techniques involve processing in the data space (of dimension M) and a lower dimensional sparsity space (of dimension N or $\overline{N}$). For CLEAN-based algorithms this can be seen in the move between 'major' cycles in the data dimension M, 'minor' cycles with gridded visibilities of dimension $\overline{N}$, and the image space of dimension N with an implicit sparsity assumption. For convex optimization-based algorithms like the one used in this work, this is typically seen in the concurrent computation of a data fidelity term with vectors of dimension M, and a sparsity prior of lower dimension N.

The goal of dimensionality reduction as described here is to reduce the computational load of imaging methods for next-generation radio interferometers where M is very large, on the order of 10^{10}. The proposed dimensionality reduction method using $\mathbf{R}_{\text{sing}}$ reduces data size by significant amounts to $N_0 \leq N \ll M$, and these lower-dimensional data cause a smaller memory footprint in imaging pipelines. The existence of fast sub-operator implementations and a low-sized full measurement operator translate to faster computations per iteration of the convex optimization algorithms. The properties of the full measurement operator $\boldsymbol{\Phi}'$ as listed in Table. 5.1 are good indicators of the computational savings afforded by $\mathbf{R}_{\text{sing}}$ as compared to other reduction methods. The applicability of the proposed dimensionality reduction method using $\mathbf{R}_{\text{sing}}$ has the advantage of resulting in a reduced data dimension that is independent of the initial data size (this is also true for $\mathbf{R}_{\text{dirt}}$ and $\mathbf{R}_{\text{grid}}$, which results in reduced data sizes of N and $\overline{N}$ respectively).

The initial set-up of the imaging method shall indeed be affected by an increase in the initial dimension – in particular, the pre-computation of the holographic matrix and the appropriate weights to be used in the imaging algorithm. However, these pre-computations have a one-time cost, and subsequent imaging is unaffected, depending only on the embedded data and thus using fewer resources in terms of memory and computing time.

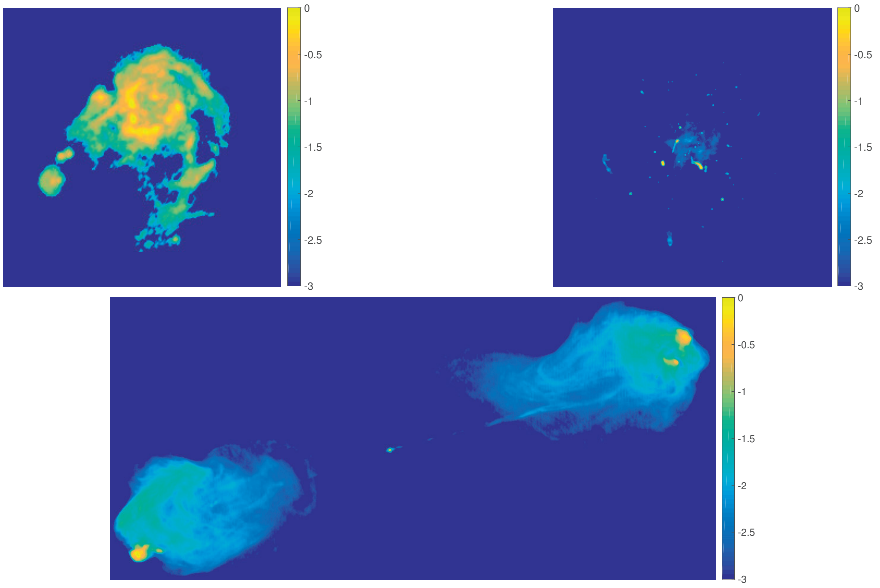


Figure 5.2 – The test images in $\log_{10}$ scale, clockwise from top left: M31 (256 × 256 pixels), a simulated galaxy cluster (512 × 512 pixels) and Cygnus A (477 × 1025 pixels).

5.2 Simulations and results

5.2.1 Simulation settings

The effectiveness of the proposed dimensionality reduction method was demonstrated through simulations. Image quality comparisons were made between reconstructions through the dimensionality reduction methods $\mathbf{R}_{\text{sing}}$, $\mathbf{R}_{\text{dirt}}$ and $\mathbf{R}_{\text{grid}}$, and their respective variants $\mathbf{R} = \mathbf{F}\boldsymbol{\Phi}^{\dagger}$, $\mathbf{R} = \boldsymbol{\Phi}^{\dagger}$ and $\mathbf{R} = \mathbf{G}^{\dagger}$. In a first setting, a conservative dimensionality reduction was performed for each case – $\mathbf{R}_{\text{sing}}$ accounting for dimensions corresponding to all non-zero singular values of $\boldsymbol{\Phi}$, and $\mathbf{R}_{\text{grid}}$ accounting for dimensions corresponding to all discrete Fourier grid points that have non-zero contribution to the continuous visibilities through interpolation kernels. The final data dimension after reduction in this setting was seen to be $N_0 \approx N$ for $\mathbf{R}_{\text{sing}}$, $N_0 = N$ for $\mathbf{R}_{\text{dirt}}$, and $\overline{N} \approx 4N$ for $\mathbf{R}_{\text{grid}}$.

Simulations were performed on different test images chosen for their varied characteristics: (i) the classic 'M31' image has a compact structure showing an HII region of the M31 galaxy (256 × 256 pixels); (ii) an image of a galaxy cluster (512 × 512 pixels) simulated using the 'FARADAY' tool (courtesy M. Murgia and F. Govoni [Murgia et al., 2004]), has high dynamic range by design; (iii) a partial image of the Cygnus A radio galaxy (477 × 1025 pixels) includes a strong central core, two strong jets and lobes of diffuse structure with bright hotspots. These test images are shown in Fig. 5.2.

Two categories of uv coverages were used to simulate telescope measurements. One with synthetic coverages with a random Gaussian sampling profile with missing frequency regions, and another with more realistic SKA-like coverages generated with a simulated telescope

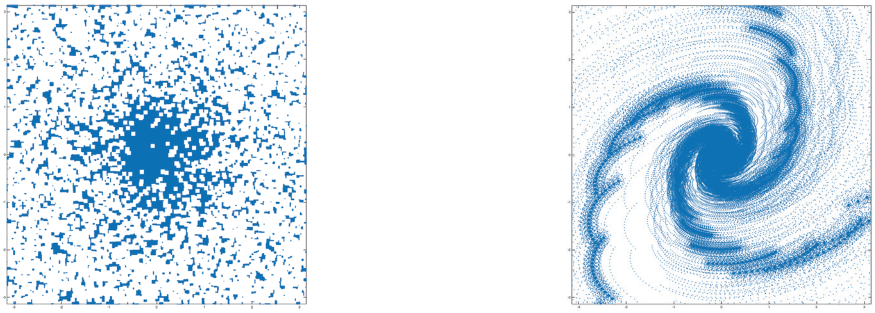


Figure 5.3 – Illustration of simulated uv coverages over normalized frequency axes $[-\pi, \pi]$. Left: Random Gaussian density profile with the uv plane increasingly sparsely covered at higher frequencies (approximately 3,000,000 uv points); Right: Profile generated with SKA-like baselines, with partial ellipses simulating dense coverage at lower frequencies and sparser coverage at higher frequencies. A telescope configuration of 254 antennas was used to obtain approximately 650,000 uv points.

configuration of 254 antennas (Fig. 5.3), using the 'MEQTREES' tool[1] [Noordam and Smirnov, 2010] that uses the 'CASACORE' software suite.[2] The SKA-like coverages correspond to observation times ranging from 30 minutes to 8 hours, depending on the image and the initial data dimension. The frequencies were normalized to lie in the interval $[-\pi, \pi]$. The $(0,0)$ component of the uv plane was not included in generating visibilities, as explained in section 4.4.2.

For each coverage, complex visibilities were generated over the continuous uv plane using the measurement model described in equation (2.8), and these were perturbed with complex additive white Gaussian noise, resulting in an input signal-to-noise ratio (SNR) of 30 dB. The input SNR was defined as $\text{SNR}_\text{i} = 20\log_{10}(\|\boldsymbol{\Phi x}\|_2/\|\boldsymbol{n}\|_2)$ as given in section 4.4.2. The continuous visibilities dimension was varied over a wide range, in multiples of image size, from 10 to 100. This corresponds to an approximate range of 650,000 to 26 million visibilities over different test images. Continuous visibilities were generated with a non-uniform oversampled Fourier transform (2× oversampling in each image dimension) using 8×8 Kaiser-Bessel interpolation kernels as described and implemented by Fessler and Sutton [2003]. ℓ_1-minimization was performed using the SARA algorithm implemented in PURIFY. SARA regularizes the inverse problem by imposing 'average sparsity' of the signal over a set of bases, as explained in section 3.3.1. The resulting minimization problem was solved using an ADMM-based proximal splitting method – Carrillo et al. [2012, 2014] provide further details on SARA; Carrillo et al. [2015], Onose et al. [2016] present implementation details of the ADMM-based minimizer.

Additionally, a more aggressive dimensionality reduction was performed for $\mathbf{R}_\text{sing}$ and $\mathbf{R}_\text{grid}$ by only retaining dimensions corresponding to significant values of embedded data, as described in Section 5.1.4. This led to a final data dimension size of $N_0 \ll N$ and $\overline{N} \ll 4N$ respectively.

---

[1] MEQTREES is available at http://meqtrees.net/.

[2] CASACORE is available at https://github.com/casacore.

Image reconstruction was performed for SKA-like coverages using these further reduced data.

As a comparison baseline, images were also recovered using the complete visibilities set. This corresponds to the trivial 'embedding' $\mathbf{R} = \mathbf{I}$. Finally, the reconstruction was compared with model images obtained using the Multiscale CLEAN (MS-CLEAN) algorithm [Cornwell, 2008] as provided in the 'WSCLEAN' software program[3] [Offringa et al., 2014]. MS-CLEAN was run on continuous visibilities simulated from SKA-like uv coverages. The synthetic Gaussian-profile uv coverages were generated on the fly during simulations, and since WSCLEAN takes measurement sets as input, MS-CLEAN was not run on data simulated using these coverages.

We used SNR and dynamic range (DR) of the reconstructed image as measures of image quality, and compared these values across different methods in our simulations. 'Reconstruction SNR' is defined in equation (4.8). We define 'Reconstruction DR' as

$$\mathrm{DR}_{\hat{\boldsymbol{x}}} = \sqrt{N}\|\boldsymbol{\Phi}\|^2 \frac{max(\hat{\boldsymbol{x}})}{\|\boldsymbol{\Phi}^\dagger(\boldsymbol{y} - \boldsymbol{\Phi}\hat{\boldsymbol{x}})\|_2}, \tag{5.22}$$

where $\boldsymbol{x}$ is the underlying test image measured through the operator $\boldsymbol{\Phi}$ (with spectral norm $\|\boldsymbol{\Phi}\|$) to give visibilities $\boldsymbol{y}$, and $\hat{\boldsymbol{x}}$ is the reconstructed image. For the methods $\mathbf{R}_{\mathrm{sing}}$, $\mathbf{R}_{\mathrm{dirt}}$, $\mathbf{R}_{\mathrm{grid}}$ and their respective variants, the DR was computed using the corresponding measurement operators as given in equations (5.11), (5.14) and (5.17), along with the respective embedded visibilities. For MS-CLEAN, the DR is computed by using the peak of the model image instead of the restored image.

Since reconstruction results from MS-CLEAN are either in the form of a restored image containing added residual, or a model image containing extended components, direct SNR and DR computations are not readily apparent and no longer remain a valid way to compare reconstruction performance between CLEAN and the proposed compressed sensing-based imaging methods. Therefore, the MS-CLEAN reconstructions are presented here as model images (without being convolved with the beam and adding the residual) for visual comparison with the other methods described in this work. Additionally, we note that MS-CLEAN model images contain prominent negative-valued components – this does not have a physical representation for an intensity distribution, but is crucial for MS-CLEAN to produce 'restored' images after convolution of the model image with the CLEAN beam.

5.2.2 Image reconstruction results

Accounting for all non-zero singular values

Image reconstruction performance of the different methods over varying simulation settings is discussed here for each test image. Graphs showing SNR and DR comparisons over the two types of coverage are shown in Figs. 5.4 and 5.5 respectively. For the SKA-like coverages, a visual comparison is also made between the methods $\mathbf{R} = \mathbf{I}$, $\mathbf{R}_{\mathrm{sing}}$, $\mathbf{R}_{\mathrm{dirt}}$, $\mathbf{R}_{\mathrm{grid}}$, and MS-CLEAN

---

[3] WSCLEAN is available at https://sourceforge.net/projects/wsclean.

(run on the full data set) by showing the reconstructed, error and residual images in $\log_{10}$ scale. These visual comparisons for the three test images are shown in Figs. 5.6, 5.7 and 5.8.

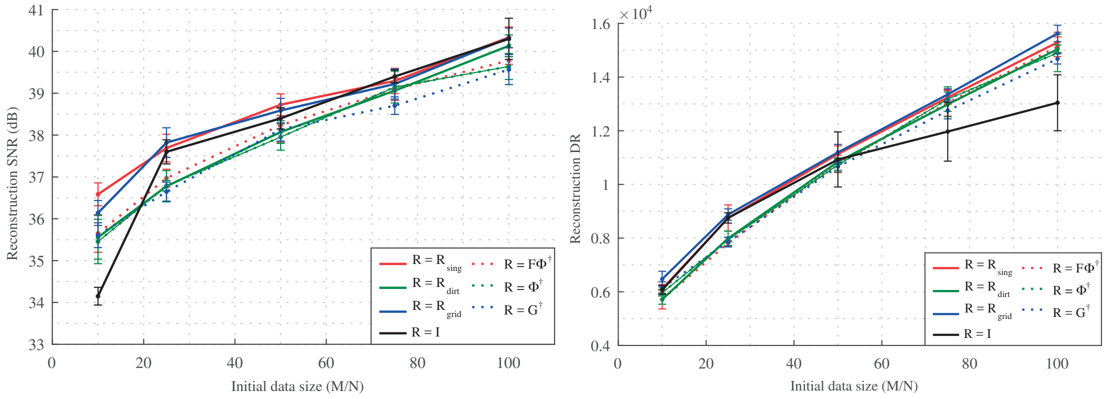
The reconstructed images shown for MS-CLEAN are obtained by cropping from a 3-4 times larger model image output by WSCLEAN. For all test images, MS-CLEAN was run with a uniform weighting scheme, set to iterate down to an automatically calculated threshold of two standard deviations of the noise, and with a major loop gain of 0.8. The model image in each case was renormalized to have a maximum pixel value matching that of the MS-CLEAN model. It may be noted that this does not change the overall visual appearance of the reconstruction and error images as shown in Figs. 5.6, 5.7 and 5.8. The $\log_{10}$ scale was used to highlight the smallest variations and structures, which inadvertently emphasizes the low-valued artefacts in MS-CLEAN output models; a linear scale would render these artefacts visually indistinguishable from the background.

Uniformly weighted dirty residual images were generated using WSCLEAN for each output model from the different image recovery methods. Absolute Jy/beam residual values were plotted on a $\log_{10}$ scale, highlighting small variations in structure. The comparatively significant structure seen in the residual images for $\mathbf{R} = \mathbf{I}$, $\mathbf{R}_{\text{sing}}$, $\mathbf{R}_{\text{dirt}}$ and $\mathbf{R}_{\text{grid}}$ may be due to the absence of negative-valued model components, which are present in the model generated by MS-CLEAN and are compensated for during the computation of the dirty image by WSCLEAN.

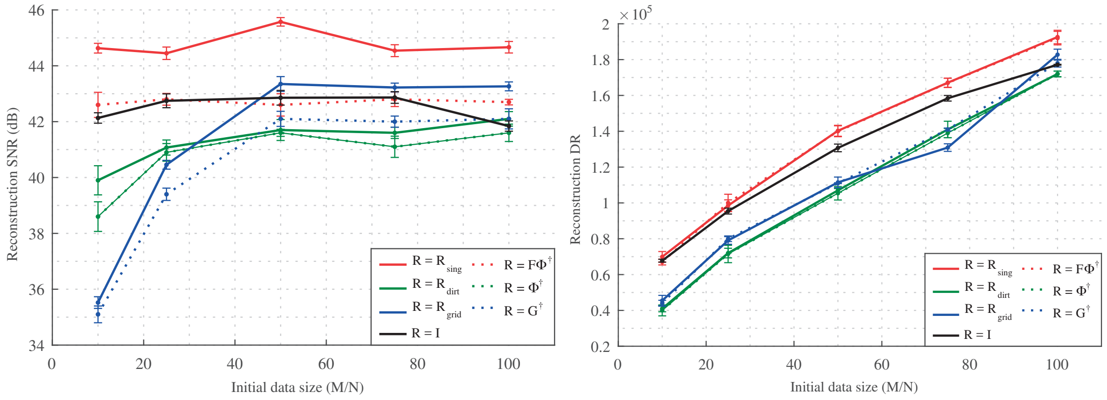
The M31 test image is reconstructed accurately for all the imaging methods. Reconstruction with the complete visibilities set reaches 40 dB for Gaussian random coverages and 28 dB for SKA-like coverages at data sizes of $100N$. Fig. 5.4a shows that $\mathbf{R}_{\text{sing}}$ and $\mathbf{R}_{\text{grid}}$ perform equally well over Gaussian random coverages, reaching output SNRs of around 40 dB. Over SKA-like coverages, $\mathbf{R}_{\text{grid}}$ results in output SNRs noticeably higher than $\mathbf{R}_{\text{sing}}$ for most data sizes, e.g., around 25 dB for data sizes of $75 - 100N$ as shown in Fig. 5.5a. When comparing DR over Gaussian random coverages, images reconstructed after applying $\mathbf{R}_{\text{sing}}$ reach 1.6×10^4, $\mathbf{R}_{\text{grid}}$ reaches 1.5×10^4 and the complete visibilities set leads to a DR of 1.3×10^4. The corresponding values over SKA-like coverages are 7.2×10^5, 6.5×10^5 and 6.3×10^5 respectively. Fig. 5.6 shows a visual comparison of images reconstructed over SKA-like coverages that confirms this trend, where $\mathbf{R}_{\text{grid}}$ results in the lowest error among all methods.

The computation time of the ADMM-based algorithm used for image reconstruction shows a clear advantage of using $\mathbf{R}_{\text{sing}}$, which takes ≈1.5 seconds per iteration as opposed to ≈18 seconds per iteration without dimensionality reduction, when using all $M = 100N$ visibilities. MS-CLEAN output model images of size 1024×1024 pixels were cropped to 256×256 pixels. Image deconvolution took 4 major iterations, and the output model shown In Fig. 5.6 contains 8458 components.

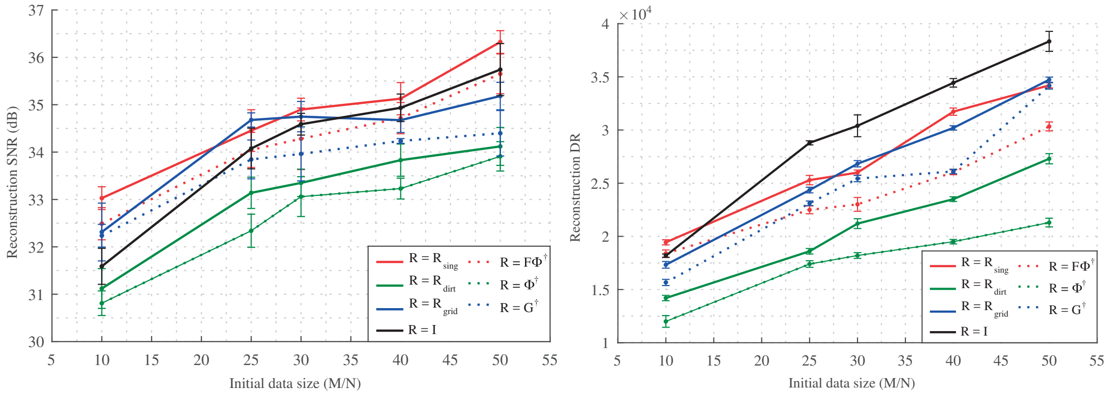
The galaxy cluster test image was chosen for its high dynamic range, and simulations show that $\mathbf{R}_{\text{sing}}$ results in much better reconstruction than all other methods, both in terms of DR and SNR. As seen in Fig. 5.4b, the SNR from $\mathbf{R}_{\text{sing}}$ on Gaussian random coverages is more than 2 dB higher than the complete visibilities set, on average, reaching consistently up to 45 dB



(a) M31

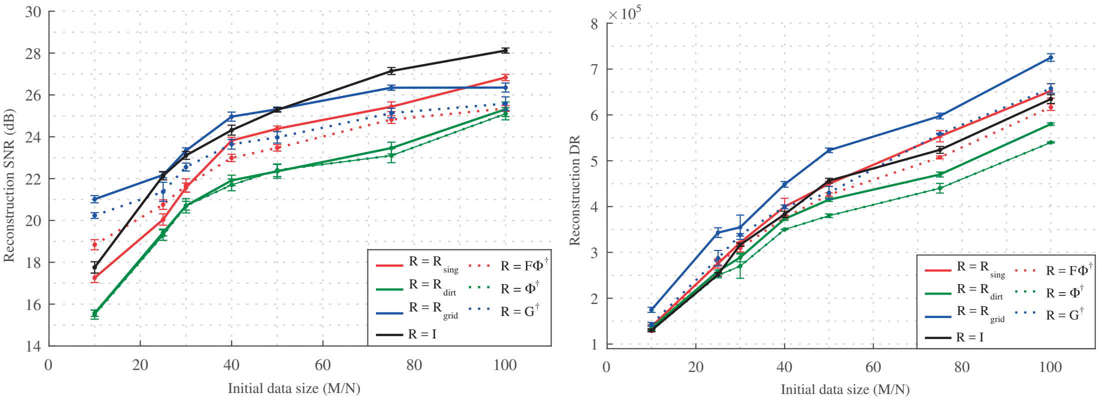


(b) Galaxy cluster

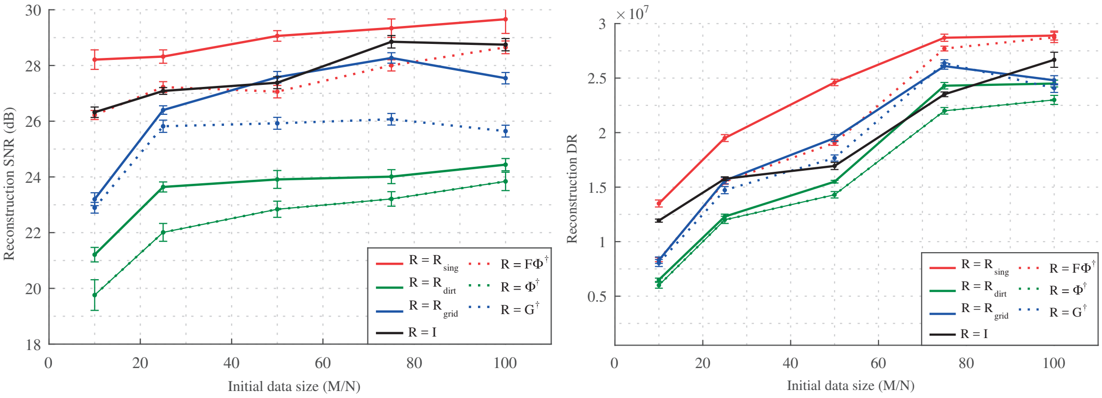


(c) Cygnus A

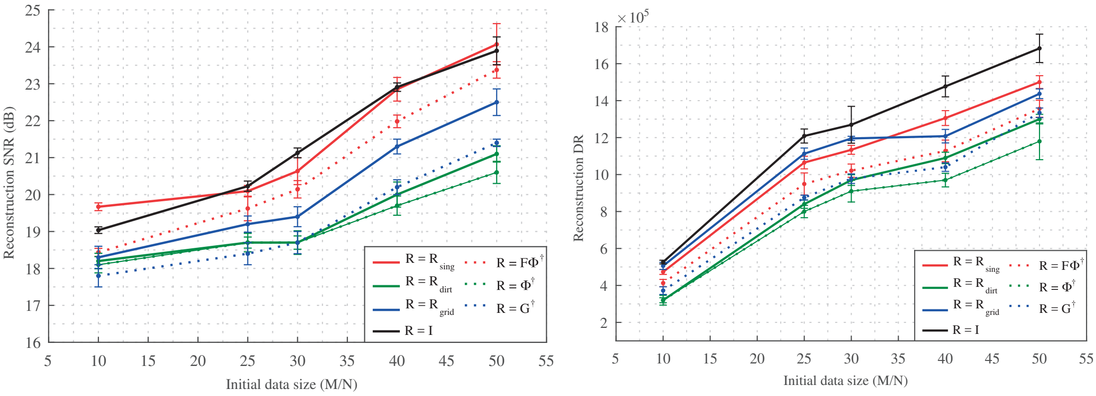
Figure 5.4 – Image reconstruction from visibilities using Gaussian random coverage: comparison of different dimensionality reduction methods. Left: SNR; Right: DR, for the recovered image over a range of initial continuous visibilities. Top row: M31 image, initial data size varies from 650,000 to 6.5 million visibilities. Middle row: Galaxy cluster image, initial data size varies from 2.6 million to 26 million visibilities. Bottom row: Cygnus A image, initial data size varies from 4.8 million to 24 million visibilities. Error bar lengths correspond to one standard deviation around the mean over ~ 15 simulations.



(a) M31



(b) Galaxy cluster



(c) Cygnus A

Figure 5.5 – Image reconstruction from visibilities using SKA-like coverages: comparison of different dimensionality reduction methods. Left: SNR; Right: DR, for the recovered image over a range of initial continuous visibilities. (a): M31 image, initial data size varies from 650,000 to 6.5 million visibilities. (b): Galaxy cluster image, initial data size varies from 2.6 million to 26 million visibilities. (c): Cygnus A image, initial data size varies from 4.8 million to 24 million visibilities. Error bar lengths correspond to one standard deviation around the mean over ~ 15 simulations.

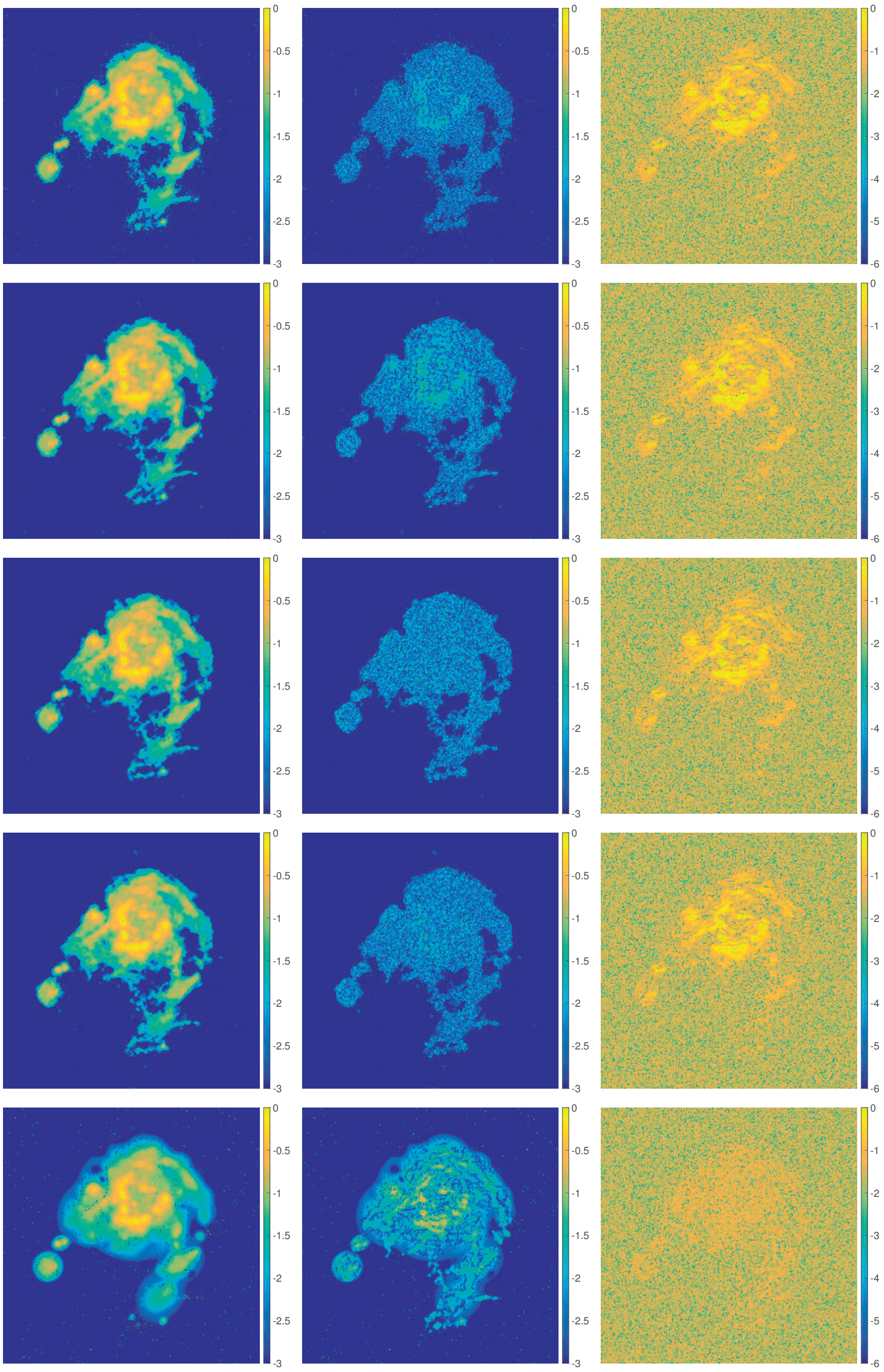


Figure 5.6 – Visual comparison of reconstruction results for the M31 test image with $M = 100N$. From left to right: reconstructed, error and residual images in $\log_{10}$ scale. First four rows from top to bottom: reconstruction performed with all visibilities, 'reduced' visibilities after performing dimensionality reduction with $\mathbf{R}_{\text{sing}}$, $\mathbf{R}_{\text{dirt}}$, and $\mathbf{R}_{\text{grid}}$ respectively. Last row: reconstruction using MS-CLEAN with a uniform weighting scheme.

and outperforming all other dimensionality reduction methods. Over SKA-like coverages, the image reconstruction trend as shown in Fig. 5.5b is similar, with $\mathbf{R}_{\text{sing}}$ providing output SNRs reaching almost 30 dB while $\mathbf{R}_{\text{grid}}$ usually has SNRs 2 dB below this value. The DR curves confirm this trend, and high DR values of 2.8×10^7 are reached with $\mathbf{R}_{\text{sing}}$ over SKA-like coverages. In comparison, $\mathbf{R}_{\text{grid}}$ and the complete visibilities set reach 2.5×10^7 and 2.6×10^7 respectively. The DR values are comparatively low for Gaussian random coverages, but the relative improvement is maintained. The respective values are 1.9×10^5, 1.8×10^5 and 1.7×10^5. A visual comparison in Fig. 5.7 highlights the sharper structure of the bright sources and detail in the reconstructed image corresponding to $\mathbf{R}_{\text{sing}}$ when compared to the other methods. Fig. 5.7 also shows very low values over the residual images in general across the methods, but an extremely sensitive scale is set to forcibly show the structure present in the residual.

The actual residual image values are close to the numerical precision for these computations. Reconstruction using $M = 50N$ visibilities took ≈12 seconds per iteration without dimensionality reduction, whereas it only took ≈1.8 seconds per iteration using $\mathbf{R}_{\text{sing}}$. Image deconvolution with MS-CLEAN took 6 major iterations, and the model shown in Fig. 5.7 contains 30649 components. Output model images of size 2048 × 2048 pixels were cropped to 512 × 512 pixels. The reconstruction is closer to the other methods, and the residual image contains only minor structure. Some high-flux regions are visible in the error image that were not appropriately modelled in the reconstruction.

The Cygnus A image was chosen for the varied structure present in the different parts of the image. Fig. 5.4c shows output SNR for reconstruction with $\mathbf{R}_{\text{sing}}$ marginally higher than that with the complete visibilities set over Gaussian random coverages, at 36.3 dB and 35.7 dB respectively. $\mathbf{R}_{\text{grid}}$ leads to 35.2 dB output SNR. Over SKA-like coverages, $\mathbf{R}_{\text{sing}}$ performs better than the complete visibilities set over low data sizes, but the difference is made up for larger data sizes and the output SNRs reach comparable values, at 24 dB and 23.8 dB respectively, as seen in Fig. 5.5c. DR values are clearly highest when using the complete visibilities set. Over Gaussian random coverages, it reaches up to 3.8×10^4 whereas both $\mathbf{R}_{\text{sing}}$ and $\mathbf{R}_{\text{grid}}$ show similar trends across data sizes, reaching up to 3.4×10^4. Over SKA-like coverages, the DR values are 1.7×10^6, 1.5×10^6 and 1.4×10^6 respectively.

The error images reflect the trend seen in Figs. 5.4c and 5.5c. $\mathbf{R}_{\text{grid}}$ has higher errors, and the complete visibilities set also fails to recover all the diffuse structure in the two lobes. The recovered images show that $\mathbf{R}_{\text{sing}}$ is able to faithfully recover the diffuse structure as well as the bright point-like sources present in the image. The computation time per iteration for reconstruction using $M = 10N$ visibilities was ≈20 seconds when imaging without prior dimensionality reduction, which decreased to ≈7.5 seconds per iteration using $\mathbf{R}_{\text{sing}}$. MS-CLEAN output model images of size 1431 × 3075 pixels were cropped to 477 × 1025 pixels. The model image shown in Fig. 5.8 contains 21507 components and took 5 major iterations. The reconstruction shows smooth regions instead of the details of the diffuse structure present in the test image. The error image illustrates missing features with details of the test image that were not captured by the model.

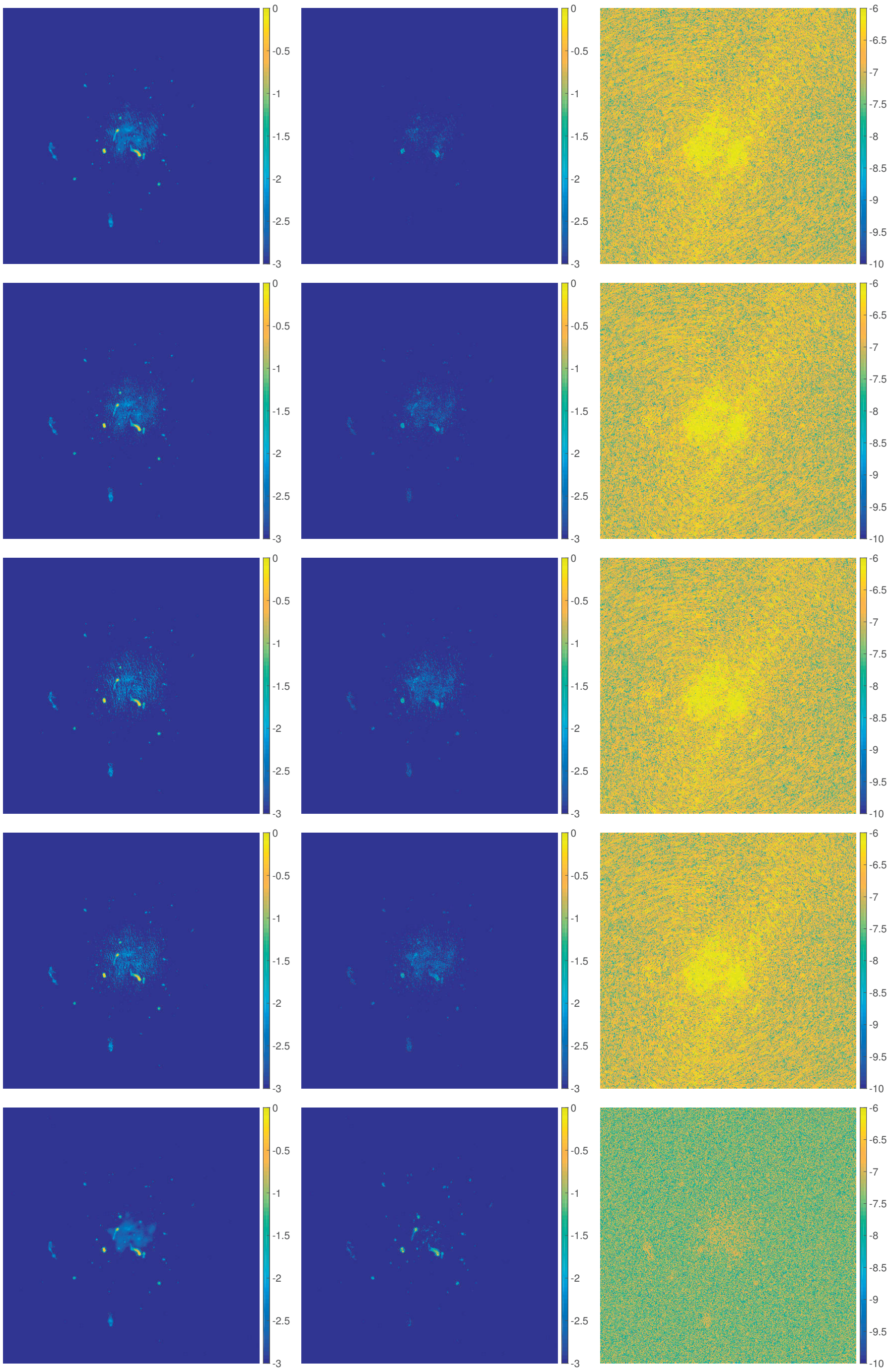


Figure 5.7 – Visual comparison of reconstruction results for the 'Galaxy cluster' test image with $M = 50N$. From left to right: reconstructed, error and residual images in $\log_{10}$ scale. From top to bottom: reconstruction performed with all visibilities using ADMM, and 'reduced' visibilities after performing dimensionality reduction with $\mathbf{R}_{\text{sing}}$, $\mathbf{R}_{\text{dirt}}$, and $\mathbf{R}_{\text{grid}}$ respectively. Last row: reconstruction using MS-CLEAN with a uniform weighting scheme.

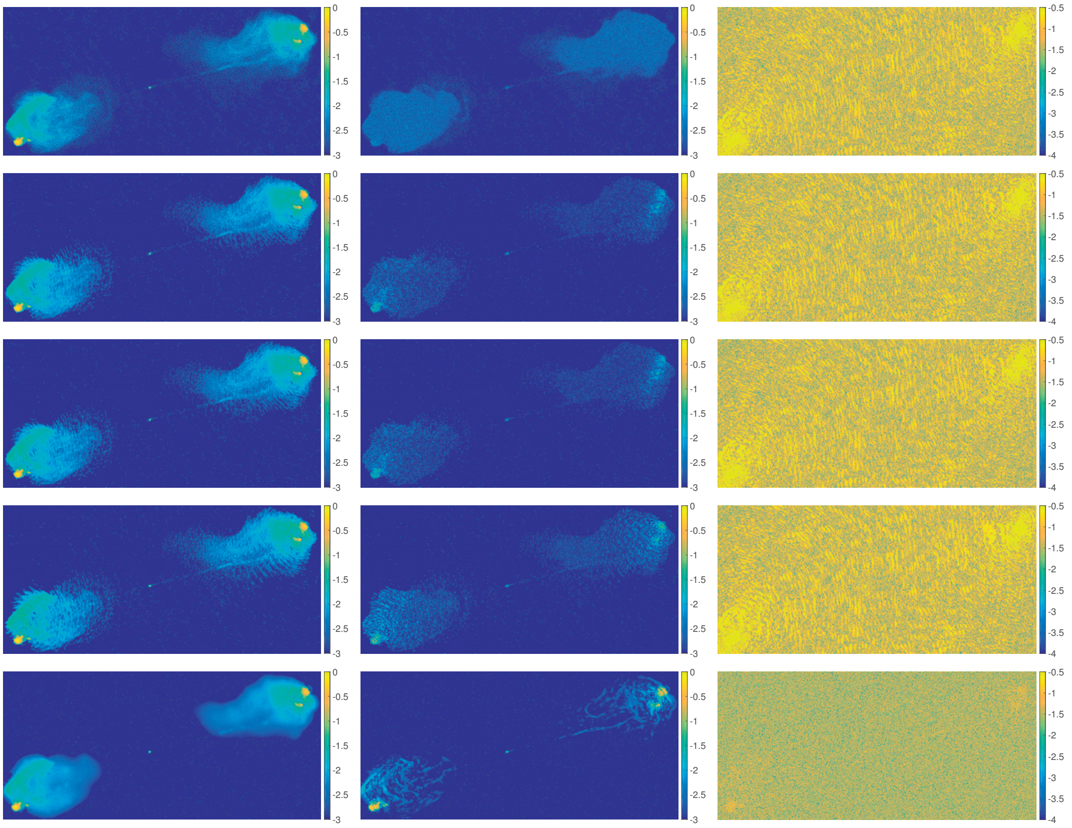


Figure 5.8 – Visual comparison of reconstruction results for the 'Cygnus A' test image with $M = 10N$. From left to right: reconstructed, error and residual images in $\log_{10}$ scale. From top to bottom: reconstruction performed with all visibilities using ADMM, and 'reduced' visibilities after performing dimensionality reduction with $\mathbf{R}_{\text{sing}}$, $\mathbf{R}_{\text{dirt}}$, and $\mathbf{R}_{\text{grid}}$ respectively. Last row: reconstruction using MS-CLEAN with a uniform weighting scheme.

The higher SNR achieved in some cases after dimensionality reduction with $\mathbf{R}_{\text{sing}}$ or $\mathbf{R}_{\text{grid}}$, as compared to $\mathbf{R} = \mathbf{I}$, is a secondary effect – of the approximated embedding to the singular vectors of the measurement operator $\boldsymbol{\Phi}$ in the case of $\mathbf{R}_{\text{sing}}$ and of an effective 'averaging' over neighbouring uv points in the case of $\mathbf{R}_{\text{grid}}$. It may be attributed to the retention of signal information through the null space of $\boldsymbol{\Phi}$ while effectively reducing noise content. It must be noted here, however, that this is a by-product of the design of $\mathbf{R}_{\text{sing}}$ and $\mathbf{R}_{\text{grid}}$. The extent of this apparent denoising depends on the type of image and the actual coverage under consideration, as seen in Figs. 5.4 and 5.5. In all cases, the reconstruction from $\mathbf{R}_{\text{sing}}$ is seen to be at least as accurate as that from the complete visibilities set.

An overarching trend across images, coverages and data sizes is that using an identity matrix to approximate the initial noise covariance matrix results in consistently poorer reconstruction, qualified both in values of output SNR and reconstruction DR. This trend supports our understanding that appropriate handling and a justified approximation of the noise covariance are essential for accurate performance of the image reconstruction algorithms used in this work.

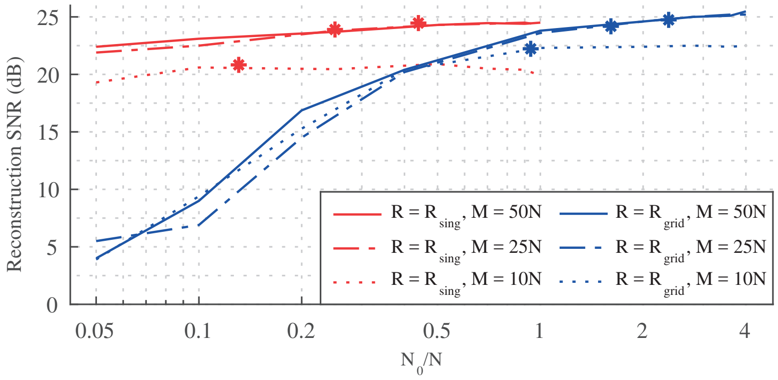
Including noise-dependent thresholding

We investigated a further dimensionality reduction to very small sizes $N_0 \ll N$ and $\overline{N} \ll 4N$ for $\mathbf{R}_{\text{sing}}$ and $\mathbf{R}_{\text{grid}}$ respectively. We performed image reconstruction using reduced data obtained through $\mathbf{R}_{\text{sing}}$ and $\mathbf{R}_{\text{grid}}$, and observed the effects of reducing dimensionality to particularly low values. Fig. 5.9 shows the SNR of reconstructed images from data reduced to sizes ranging from $4N$ all the way down to $0.05N$, which translates to a final low-dimensional data vector of approximately 4,000, 13,000 and 25,000 'reduced' visibilities for the M31, galaxy cluster and Cygnus A images respectively. Reconstruction was performed using initial continuous visibilities of size $10N$, $25N$ and $50N$, simulated over the same SKA-like coverages that were used for obtaining the results shown in Figs. 5.5, 5.6, 5.7 and 5.8. The point of 'diminishing returns' with respect to reduced data dimension and corresponding SNR can be seen as the inflection points in Fig. 5.9 where the SNR no longer remains unaffected by discarding further content.

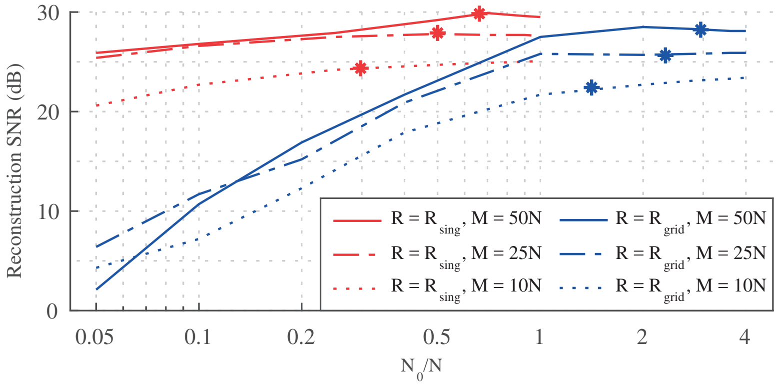
We find that the dimensionality reduction method $\mathbf{R}_{\text{sing}}$ is much more robust to reducing data size below image size, and that we are able to reduce data from an initial visibilities dimension of $50N$ to a final data size of $0.05N$ while decreasing the SNR by less than 5 dB. The method $\mathbf{R}_{\text{grid}}$, however, is seen to be affected adversely from significant dimensionality reduction, and the reconstruction quality dips strongly with decreasing data size to values much below image size. A visual comparison of the artefacts introduced in the reconstruction due to an extremely low data size can be seen in Fig. 5.10 on M31, the galaxy cluster and a zoomed-in portion of the Cygnus A image, highlighting the robustness of image reconstruction after reducing dimensionality with $\mathbf{R}_{\text{sing}}$ as compared with $\mathbf{R}_{\text{grid}}$.

The N_0 and $\overline{N}$ values marked in Fig. 5.9 were computed with $\|\boldsymbol{x}\|_2$ known a priori from the test images. The marked values correspond to $\gamma = 1$ as given in equations (5.20) and (5.21), and provide an indication of a possible estimation of the appropriate final reduced data size, based on our noise-dependent thresholding considerations. The results indicate that a threshold value corresponding to $\gamma > 1$ can probably be safely considered for more aggressive reduction with no significant cost in reconstruction quality.

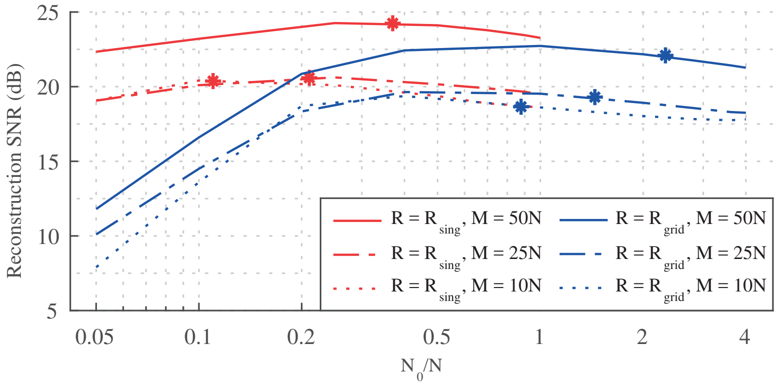
As mentioned in Section 5.1.4, the prior knowledge of $\|\boldsymbol{x}\|_2$ required to compute the thresholds is, in general, not available from interferometric data. It could, however, be approximated by using the dirty image. In the case of $\mathbf{R}_{\text{sing}}$, simple simulations on the test images with known $\boldsymbol{x}$ suggested that for the approximations given by equation (5.9), the use of the dirty image $(\boldsymbol{\Phi}'_{\text{sing}})^\dagger \mathbf{R}_{\text{sing}} \boldsymbol{y}$ instead of $\boldsymbol{x}$ leads to a discrepancy of a few orders of magnitude in the estimation of $\|\boldsymbol{x}\|_2$. This is admittedly a loose bound, and the corresponding threshold was found to be suboptimal with respect to the discarding of data dimensions. Further investigation is needed to check for better estimations of $\|\boldsymbol{x}\|_2$ from the data.



(a) M31

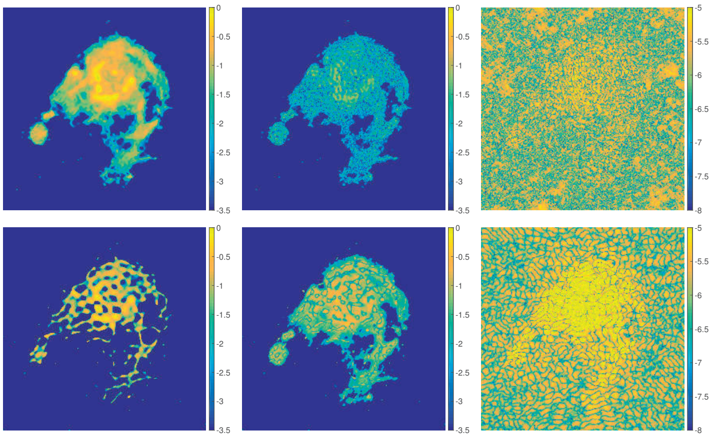


(b) Galaxy cluster

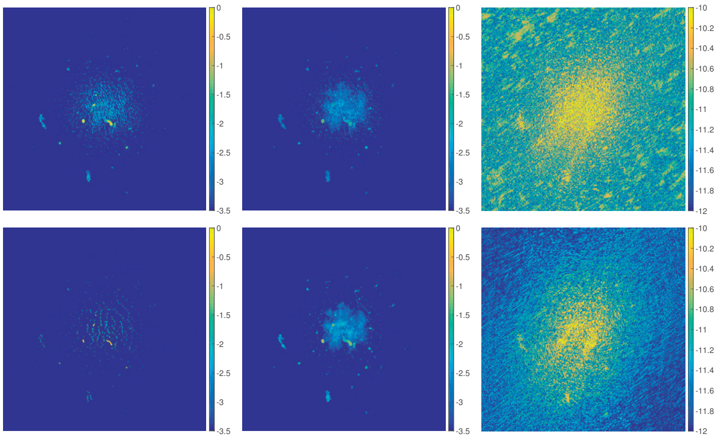


(c) Cygnus A

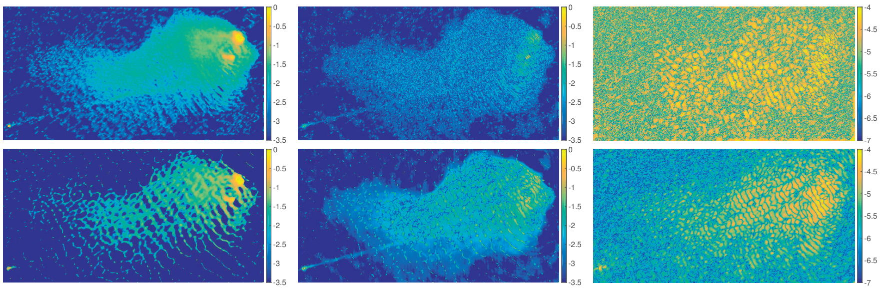
Figure 5.9 – Evolution of reconstruction quality from data reduced to below image size. (a) M31; (b) Galaxy cluster; (c) Cygnus A test images. Initial data size ranges from 650,000 to 12 million visibilities, simulated on SKA-like coverages. Continuous visibilities contain 30 dB additive noise. The analytically computed noise-based threshold marked as '*' shows the values of N_0 and $\overline{N}$ for $\gamma = 1$, which corresponds to the minimum value of N_0 or $\overline{N}$ that ensures that no discarded data is more significant than noise fluctuation as given by equations (5.20) and (5.21).



(a) M31



(b) Galaxy cluster



(c) Cygnus A

Figure 5.10 – Visual comparison of image reconstruction after data reduction to 5% of image size. From left to right: reconstructed, error and residual images in $\log_{10}$ scale. (a) M31, initial data size = 50N. (b) Galaxy cluster image, initial data size = 50N. (c) Cygnus A (partial zoom on top right), initial data size = 25N. For each test image, the top row shows reconstruction from 'reduced' visibilities after dimensionality reduction with $\mathbf{R}_{\text{sing}}$ to $0.05N$, and the bottom row shows reconstruction from 'reduced' visibilities after dimensionality reduction with $\mathbf{R}_{\text{grid}}$ to $0.05N$.

5.3 Summary

We revisited the concept of dimensionality reduction of radio-interferometric data from a compressed sensing perspective. The proposed post-gridding linear data embedding approach, R_{sing}, consists of projecting the data, assumed to be of size much larger than the image size $M \gg N$, to the space spanned by the N_0 left singular vectors of the measurement operator, thus preserving its null space. In the absence of DDEs and calibration errors, we showed that this dimensionality reduction approach consists of first mapping gridded visibilities back to image space, i.e., computing the dirty image, and then performing a weighted subsampled discrete Fourier transform to obtain the final reduced data vector with dimension below image size. The Fourier approximation model for the right singular vectors ensures a fast implementation of the full measurement operator after dimensionality reduction. R_{sing} also preserves the i.i.d. Gaussian properties of the original measurement noise thus making it directly suitable for use with convex optimization algorithms with an ℓ_2-norm data fidelity term. The number of significant singular vectors can be conservatively evaluated by retaining all non-zero singular values, or for further dimensionality reduction to $N_0 \ll N$, by retaining only singular values above a noise-based threshold, effectively introducing a low-rank approximation of the original measurement operator. This is in contrast with current gridding-based imaging in radio interferometry, which reduces data on the oversampled discrete Fourier grid of size $\overline{N}$ or to the dirty image of size N. Our proposed version of gridding, R_{grid}, is also shown to perform well down to low sizes of the order of the image size, but is less robust than R_{sing} when dealing with extremely low sizes.

We show with realistic data simulated using SKA-like uv coverages and using the SARA convex optimization method, that reconstruction quality after embedding the data with R_{sing} is at least as good as with using the complete visibility data set of size M, while being computationally less expensive, both by having a smaller memory footprint thanks to a reduced data size, and through lower running time per iteration of the imaging algorithm. It is also similar in reconstruction quality to results obtained with 'gridded' visibilities obtained through R_{grid} or from the dirty image, but again enabling significantly more reduction below image size.

Another contribution from this work is that dimensionality reduction below $\overline{N}$ can also be achieved from gridded visibilities by discarding those visibilities below a noise-dependent threshold. This reduction by thresholding is however significantly less optimal when applied on gridded visibilities, than on the singular value decomposition. Further work integrating these methods in the PURIFY software [Carrillo et al., 2014] is foreseen as part of the research towards scalable HPC-ready algorithms for radio-interferometric imaging. As we currently assume correctly calibrated data with negligible issues arising from imperfections in data acquisition, future work will include testing the robustness of the proposed methods to w-term effects and calibration errors in particular. The next chapter extends the proposed methods and demonstrates their effectiveness with real data.

6 Real-world performance of dimensionality reduction

Dimensionality reduction pertaining to radio-interferometric data has traditionally been performed using time- and frequency-averaging, with averaging bins chosen according to a combination of factors including the desired data reduction, field of view, ease of applying the FFT, and relevant science objectives among others. As mentioned in chapter 4, these averaging methods, however, can only reduce the data size to a limited extent, depending on the number of time and frequency data-points available respectively as snapshots and channels. Moreover, this type of averaging introduces ‘smearing’ artefacts in the reconstructed images. Smearing presents in the image as attenuated off-centre sources. This attenuation in itself may not always be undesirable, however. For example, a decidedly fortunate effect of smearing is the attenuation of the global background known as far sidelobe confusion noise (FSCN). In general, though, averaging-induced artefacts in an image are detrimental to image quality, since the overall apparent flux is reduced and the point spread function is distorted [Atemkeng et al., 2016]. The ill-effects of averaging visibilities are known and documented in the literature, and several approaches to mitigate them through windowing/filtering methods have been proposed, e.g. by Offringa et al. [2012], Parsons et al. [2016]. Atemkeng et al. [2016] proposed a baseline-dependent windowing method to minimize smearing artefacts, while continuing to use an averaging-based method for data size reduction.

In this chapter we show that images reconstructed using our SVD-based method, R_{sing}, in combination with the PDFB algorithm proposed in Onose et al. [2016] compare favourably to images obtained with the ‘classical’ dimensionality reduction of visibility averaging, and show that R_{sing} performs equally well or better when compared to images reconstructed from the full data set of continuous visibilities. We demonstrate the robustness of R_{sing} down to very low data sizes in a real data setting, using well-calibrated observations of the radio galaxy Cygnus A. We additionally present reconstruction results using our modified gridding-based reduction method, R_{grid}, and make comparisons with R_{sing} in the case of low data size. The mathematical background and motivation for the proposed methods can be found in sections 5.1.1 and 5.1.2.

This chapter contains work described by Kartik et al. [2017c].

6.1 Advanced dimensionality reduction

6.1.1 'On-line' SVD-based dimensionality reduction

As we have noted in section 4.4.2 in the context of Gaussian random projections and in section 5.1.3 in discussing the computational efficacy of our proposed dimensionality reduction operators $\mathbf{R}_{\mathrm{sing}}$ and $\mathbf{R}_{\mathrm{grid}}$, the initial availability of the full data set $y \in \mathbb{C}^M$ implies handling M-dimensional data for the preprocessing step. For instance, applying $\mathbf{R}_{\mathrm{sing}}$ would involve either computing the dirty image from high-dimensional data, or avoiding the higher dimension by pre-computing $\mathbf{\Phi}^{\dagger}\mathbf{\Phi}$ (or more precisely, $\mathbf{G}^{\dagger}\mathbf{G}$). This pre-computation, even though it only needs to be performed once, requires handling M-dimensional structures. This situation, while remedied from the very beginning of the imaging process by the fast application of the combined low-dimensional measurement operator $\mathbf{\Phi}'$, begs to be avoided in the first place, as it seems to counteract some of the computational savings that $\mathbf{R}_{\mathrm{sing}}$ claims to provide. However, the block-separable structure of $\mathbf{R}_{\mathrm{sing}}$ can be exploited to resolve this issue. The holographic matrix $\mathbf{H} = \mathbf{G}^{\dagger}\mathbf{G}$ [Sullivan et al., 2012] combines the steps of computing and gridding visibilities to the discrete Fourier grid into a single, pre-computed mapping. This can be split into separate blocks $\mathbf{H}_{\mathrm{i}}$, giving

$$\mathbf{H} = \mathbf{G}^{\dagger}\mathbf{G} = \sum_{\mathrm{i}} \mathbf{G}_{\mathrm{i}}{}^{\dagger}\mathbf{G}_{\mathrm{i}} = \sum_{\mathrm{i}} \mathbf{H}_{\mathrm{i}} \quad \in \mathbb{C}^{o^2N \times o^2N}. \tag{6.1}$$

The discrete Fourier grid contains o^2N points for an oversampling factor of o in each dimension. We note that the combined measurement operator $\mathbf{\Phi}'_{\mathrm{sing}}$ given in equation (5.11) can then be expressed as a sum of separate block-wise operators, giving

$$\mathbf{\Phi}'_{\mathrm{sing}} = \mathbf{\Sigma}_0^{-1}\mathbf{S}\sum_{\mathrm{i}} \mathbf{\Phi}'_{\mathrm{i}} \quad \in \mathbb{C}^{N_0 \times N}, \tag{6.2}$$

where

$$\mathbf{\Phi}'_{\mathrm{i}} = \mathbf{F}\mathbf{\Phi}_{\mathrm{i}}{}^{\dagger}\mathbf{\Phi}_{\mathrm{i}} = \mathbf{F}\overline{\mathbf{Z}}^{\dagger}\overline{\mathbf{F}}^{\dagger}\mathbf{H}_{\mathrm{i}}\overline{\mathbf{F}\mathbf{Z}} \quad \in \mathbb{C}^{N \times N}, \tag{6.3}$$

$\mathbf{\Phi}_{\mathrm{i}}$ being blocks of the original measurement operator $\mathbf{\Phi}$. We see here that the forward modelling is performed in the lower dimension N_0 through block-wise measurement operators also in the lower dimension N, without ever performing computations in the higher dimension M as outlined originally in equation (2.10). We can thus apply the combined measurement operator in blocks $\mathbf{\Phi}'_{\mathrm{i}}$, leading to the reduced data vector $\boldsymbol{y}'_{\mathrm{i}}$, given by

$$\boldsymbol{y}' = \mathbf{\Sigma}_0^{-1}\mathbf{S}\sum_{\mathrm{i}} \boldsymbol{y}'_{\mathrm{i}} \quad \in \mathbb{C}^{N_0}, \tag{6.4}$$

where

$$\boldsymbol{y}'_{\mathrm{i}} = \mathbf{F}\mathbf{\Phi}_{\mathrm{i}}{}^{\dagger}\boldsymbol{y}_{\mathrm{i}} \quad \in \mathbb{C}^{N}. \tag{6.5}$$

So, as each batch of visibilities $\boldsymbol{y}_{\mathrm{i}}$ is acquired, we can partially apply our dimensionality reduction $\mathbf{R}_{\mathrm{sing}}$ *on-line* by taking a Fourier transform of the dirty image obtained from the batch. This results in an immediate size reduction down to N, the image size – without intermediate steps in the higher dimension M. These N-sized data can be added sequentially as each batch of data is processed. As a final step after all visibilities are acquired and reduced, we can apply the subsampling and weighting through $\boldsymbol{\Sigma}_0^{-1}\mathbf{S}$, to obtain the low-dimensional embedded data that can be fed into imaging algorithms. The data size at this stage would then be lower than image size.

Calculating $\boldsymbol{\Sigma}_0^{-1}$ is crucial as these weights define the importance of the singular values of the measurement operator $\boldsymbol{\Phi}$, and are key in maintaining the reconstruction quality from embedded visibilities. As outlined in the above equations, $\boldsymbol{\Sigma}_0^{-1}$ and $\mathbf{S}$ can be applied at the end of a two-step reduction process, after reducing batches of data to image size N on-line. In the particular case of data under ideal acquisition circumstances, and given our prior knowledge of $\boldsymbol{\Phi}$ – which covers the telescope characteristics and the observation details for a given coverage – we can compute $\boldsymbol{\Sigma}_0^{-1}$ and $\mathbf{S}$ in advance, and apply them as part of the block-wise size reduction, thus further reducing data dimensionality to N_0 per batch, instead of N. However, this is not usually the case, and our prior knowledge of $\boldsymbol{\Phi}$ may need to be corrected after accounting for any anomalous antenna measurements and flagging issues. In this situation, our on-line dimensionality reduction continues to be an attractive option since $\boldsymbol{\Sigma}_0^{-1}$ and $\mathbf{S}$ can be computed and applied *after* all batches of data are acquired, while still only storing image-sized embedded data after each batch of data is processed. One of the main advantages of this on-line method of dimensionality reduction is that the full-sized measurement operator $\boldsymbol{\Phi}$ never needs to be created or handled, thus saving computational resources.

With this scheme of applying the dimensionality reduction $\mathbf{R}_{\text{sing}}$ to batches of data, we propose an avenue to handle high-dimensional data *as they are acquired*. This can potentially be plugged in as a module in the data processing pipeline, leading to an imaging step with already reduced data, while guaranteeing that the information content from the original data is retained.

6.1.2 Gridding-based dimensionality reduction

$\mathbf{R}_{\text{grid}}$ can also be applied on-the-fly to data as they are acquired in batches. We can follow the argument outlined in equations (6.2)-(6.5), and use the block sub-structure of the holographic matrix $\mathbf{H}$ (as defined in equation (6.1)) in a similar fashion, giving

$$\boldsymbol{\Phi}'_{\text{grid}} = \overline{\mathbf{W}}_0^{-1}\overline{\mathbf{S}}\sum_{\text{i}}\boldsymbol{\Phi}'_{\text{i}} \quad \in \mathbb{C}^{\overline{N}\times N}, \tag{6.6}$$

where

$$\boldsymbol{\Phi}'_{\text{i}} = \mathbf{G}_{\text{i}}{}^{\dagger}\boldsymbol{\Phi}_{\text{i}} = \mathbf{H}_{\text{i}}\overline{\mathbf{FZ}} \quad \in \mathbb{C}^{o^2N\times N}. \tag{6.7}$$

In this case, the reduced data vector $\boldsymbol{y}'_\mathrm{i}$ is then given by

$$\boldsymbol{y}' = \overline{\mathbf{W}}_0^{-1} \overline{\mathbf{S}} \sum_\mathrm{i} \boldsymbol{y}'_\mathrm{i} \quad \in \mathbb{C}^{\overline{N}}, \tag{6.8}$$

where

$$\boldsymbol{y}'_\mathrm{i} = \mathbf{G}_\mathrm{i}{}^\dagger \boldsymbol{y}_\mathrm{i} \quad \in \mathbb{C}^{o^2 N}. \tag{6.9}$$

Data can be acquired in batches, and each batch $\boldsymbol{y}_\mathrm{i}$ can be immediately embedded to the oversampled discrete Fourier grid by applying $\mathbf{G}_\mathrm{i}^\dagger$. This reduces the data dimensionality as a first step. Further reduction can possibly be applied by subsampling from this reduced data set. We note from equations (6.7) and (6.9) that the reduction is achieved without ever performing computations in the higher dimension M as modelled in the original measurement operator given by equation (2.10). Moreover, as for $\mathbf{R}_\mathrm{sing}$, $\overline{\mathbf{W}}_0^{-1}$ and $\overline{\mathbf{S}}$ may be computed in advance and applied to each batch of data, thus reducing data size to $\overline{N}$ instead of $o^2 N$ from the beginning, thus avoiding a two step reduction process.

6.1.3 Visibility averaging

Here we continue from our discussion in section 4.4.1 of the conventional method of reducing data dimensionality in radio interferometry through time- and frequency-averaging of continuous visibilities. Time-averaging refers to averaging, across consecutive snapshots, the visibilities that correspond to the same baseline. Increasing the number of snapshots that one includes in an averaging bin leads to a bigger reduction in data size but comes at the cost of a coarser and less accurate coverage of the uv space. Frequency-averaging is performed across spectral channels, averaging over visibilities corresponding to the same baseline for a given snapshot. In the case of narrow-bandwidth channels in an averaging bin, the reduced data may remain a good approximation of the original data, and is indeed a quick and easy dimensionality reduction method. Time- and frequency-averaging, however, have limitations. Due to the limited number of snapshots in typical data sets, time-averaging cannot lead to drastic data dimensionality reduction. As mentioned in section 4.4.1, a major cause of loss of reconstruction quality, however, is the fact that time- and frequency-averaging are typically performed without being appropriately modelled in the measurement operator that is ultimately used for image reconstruction. The measurement operator does not take into account the averaging operation performed, relying only on (now inaccurate) degridding kernels over the Fourier grid which do not correspond to the 'reduced' data. Instead, standard practice is to continue using the measurement operator $\boldsymbol{\Phi}$ instead of $\boldsymbol{\Phi}' = \mathbf{R}_\mathrm{avg}\boldsymbol{\Phi}$ in the imaging process. The side-effects of the averaging can be seen in lower image reconstruction quality. In particular, the effect of averaging over identical bins in all baselines is seen in reconstructed images in the form of 'smearing'. To alleviate the side-effects of the tapering window function on images described in section 4.4.1, Atemkeng et al. [2016] propose baseline-dependent windowing functions which minimize smearing, and further suggest that choosing larger time-averaging bins for shorter baselines (and vice versa) would reduce smearing in the image domain.

Time- and frequency-averaging cannot be performed indefinitely to reach arbitrarily low data sizes. The absolute minimum reachable sizes are governed by initial conditions of the data acquisition, mainly the time intervals between snapshots, the number of channels and the overall uv coverage. In addition, averaging data to achieve very low data sizes may lead to a decline in the reconstruction quality – both in itself and with respect to other data reduction methods. Results of image recovery tests with extremely low-sized data support this conjecture, and visual comparisons between images recovered using visibilities by applying $\mathbf{R}_{\text{sing}}$, $\mathbf{R}_{\text{grid}}$ and simple averaging are shown in section 6.2.

Averaging lends itself readily to on-the-fly application. Indeed, on-line batch processing of acquired data would be the simplest method of data reduction through averaging. The ease of using batch-wise averaging is, however, tempered by the loss in image reconstruction quality that accompanies it. This is particularly relevant for averaging aggressively to reach lower data sizes, as we show through image reconstruction results on real data. In the following section we can see the effects of averaging on recovered images, particularly in contrast with the dimensionality reduction methods that we propose, $\mathbf{R}_{\text{sing}}$ and $\mathbf{R}_{\text{grid}}$.

6.2 Image reconstruction results

6.2.1 Data set details

To test and compare the different data reduction methods described in section 6.1, we consider real data sets of observations of the radio galaxy Cygnus A. The data consist of complex visibilities acquired as part of wideband observations performed in 2015-2016 by the VLA. The data correspond to observations centred at 6680 MHz (the ‘C’ band), over a narrow spectral window of 128 MHz acquired over sixty-four 2 MHz wide channels. Measurements were recorded using the VLA in configuration C, pointing at the phase centre given by RA= 19h 59mn 28.356s ($J2000$) and DEC= $+40°44'2.07''$.[1]

Testing dimensionality reduction methods requires high-dimensional data. For the considered data set, given the relatively small number of data points per channel ($\approx 2 \times 10^5$) and the very narrow spectral window of observations, we decided to collate data from several channels together to form one single uv coverage, from which the aim is to recover a single image. Collating data from all 64 channels, however, was impractical due to computational limitations on (i) reconstructing without dimensionality reduction, and (ii) pre-computing the holographic matrix $\mathbf{H}$ to enable application of $\mathbf{R}_{\text{grid}}$ and $\mathbf{R}_{\text{sing}}$. We note here that this issue can be avoided in the future by applying on-line dimensionality reduction, which would ensure that we never handle the full data set, instead always taking per-block data as input, leading to manageable data sizes at each step of the imaging process. Therefore, we chose 10 separate channels between 6630 MHz and 6720 MHz and concatenated their visibilities together, yielding a uv

---

[1] Right Ascension (RA) and Declination (DEC) are equatorial coordinates that define the direction of a source on the celestial sphere, and function as longitude and latitude equivalents, respectively.

coverage with about 2×10^6 data points. Since the spectral slope in the data set was mild enough to be negligible over the observed bandwidth, we performed single frequency imaging and did not treat different channels separately as one would for hyperspectral imaging. The uv points were normalized to the maximum baseline and subsequently scaled to lie within $[-\pi, \pi]$. An illustration of the uv coverage is shown in Fig. 6.1, with visibilities from three channels.

6.2.2 Image recovery from VLA data

The VLA data were used to reconstruct 256×256 images with a pixel width of 0.5 arcseconds – this corresponds to recovering a signal of up to 2.5 times the band-limit of the observations. The full data set of 2×10^6 continuous visibilities is thus ≈ 30 times larger than the size of reconstructed images ($\approx 6.5 \times 10^4$). Images were reconstructed from data of varying dimensions, obtained through one of the following methods: (i) full data with no dimensionality reduction, (ii) simple averaging over time and frequency bins, (iii) dimensionality reduction by applying $\mathbf{R}_{\text{grid}}$, and (iv) dimensionality reduction by applying $\mathbf{R}_{\text{sing}}$. In the first two cases, image recovery was performed using both MS-CLEAN and PDFB algorithms. In the latter two cases, only PDFB was used.

Fig. 6.2 shows image reconstruction using MS-CLEAN on data of varying sizes, obtained by time- and frequency-averaging the initial data set. The left column shows restored images, where the model, smoothed with the CLEAN beam, is added to the residual image. We do not show the model images as they are not physically realistic. MS-CLEAN was run with Briggs weighting (robust weighting parameter set to -1) and a major loop gain of 0.8, and took 6 major iterations on the full data set to produce a model image of size 2048×2048 from which a 256×256 image was cropped. Reducing the full data set to lower sizes corresponding to $4N$ and N through a simple time-averaging, over 10 and 20 snapshots respectively, led to increasing artefacts in the reconstructed image. We can see regular structures in the residuals corresponding to smearing effects in the reconstructed image. A much lower data size of $0.2N$ was reached by first time-averaging the full data over 30 snapshots and subsequently frequency-averaging over 10 channels. Running MS-CLEAN on this reduced data led to poor image quality, and may be attributed to the unrealistic averaging needed to reach low data size. In the first column of Fig. 6.3, we note that reconstructing images using the PDFB algorithm on the same reduced data sets produced images of better visual quality. This agrees with results previously reported on data without dimensionality reduction [Onose et al., 2016, 2017, Kartik et al., 2017a]. We can nevertheless observe the adverse effects of drastic averaging methods on PDFB, in the form of artefacts for images reconstructed from very low data sizes, like $0.2N$.

We can thus see that averaging has several limitations as a dimensionality reduction method. The final data sizes that can be achieved using averaging are limited by the initial number of snapshots and the number of channels in the data set. The time-averaged data offers a uv coverage that is more incomplete than the original data set and, in addition, the measurement

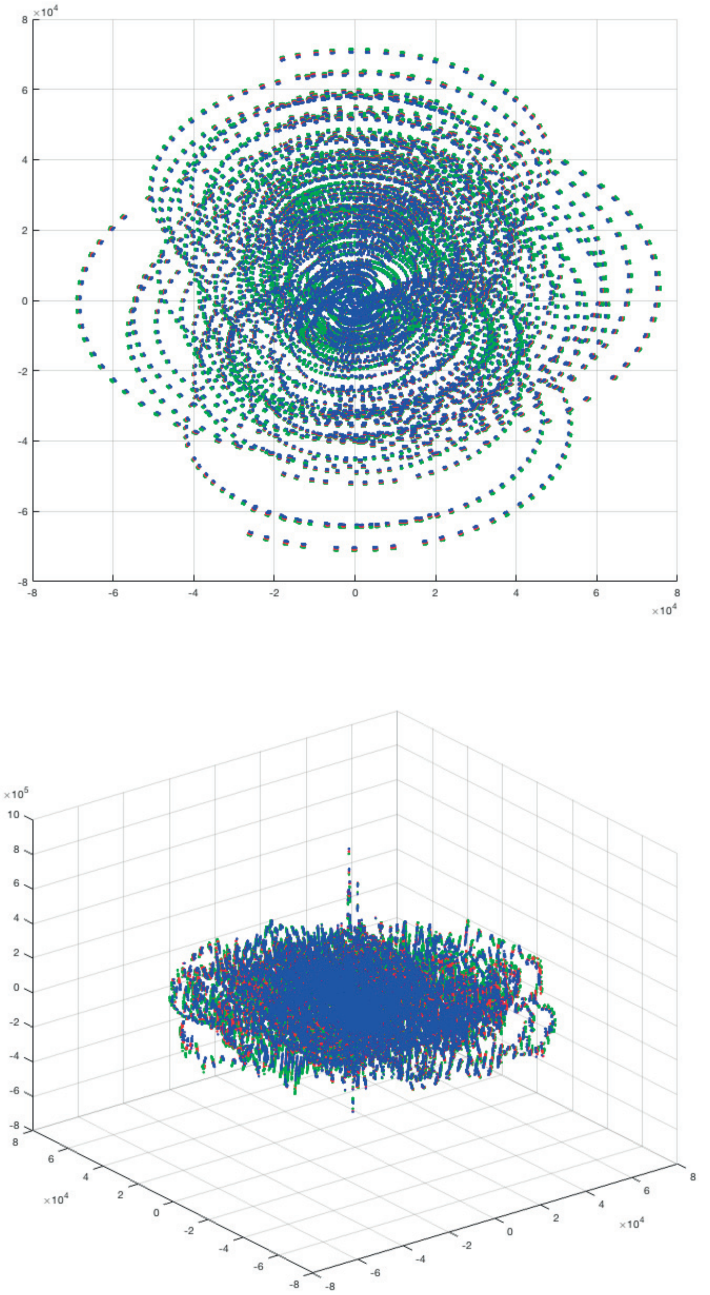


Figure 6.1 – Illustration of VLA uv coverage used in the tests. Visibilities from different channels are collated to cover the same uv plane. The three colours shown here signify three different channels. All uv points were normalised to the maximum baseline. The values shown here are before scaling to lie between $[-\pi, \pi]$, to keep individual channel data visually distinct in spite of the narrow spectral window of the data set. Top: 2-D view, showing the uv coverage; bottom: 3-D view, highlighting the magnitude of the (complex) visibilities. Higher amplitude data points are concentrated in lower frequencies at the centre, as expected.

model is inaccurately approximated due to the omission of the averaging operation $\mathbf{R}_{\text{avg}}$. Consequently, reconstructed images from both MS-CLEAN and PDFB contain related artefacts. The low data size of $0.2N \approx 13\,000$ visibilities is reached by averaging over arbitrarily large bins. Reducing data size in this manner is not meaningful, however, since the corresponding loss of information cannot be compensated for by a simple averaging procedure. Averaging is clearly limited by the need to critically sample the uv plane, and ignoring these hard limits has severe ill-effects on the reconstruction. Averaging produces reasonable images only if the final data size is much higher than that shown here. We were able to reconstruct an image with negligible artefacts with a reduced data size of $7N \approx 455\,000$ visibilities – obtained by time-averaging the collated data set over 5 snapshots – which is much higher than the most conservative reduction performed by $\mathbf{R}_{\text{grid}}$ $(4N)$ or $\mathbf{R}_{\text{sing}}$ (N). Image reconstruction from averaged data may also perform better if the correct measurement model were taken into account, by including the averaging operation $\mathbf{R}_{\text{avg}}$ as mentioned in section 6.1.3. The current work, however, mimics the state-of-the-art averaging method which ignores $\mathbf{R}_{\text{avg}}$ at the expense of inaccurate image recovery.

The second and third columns in Figs. 6.3 and 6.4 show a visual comparison of the reconstructed and residual images, respectively, from data reduced using the dimensionality reduction methods $\mathbf{R}_{\text{grid}}$ and $\mathbf{R}_{\text{sing}}$, and then imaged using the PDFB algorithm. The largest data size that can be achieved after applying $\mathbf{R}_{\text{grid}}$ is $4N$, and that for $\mathbf{R}_{\text{sing}}$ is N, which is why target data sizes above these values are shown as blank spaces in the corresponding columns of Figs. 6.3 and 6.4. On the other hand, both $\mathbf{R}_{\text{grid}}$ and $\mathbf{R}_{\text{sing}}$ allow us – by construction – to attain arbitrarily low data dimensionality by choosing to discard dimensions based either on the significance of the contribution of interpolation kernels to the discrete Fourier grid points (in the case of $\mathbf{R}_{\text{grid}}$) or on the significance of the singular values of the original measurement operator (in the case of $\mathbf{R}_{\text{sing}}$). We see in Fig. 6.3 that both these dimensionality reduction methods outperform averaging for the same target data sizes. We note that for final data sizes of approximately the same order as the image size, i.e., $4N, N, 0.2N$, $\mathbf{R}_{\text{grid}}$ performs as well as $\mathbf{R}_{\text{sing}}$. The robustness of $\mathbf{R}_{\text{sing}}$, however is apparent when data size is aggressively reduced to as low as $0.05N$ and $0.02N$. At these extremely low sizes, we can see that data reduced using $\mathbf{R}_{\text{sing}}$ continue to retain much of the original features of the image (as can be seen in the side lobes in particular) whereas $\mathbf{R}_{\text{grid}}$ appears to recover only the overall structure, producing an overly smooth appearance lacking detail. We note here that the final data size of $0.02N$ is achieved by reducing from an initial data size of $30N$, which represents a dimensionality reduction factor of ≈ 1500, i.e., three orders of magnitude.

Residuals shown in Fig. 6.2 were computed with Briggs weighting using MS-CLEAN. Residuals shown in Fig. 6.4 were computed using the original measurement operator for PDFB. To enable visual comparison of CLEAN and PDFB residual images across columns of Figs. 6.2 and 6.4, residual images obtained using PDFB have been scaled by the peak of the point spread function (PSF).[2] Unsurprisingly, we see an increase in residual structures as we decrease the size of the

---

[2] PSF_{max}, the peak of the instrument response to a point source image at the phase centre with a value 1 at the central pixel and zero otherwise, i.e., $\text{PSF}_{\text{max}} = \max_{\text{i}}(\boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi}\boldsymbol{\delta})_{\text{i}}$, where $\boldsymbol{\delta}$ is the point source image.

data used for image recovery (top to bottom). We note that with R_{sing} (last column of Fig. 6.4), we were able to maintain the absence of regular structures down to very low sizes.

An interesting observation is the similarity in the residual images for R_{grid} for the data sizes $4N$ and N. The corresponding reconstructed images are also very similar to each other. This may be due to the fact that the initial uv coverage was concentrated in the lower frequencies of the (oversampled) Fourier plane, leaving much of the uv plane empty. Consequently, the number of *effective* discrete grid points containing contributions from interpolation kernels was much lower than $4N$ ($\approx 0.6N$ in this particular case). Reducing data size from $4N$ to N, therefore, has no effect on the amount of information contained in the 'reduced' data since the discarded dimensions would correspond to discrete grid points with zero contributions anyway. Thus, the reconstructed images look very similar, and a dip in reconstruction quality is only seen when the data size is reduced below the number of effective discrete grid points.

Running PDFB on the full set of visibilities took ≈ 2 seconds per iteration. Applying R_{grid} or R_{sing} reduced the running time of PDFB to ≈ 0.2 seconds per iteration. This may be attributed to the sparse nature of the constituent operators in R_{grid} and R_{sing}. Additionally, the reduction in data size potentially entails lower memory usage, but this was not directly quantified in our tests. We see a clear computational advantage of performing dimensionality reduction on the initial data set before invoking the imaging algorithm. We also note that the quality of images recovered from reduced data produced with R_{sing} and R_{grid} is comparable to that obtained with the complete set of initial visibilities.

6.3 Summary

We have shown the effectiveness of our proposed dimensionality reduction method, R_{sing}, to handle the large volumes of data expected to be acquired in next-generation radio interferometers like the SKA. It is based on retaining the original information content of the data, and leverages the singular value decomposition of the original measurement operator to achieve this. An alternative reduction method, R_{grid}, is closely related to the familiar method of 'gridding' continuous visibilities to the discrete Fourier grid, and works well when reducing to data sizes close to the image size. We have shown through Cygnus A image reconstruction using VLA data that both R_{sing} and R_{grid} outperform the current standard method of reducing data dimension through simple time- and frequency-averaging. R_{sing} is particularly robust down to extremely low embedded sizes, and is a good candidate for reducing very high-dimensional data. In our case of reconstructing 256×256 size images from well-calibrated VLA data, a final data size of up to 2 per cent of the image size was reached with reasonably low loss in image reconstruction quality. Given our starting data size of 30 times image size ($30N$), a final data size of 2 per cent of image size ($0.02N$) represents a dimensionality reduction factor of ≈ 1500, i.e., more than three orders of magnitude. One can expect significantly higher dimensionality reduction ratios for SKA data when the initial data sizes could be many orders of magnitude larger than image size, while the final data size using R_{sing} would always be lower than image size, potentially reaching much lower, depending on uv coverage and other data acquisition characteristics. In addition to not having the same limitation as averaging methods to reach very low data sizes, R_{sing} and R_{grid} also produce images with fewer reconstruction artefacts for a comparable data size. Owing to the modular nature of the constituent operators of R_{sing} and R_{grid}, we propose a mechanism that enables dimensionality reduction to be applied on-the-fly on data as they are being acquired. This ensures that data size is reduced from the very beginning, thus precluding any issues related to storing or processing large amounts of data in real-time. This could be a possible addition in the data pipelines for the SKA, which currently estimates handling the massive amounts of data flow to be a serious challenge. Further work with R_{sing} is foreseen towards addressing calibration issues, and the suitability of dimensionality reduction in the presence of large w-terms.

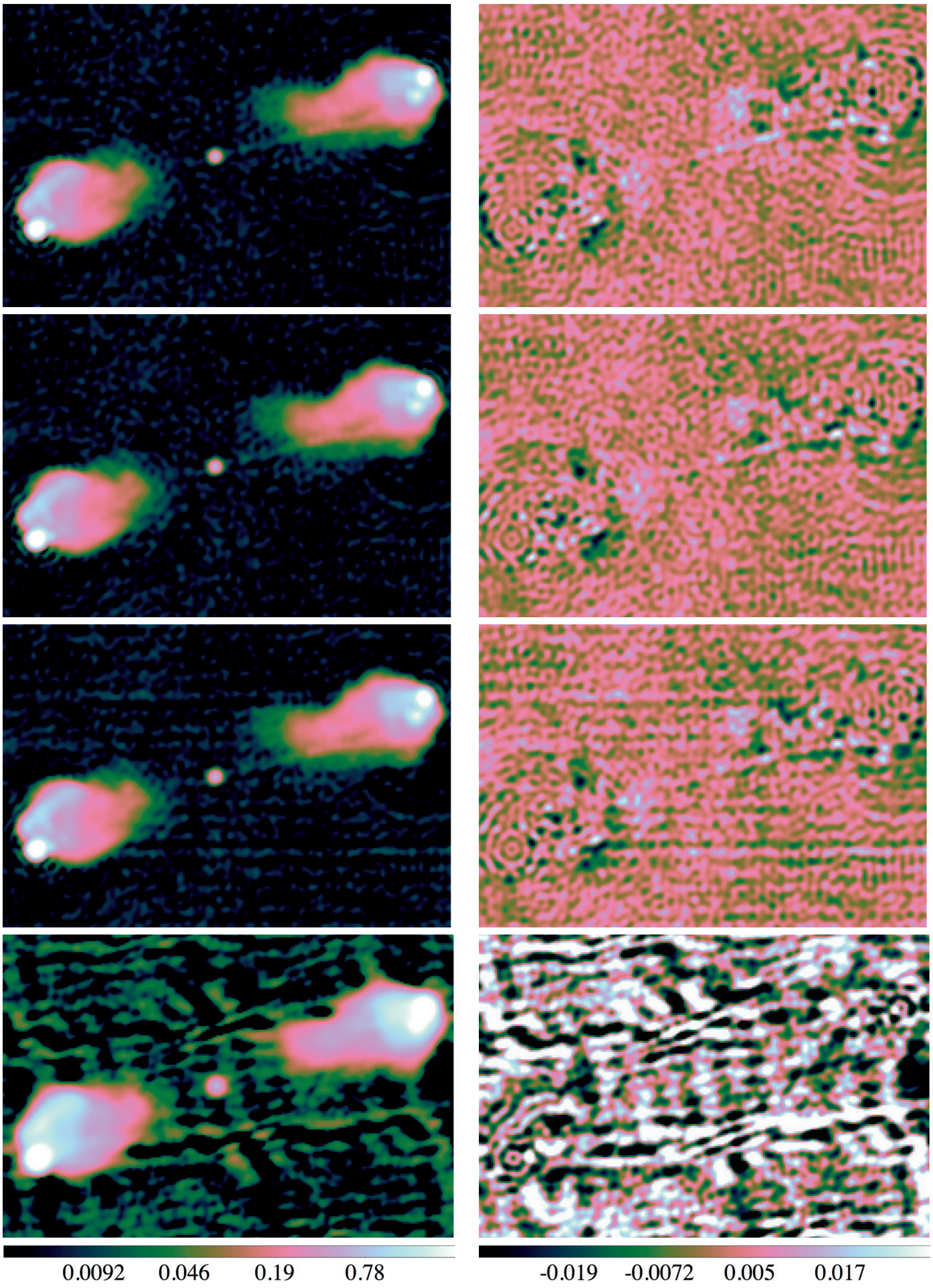
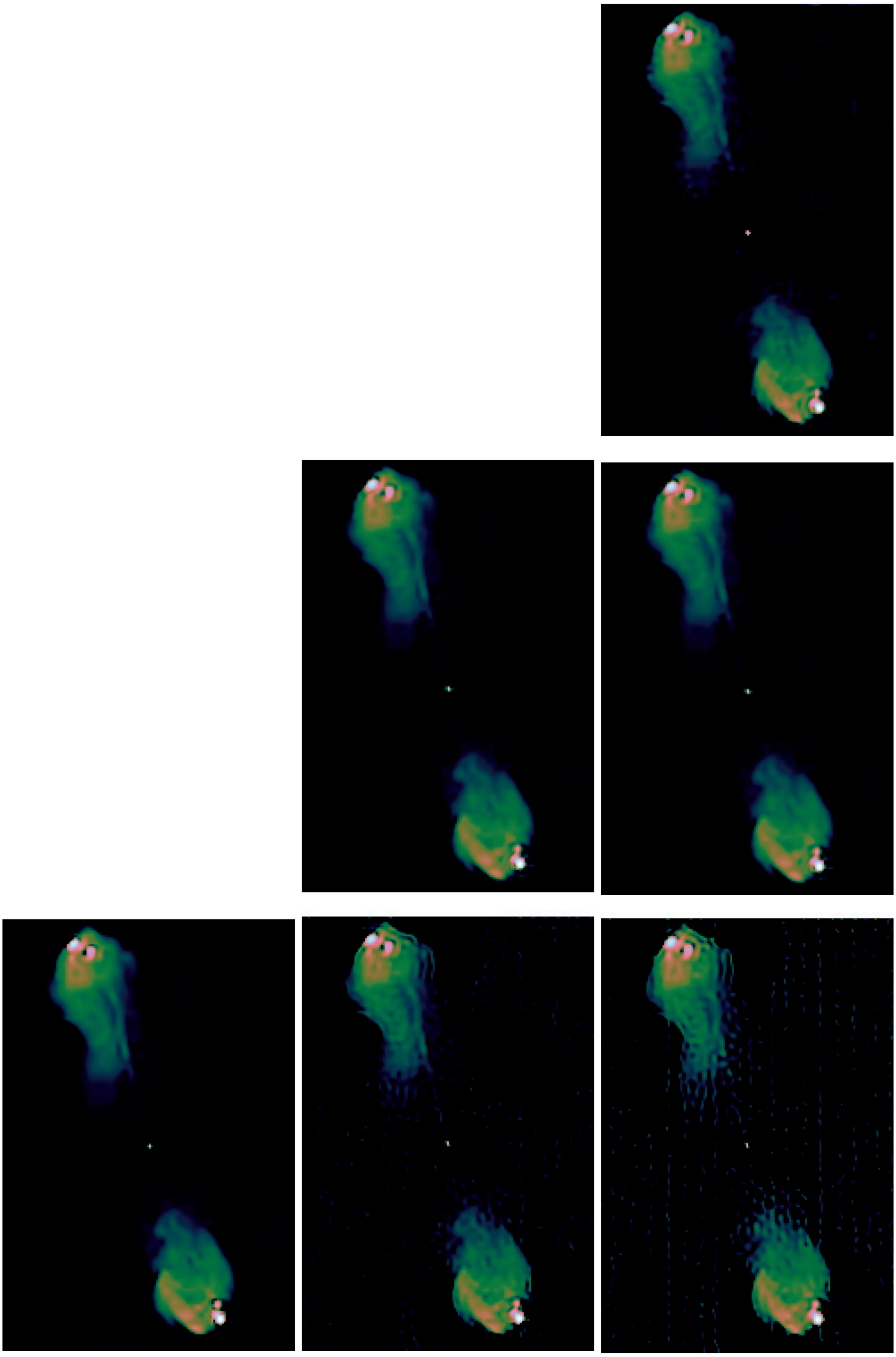


Figure 6.2 – MS-CLEAN image reconstructions using averaged visibilities. Left column: restored images in $\log_{10}$ scale. Right column: Briggs weighted residual images in linear scale. Rows denote final data size achieved after visibility averaging – from top to bottom, $30N \approx 2\,000\,000$ visibilities ($\equiv$ full data, no averaging), $4N \approx 260\,000$ visibilities (time-averaging over 10 snapshots), $N \approx 65\,000$ visibilities (time-averaging over 20 snapshots), and $0.2N \approx 13\,000$ visibilities (time-averaging over 30 snapshots and frequency-averaging over 10 channels). MS-CLEAN was run with Briggs weighting (robust weighting parameter set to -1).



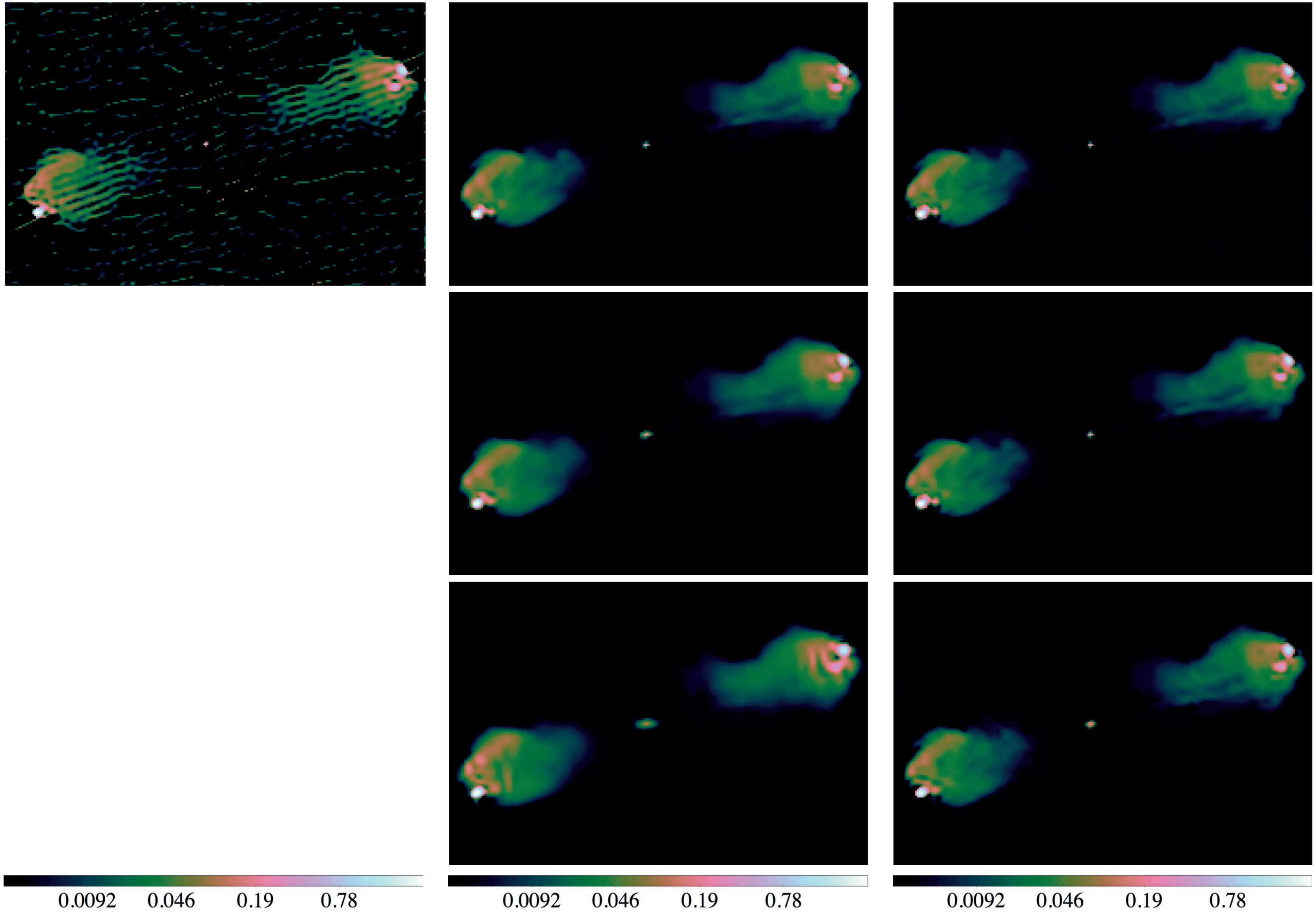
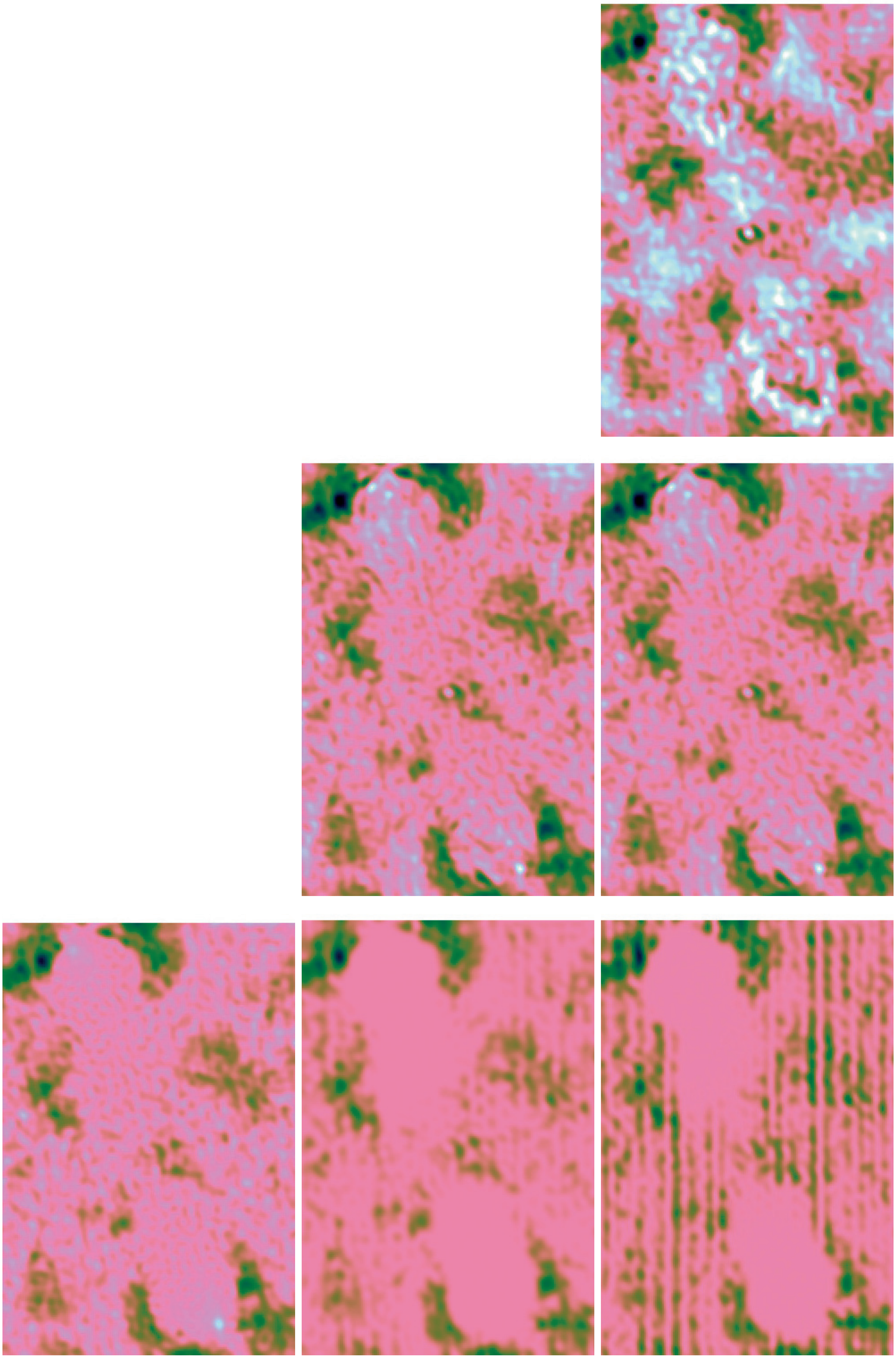


Figure 6.3 – (N.B.: Figure appears over two pages) Reconstructed images using the PDFB algorithm, shown in $\log_{10}$ scale. Columns denote different dimensionality reduction methods – from left to right, (i) time- and frequency-averaging, (ii) $\mathbf{R}_{\text{grid}}$, and (iii) $\mathbf{R}_{\text{sing}}$. Rows denote final data sizes achieved after dimensionality reduction – from top to bottom, (i) $30N \approx 2\,000\,000$ visibilities ($\equiv$ full data, no reduction), (ii) $4N \approx 260\,000$ visibilities, (iii) $N \approx 65\,000$ visibilities, (iv) $0.2N \approx 13\,000$ visibilities, (v) $0.05N \approx 3\,200$ visibilities, and (vi) $0.02N \approx 1\,300$ visibilities. Blank spaces in a column represent data sizes that could not be reached with the corresponding dimensionality reduction method.



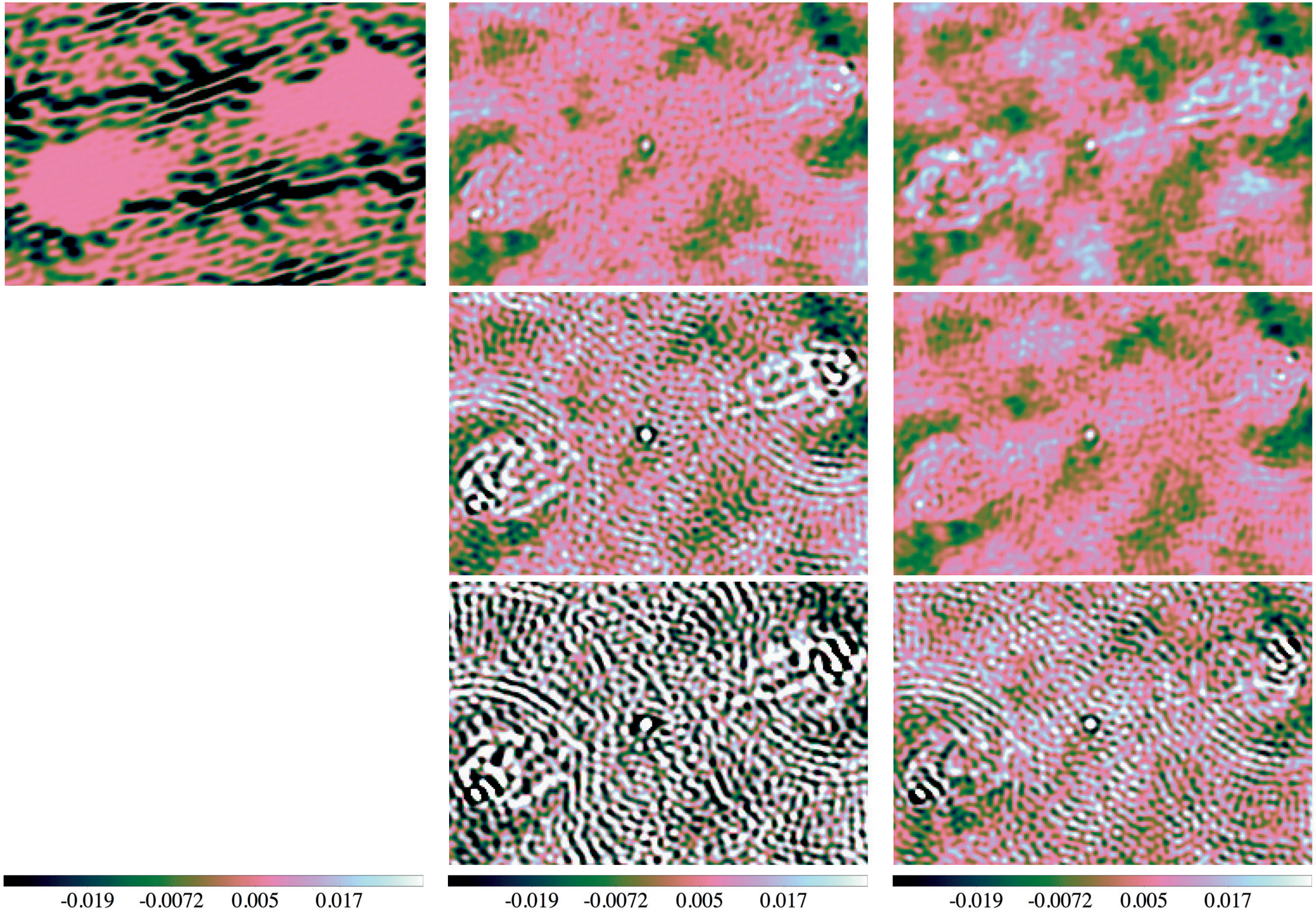


Figure 6.4 – (N.B.: Figure appears over two pages) Naturally weighted residual images for the PDFB algorithm, shown in linear scale. Columns denote different dimensionality reduction methods – from left to right, (i) time- and frequency-averaging, (ii) $\mathbf{R}_{\text{grid}}$, and (iii) $\mathbf{R}_{\text{sing}}$. Rows denote final data sizes achieved after dimensionality reduction – from top to bottom, (i) $30N \approx 2\,000\,000$ visibilities ($\equiv$ full data, no reduction), (ii) $4N \approx 260\,000$ visibilities, (iii) $N \approx 65\,000$ visibilities, (iv) $0.2N \approx 13\,000$ visibilities, (v) $0.05N \approx 3\,200$ visibilities, and (vi) $0.02N \approx 1\,300$ visibilities. Blank spaces in a column represent data sizes that could not be reached with the corresponding dimensionality reduction method.

7 Conclusions

We have proposed novel dimensionality reduction methods for radio-interferometric data. These methods are based on analyzing the singular values of the original measurement operator and a modified visibility gridding, respectively. By embedding the higher-dimensional data into a lower-dimensional Fourier space which preserves all significant singular values, we are able to retain all the information present in the original data. Our proposed reduction methods adhere to principles from compressed sensing, and as a consequence, we are able to reduce data to very low sizes, well below the size of the images to be reconstructed. The adherence to compressed sensing theory guarantees accurate recovery of signals, and also allows us to use convex optimization techniques to reconstruct images from the data. Thanks to recently developed convex optimization algorithms, the image reconstruction process is computationally efficient, while being numerically stable with analytically proven convergence.

7.1 Use of dimensionality reduction in next-generation radio interferometers

Next-generation radio interferometers like the SKA will produce massive amounts of data at an extremely high rate. Dimensionality reduction is one of the most crucial steps that need to be taken to enable appropriate handling of this data explosion, while retaining all information necessary to meet the science goals for these ambitious projects. With our proposed dimensionality reduction methods, we can guarantee that no relevant data are discarded, and that images are reconstructed with scalable, parallelizable algorithms that work across computing nodes in a true HPC setting. The three main advantages of our reduction methods are:

- Extremely fast embedding, thanks to constituent operators being sparse (the subsampling, weighting operators), fast (the FFT), or precomputed (the holographic mapping).
- Robustness to aggressive data size reduction: our SVD-based method performs very well even with extremely low data sizes; experiments on pre-calibrated data from the

VLA have shown that this method can decrease data size by more than three orders of magnitude while maintaining image reconstruction quality. Our gridding-based method produces data with size comparable to the image size while maintaining the information content. It is not as effective for extremely low data sizes, however, and is more suited for cases where aggressive size reduction is not crucial.

- Modular structure, leading to a possible block-wise application to data on-the-fly. This can potentially solve issues with data acquisition and buffering for the SKA, by performing on-line data dimensionality reduction to batches of incoming data, providing an immediate size reduction and thus circumventing any cumbersome large scale data handling.

7.2 Next steps and directions

This thesis proposes dimensionality reduction methods and provides the mathematical development of these methods that enable their use as fast, scalable operations [Kartik et al., 2017a]. The effectiveness of our proposed methods is demonstrated through extensive image reconstruction exercises with various settings, with both simulated data of multiple realistic models and actual observations of a bright radio source, taken with a state-of-the-art interferometer. Our reconstruction results – as compared to those obtained with other existing dimensionality reduction methods – show much improved performance and very promising robustness to drastic size reductions. Kartik et al. [2017c] show image reconstruction results on recent VLA data which demonstrate that our methods make it possible to reach very low data sizes that are beyond the reach of state-of-the art averaging methods used in radio interferometry today.

These favourable results are, however, shown under some assumptions made on the data. One direction to extend the current work is to extend its applicability to cases with fewer such assumptions. For example, we assume that data made available to us are free of calibration errors, and we concentrate on the imaging of these data. While this is useful to showcase the improvements obtained purely in the imaging step, the complete data processing pipeline also involves a calibration step which cannot be ignored. Incorporating our reduction methods in a two-step imaging+calibration set-up is a natural extension to achieve this. Our proposal would need to be upgraded to accept calibration errors – this involves reformulating certain portions, notably involving the generation of the gridding/degridding operators that move data from a continuous visibility space to a discrete Fourier space. Using an extended version, reduced data vectors could be used during both calibration and imaging steps, speeding up the processes individually. A combined set-up would ideally provide a self-contained solution for the data processing pipeline, right from raw interferometer data handling, to the final image ‘products’.

Another real-world issue that has not been addressed in the current work is the presence of large w-terms and other calibrated DDEs. These need to be taken into account during gridding and degridding, as they affect the final image obtained. Including this change to the proposed methods makes them more appropriate for use with actual interferometer data, since such data would indeed contain large DDEs and related corrections.

A major advantage of our methods are their modular structure, allowing a blockwise application to streams of data. This feature can be exploited to provide on-line reduction of data in blocks, as they are acquired from the instrument. This on-the-fly method is perfectly suited for upcoming telescopes where handling, processing and storing all of the incoming data is a serious challenge. While the modified on-line reduction method has been proposed by Kartik et al. [2017c] and explained in chapter 6, its implementation remains to be tested on actual data, to confirm the claimed advantages. One direction for further work in the near future, therefore, is to benchmark this on-line reduction method.

The current work on dimensionality reduction takes data obtained through a sweeping acquisition integrated in time, but at a fixed frequency. As a next step, it also needs to be extended for application to hyperspectral imaging. Embedding visibilities across the frequency dimension in addition to the fixed uv-plane in a given slice of the data cube is essential in handling the complete data acquired by telescopes. A more involved application entails imaging transient sources, which would then necessitate the dimensionality reduction methods to preserve information not only across the uv-plane at one or several frequencies, but also along a temporal dimension at each frequency. Extensions to our proposed reduction methods are, therefore, important to enable its wider use and acceptance in real use cases.

7.3 Closing remarks

Our proposal to perform on-line dimensionality reduction at the same time as data acquisition harks back to the very essence of compressed sensing, promoting a ‘reduced’ sensing paradigm over a more traditional process of acquiring large amounts of data first and then reducing it at a later stage. We have shown that, even in the traditional case of data acquisition followed by data reduction, our proposed methods work better than the state-of-the-art technique of averaging data over time and frequency bins – they not only produce better image reconstruction, but also allow us to reach much lower data sizes. Our work needs to be extended to apply for more use cases seen commonly in radio interferometry, particularly for hyperspectral imaging and imaging of transient sources. In addition, including dimensionality reduction in the data calibration step would be a welcome addition to an otherwise computationally intensive process. Our data reduction methods are tightly coupled with reconstruction methods as proposed by CS practice, and this allows us to exploit algorithmic advancements in the field of convex optimization. We would benefit from a parallel implementation of the reduction methods, and in combination with scalable, parallel imaging methods, we can lay a reasonable claim to be ready for the imminent deluge of big data.

Bibliography

D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo. Fast Image Recovery Using Variable Splitting and Constrained Optimization. *IEEE Transactions on Image Processing*, 19(9): 2345–2356, 2010.

M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo. An Augmented Lagrangian Approach to the Constrained Optimization Formulation of Imaging Inverse Problems. *IEEE Transactions on Image Processing*, 20(3):681–695, 2011.

M. T. Atemkeng, O. Smirnov, C. Tasse, G. Foster, and J. Jonas. Using baseline-dependent window functions for data compression and field-of-interest shaping in radio interferometry. *Monthly Notices of the Royal Astronomical Society*, 462(3):2542–2558, 2016.

A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin. Certifying the Restricted Isometry Property is Hard. *IEEE Transactions on Information Theory*, 59(6):3448–3450, 2013.

R. G. Baraniuk, M. Davenport, R. DeVore, and M. B. Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation*, 28(3):253–263, 2008.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

S. Bhatnagar and T. J. Cornwell. Scale sensitive deconvolution of interferometric images I. Adaptive Scale Pixel (ASP) decomposition. *Astronomy and Astrophysics*, 426(2):747–754, 2004.

S. Bhatnagar, U. Rau, and K. Golap. Wide-field wide-band Interferometric Imaging: The WB A-Projection and Hybrid Algorithms. *The Astrophysical Journal*, 770(2):91, 2013.

T. Blumensath and M. E. Davies. Gradient Pursuits. *IEEE Transactions on Signal Processing*, 56 (6):2370–2382, 2008a.

T. Blumensath and M. E. Davies. Iterative Thresholding for Sparse Approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008b.

J. H. Blythe. A New Type of Pencil Beam Aerial for Radio Astronomy. *Monthly Notices of the Royal Astronomical Society*, 117(6):644–651, 1957.

F. Boone. Weighting interferometric data for direct imaging. *Experimental Astronomy*, 36(1-2): 77–104, 2012.

S. Boyd. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2010.

R. N. Bracewell. Strip Integration in Radio Astronomy. *Australian Journal of Physics*, 9(2): 198–217, 1956.

A. H. Bridle and F. R. Schwab. Bandwidth and Time-Average Smearing. In *Synthesis Imaging in Radio Astronomy II*, volume 180 of *ASP Conference Series*, page 371, 1999.

D. Briggs. High fidelity deconvolution of moderately resolved sources (doctoral thesis). *New Mexico Insitute of Mining Technology*, 1995.

P. C. Broekema, R. V. van Nieuwpoort, and H. E. Bal. The Square Kilometre Array Science Data Processor. Preliminary compute platform design. *Journal of Instrumentation*, 10(07): C07004, 2015.

A. Bruckstein, D. L. Donoho, and M. Elad. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review*, 51(1):34–81, 2009.

G. Brumfiel. High-energy physics: Down the petabyte highway. *Nature News*, 469(7330): 282–283, 2011.

E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969, 2007.

E. J. Candès and M. B. Wakin. An Introduction To Compressive Sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006a.

E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006b.

E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing Sparsity by Reweighted l1 Minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.

R. E. Carrillo, J. D. McEwen, and Y. Wiaux. Sparsity Averaging Reweighted Analysis (SARA): a novel algorithm for radio-interferometric imaging: SARA for radio-interferometric imaging. *Monthly Notices of the Royal Astronomical Society*, 426(2):1223–1234, 2012.

R. E. Carrillo, J. D. McEwen, D. Van De Ville, J.-P. Thiran, and Y. Wiaux. Sparsity Averaging for Compressive Imaging. *IEEE Signal Processing Letters*, 20(6):591–594, 2013.

R. E. Carrillo, J. D. McEwen, and Y. Wiaux. PURIFY: a new approach to radio-interferometric imaging. *Monthly Notices of the Royal Astronomical Society*, 439(4):3591–3604, 2014.

R. E. Carrillo, S. Vijay Kartik, J.-P. Thiran, and Y. Wiaux. A scalable algorithm for radio-interferometric imaging. In *Proceedings of Signal Processing with Adaptive Sparse Structured Representations Workshop*, Cambridge, 2015. Cambridge University Press.

B. G. Clark. An efficient implementation of the algorithm 'CLEAN'. *Astronomy and Astrophysics*, 89:377, 1980.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, 2009.

P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford Splitting Approach to Nonsmooth Convex Variational Signal Recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1 (4):564–574, 2007.

P. L. Combettes and J.-C. Pesquet. Proximal Splitting Methods in Signal Processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, number 49 in Springer Optimization and Its Applications, pages 185–212. Springer New York, 2011.

J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

T. J. Cornwell. Multiscale CLEAN Deconvolution of Radio Synthesis Images. *IEEE Journal of Selected Topics in Signal Processing*, 2(5):793–801, 2008.

A. Dabbech, C. Ferrari, D. Mary, E. Slezak, O. Smirnov, and J. S. Kenyon. MORESANE: MOdel REconstruction by Synthesis-ANalysis Estimators. A sparse deconvolution algorithm for radio interferometric imaging. *Astronomy & Astrophysics*, 576:A7, 2015.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57 (11):1413–1457, 2004.

A. De Mauro, M. Greco, and M. Grimaldi. A formal definition of Big Data based on its essential features. *Library Review*, 65(3):122–135, 2016.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006a.

D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006b.

M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Publishing Company, Incorporated, 1st edition, 2010.

M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947, 2007.

M. Elad, M. A. T. Figueiredo, and Y. Ma. On the Role of Sparse and Redundant Representations in Image Processing. *Proceedings of the IEEE*, 98(6):972–982, 2010.

A. Ferrari, D. Mary, R. Flamary, and C. Richard. Distributed image reconstruction for very large arrays in radio astronomy. In *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 389–392, 2014.

A. Ferrari, J. Deguignet, C. Ferrari, D. Mary, A. Schutz, and O. Smirnov. Multi-frequency image reconstruction for radio interferometry. A regularized inverse problem approach. In *SKA PAthfinder Radio Continuum Survey (SPARCS)*, Skukuza, South Africa, 2015.

J. A. Fessler and B. P. Sutton. Nonuniform fast Fourier transforms using min-max interpolation. *IEEE Transactions on Signal Processing*, 51(2):560–574, 2003.

M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak. Majorization-Minimization Algorithms for Wavelet-Based Image Restoration. *IEEE Transactions on Image Processing*, 16(12): 2980–2991, 2007a.

M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007b.

A. Flinth and G. Kutyniok. PROMP: A sparse recovery approach to lattice-valued signals. *Applied and Computational Harmonic Analysis*, 2017.

I. K. Fodor. *A survey of dimension reduction techniques*. Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, 2002.

M. Fornasier and H. Rauhut. Compressive Sensing. In O. Scherzer, editor, *Handbook of Mathematical Methods in Imaging*, pages 187–228. Springer New York, 2011.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

H. Garsden, J. N. Girard, J. L. Starck, S. Corbel, C. Tasse, A. Woiselle, J. P. McKean, A. S. van Amesfoort, J. Anderson, I. M. Avruch, R. Beck, M. J. Bentum, P. Best, F. Breitling, J. Broderick, M. Brüggen, H. R. Butcher, B. Ciardi, F. de Gasperin, E. de Geus, M. de Vos, S. Duscha, J. Eislöffel, D. Engels, H. Falcke, R. A. Fallows, R. Fender, C. Ferrari, W. Frieswijk, M. A. Garrett, J. Griessmeier, A. W. Gunst, T. E. Hassall, G. Heald, M. Hoeft, J. Hörandel, A. van der Horst, E. Juette, A. Karastergiou, V. I. Kondratiev, M. Kramer, M. Kuniyoshi, G. Kuper, G. Mann, S. Markoff, R. McFadden, D. McKay-Bukowski, D. D. Mulcahy, H. Munk, M. J. Norden, E. Orru, H. Paas, M. Pandey-Pommier, V. N. Pandey, G. Pietka, R. Pizzo, A. G. Polatidis, A. Renting, H. Röttgering, A. Rowlinson, D. Schwarz, J. Sluman, O. Smirnov, B. W. Stappers, M. Steinmetz, A. Stewart, J. Swinbank, M. Tagger, Y. Tang, C. Tasse, S. Thoudam, C. Toribio, R. Vermeulen, C. vocks, R. J. van Weeren, S. J. Wijnholds, M. W. Wise, O. Wucknitz, S. Yatawatta, P. Zarka, and A. Zensus. LOFAR Sparse Image Reconstruction. *Astronomy & Astrophysics*, 575:A90, 2015.

R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies. Greedy-like algorithms for the cosparse analysis model. *Linear Algebra and its Applications*, 441(Supplement C):22–60, 2014.

G. H. Golub and C. F. van Loan. *Matrix Computations*. JHU Press, 1996.

J. P. Hamaker, J. D. Bregman, and R. J. Sault. Understanding radio polarimetry. I. Mathematical foundations. *Astronomy and Astrophysics Supplement Series*, 117(1):137–147, 1996.

S. J. Hardy. Direct deconvolution of radio synthesis images using l1 minimisation. *Astronomy & Astrophysics*, 557:A134, 2013.

K. K. Herrity, A. C. Gilbert, and J. A. Tropp. Sparse Approximation Via Iterative Thresholding. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 3, pages III–III, 2006.

J. A. Högbom. Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines. *Astronomy and Astrophysics Supplement Series*, 15:417, 1974.

L. Hood and D. Galas. The digital code of DNA. *Nature*, 421(6921), 2003.

D. R Hunter and K. Lange. A Tutorial on MM Algorithms. *The American Statistician*, 58(1): 30–37, 2004.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26(189):189–206, 1984.

H. Junklewitz, M. A. Bell, and T. Enßlin. A new approach to multi-frequency synthesis in radio interferometry. *Astronomy & Astrophysics*, 581:A59, 2015.

N. Karmarkar. A New Polynomial-time Algorithm for Linear Programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, pages 302–311, New York, 1984. ACM.

S. Vijay Kartik, R. E. Carrillo, and Y. Wiaux. Dimension embedding for big data in radio interferometry. In *Proceedings of the International BASP Frontiers Workshop 2015*, Villars-sur-Ollon, 2015.

S. Vijay Kartik, R. E. Carrillo, J.-P. Thiran, and Y. Wiaux. A Fourier dimensionality reduction model for big data interferometric imaging. *Monthly Notices of the Royal Astronomical Society*, 468(2):2382–2400, 2017a.

S. Vijay Kartik, R. E. Carrillo, J.-P. Thiran, and Y. Wiaux. Fourier dimensionality reduction of radio-interferometric data. In *Proceedings of the International BASP Frontiers Workshop 2017*, Villars-sur-Ollon, 2017b.

S. Vijay Kartik, A. Dabbech, J.-P. Thiran, and Y. Wiaux. Robust dimensionality reduction for interferometric imaging of Cygnus A. *Submitted to Monthly Notices of the Royal Astronomical Society*, 2017c.

S. Kaski. Dimensionality reduction by random mapping: fast similarity computation for clustering. In *The 1998 IEEE International Joint Conference on Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence*, volume 1, pages 413–418, 1998.

N. Kingsbury. Complex Wavelets for Shift Invariant Analysis and Filtering of Signals. *Applied and Computational Harmonic Analysis*, 10(3):234–253, 2001.

N. Komodakis and J. C. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015. ISSN 1053-5888.

F. Krahmer and R. Ward. New and Improved Johnson–Lindenstrauss Embeddings via the Restricted Isometry Property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

L. Landweber. An Iteration Formula for Fredholm Integral Equations of the First Kind. *American Journal of Mathematics*, 73(3):615–624, 1951.

A. Lannes, E. Anterrieu, and P. Maréchal. CLEAN and WIPE. *Astronomy and Astrophysics Supplement Series*, 123(1):183–198, 1997.

C. Lemaréchal, A. Nemirovskii, and Y. Nesterov. New variants of bundle methods. *Mathematical programming*, 69(1):111–147, 1995.

F. Li, T. J. Cornwell, and F. de Hoog. The application of compressive sampling to radio astronomy-I. Deconvolution. *Astronomy & Astrophysics*, 528:A31, 2011.

P. Li, T. J. Hastie, and K. W. Church. Very Sparse Random Projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’06, pages 287–296, New York, NY, USA, 2006. ACM.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

L. L. McCready, J. L. Pawsey, and Ruby Payne-Scott. Solar radiation at radio frequencies and its relation to sunspots. *Proceedings of the Royal Society of London. Series A*, 190(1022):357–375, 1947.

J. D. McEwen and Y. Wiaux. Compressed sensing for wide-field radio interferometric imaging. *Monthly Notices of the Royal Astronomical Society*, 413(2):1318–1332, 2011.

M. Murgia, F. Govoni, L. Feretti, G. Giovannini, D. Dallacasa, R. Fanti, G. B. Taylor, and K. Dolag. Magnetic fields and Faraday rotation in clusters of galaxies. *Astronomy and Astrophysics*, 424:429–446, 2004.

S. Nam, M. E. Davies, M. Elad, and R. Gribonval. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56, 2013.

B. Natarajan. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

J. Nelson, E. Price, and M. Wootters. New Constructions of RIP Matrices with Fast Multiplication and Fewer Rows. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’14, pages 1515–1528, Portland, Oregon, 2014. SIAM.

Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, Studies in Applied and Numerical Mathematics, 1994.

J. E. Noordam and O. Smirnov. The MeqTrees software system and its use for third-generation calibration of radio interferometers. *Astronomy & Astrophysics*, 524:A61, 2010.

A. R. Offringa, G. A. de Bruyn, and S. Zaroubi. Post-correlation filtering techniques for off-axis source and RFI removal. *Monthly Notices of the Royal Astronomical Society*, 422(1):563–580, 2012.

A. R. Offringa, B. McKinley, N. Hurley-Walker, F. H. Briggs, R. B. Wayth, D. L. Kaplan, M. E. Bell, L. Feng, A. R. Neben, J. D. Hughes, J. Rhee, T. Murphy, N. D. R. Bhat, G. Bernardi, J. D. Bowman, R. J. Cappallo, B. E. Corey, A. A. Deshpande, D. Emrich, A. Ewall-Wice, B. M. Gaensler, R. Goeke, L. J. Greenhill, B. J. Hazelton, L. Hindson, M. Johnston-Hollitt, D. C. Jacobs, J. C. Kasper, E. Kratzenberg, E. Lenc, C. J. Lonsdale, M. J. Lynch, S. R. McWhirter, D. A. Mitchell, M. F. Morales, E. Morgan, N. Kudryavtseva, D. Oberoi, S. M. Ord, B. Pindor, P. Procopio, T. Prabu, J. Riding, D. A. Roshi, N. Udaya Shankar, K. S. Srivani, R. Subrahmanyan, S. J. Tingay, M. Waterson, R. L. Webster, A. R. Whitney, A. Williams, and C. L. Williams. WSClean: an implementation of a fast, generic wide-field imager for radio astronomy. *Monthly Notices of the Royal Astronomical Society*, 444(1):606–619, 2014.

A. Onose, R. E. Carrillo, A. Repetti, J. D. McEwen, J.-P. Thiran, J.-C. Pesquet, and Y. Wiaux. Scalable splitting algorithms for big-data interferometric imaging in the SKA era. *Monthly Notices of the Royal Astronomical Society*, 462(4):4314–4335, 2016.

A. Onose, A. Dabbech, and Y. Wiaux. An accelerated splitting algorithm for radio-interferometric imaging: when natural and uniform weighting meet. *Monthly Notices of the Royal Astronomical Society*, 469(1):938–949, 2017.

A. R. Parsons, A. Liu, Z. S. Ali, and C. Cheng. Optimized Beam Sculpting with Generalized Fringe-rate Filters. *The Astrophysical Journal*, 820(1):51, 2016.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

R. A. Perley, P. Napier, J. Jackson, B. Butler, B. Carlson, D. Fort, P. Dewdney, B. Clark, R. Hayward, S. Durand, M. Revnell, and M. McKinnon. The Expanded Very Large Array. *Proceedings of the IEEE*, 97(8):1448–1462, 2009.

H. Rauhut. *Compressive Sensing and Structured Random Matrices*, volume Theoretical foundations and numerical methods for sparse recovery of *Radon Series on Computational and Applied Mathematics*. de Gruyter, Berlin, 2011.

S. Roberts and R. Everson, editors. *Independent component analysis: principles and practice*. Cambridge University Press, Cambridge ; New York, 2001. ISBN 978-0-521-79298-1.

J. Romberg. Compressive Sensing by Random Convolution. *SIAM Journal on Imaging Sciences*, 2(4):1098–1128, 2009.

M. Ryle. A new radio interferometer and its application to the observation of weak radio stars. *Proceedings of the Royal Society of London. Series A*, 211(1106):351–375, 1952.

M. Ryle and A. Hewish. The Synthesis of Large Radio Telescopes. *Monthly Notices of the Royal Astronomical Society*, 120(3):220–230, 1960.

M. Ryle and D. D. Vonberg. Solar Radiation on 175 Mc./s. *Nature*, 158:339–340, 1946.

M. Ryle and D. D. Vonberg. An investigation of radio-frequency radiation from the sun. *Proceedings of the Royal Society of London. Series A*, 193(1032):98–120, 1948.

A. M. Sardarabadi, AA.mir Leshem, and A.-J. van der Veen. Radio Astronomical Image Formation using Constrained Least Squares and Krylov Subspaces. *Astronomy & Astrophysics*, 588: A95, 2016.

R. J. Sault, J. P. Hamaker, and J. D. Bregman. Understanding radio polarimetry. II. Instrumental calibration of an interferometer array. *Astronomy and Astrophysics Supplement Series*, 117 (1):149–159, 1996.

F. R. Schwab. Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry. *The Astronomical Journal*, 89:1076–1081, 1984.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.

S. Setzer, G. Steidl, and T. Teuber. Deblurring Poissonian images by split Bregman techniques. *Journal of Visual Communication and Image Representation*, 21(3):193–199, 2010.

N. Z. Shor and N. G. Zhurbenko. The minimization method using space dilatation in direction of difference of two sequential gradients. *Kibernetika*, 7(3):51–59, 1971.

O. Smirnov. Revisiting the radio interferometer measurement equation - I. A full-sky Jones formalism. *Astronomy & Astrophysics*, 527:A106, 2011a.

O. Smirnov. Revisiting the radio interferometer measurement equation - II. Calibration and direction-dependent effects. *Astronomy & Astrophysics*, 527:A107, 2011b.

O. Smirnov. Revisiting the radio interferometer measurement equation - III. Addressing direction-dependent effects in 21 cm WSRT observations of 3c 147. *Astronomy & Astrophysics*, 527:A108, 2011c.

J.-L. Starck, J. Fadili, and F. Murtagh. The Undecimated Wavelet Decomposition and its Reconstruction. *IEEE Transactions on Image Processing*, 16(2):297–309, 2007.

I. S. Sullivan, W. Arcus, D. Barnes, G. Bernardi, F. H. Briggs, J. D. Bowman, J. D. Bunton, R. J. Cappallo, B. E. Corey, A. Deshpande, L. deSouza, D. Emrich, B. M. Gaensler, R. Goeke, L. J. Greenhill, D. Herne, J. N. Hewitt, M. Johnston-Hollitt, D. L. Kaplan, J. C. Kasper, B. B. Kincaid, R. Koenig, E. Kratzenberg, C. J. Lonsdale, M. J. Lynch, S. R. McWhirter, D. A. Mitchell, E. Morgan, D. Oberoi, S. M. Ord, J. Pathikulangara, T. Prabu, R. A. Remillard, A. E. E. Rogers, A. Roshi, J. E. Salah, R. J. Sault, N. Udaya Shankar, K. S. Srivani, J. Stevens, R. Subrahmanyan, S. J. Tingay, R. B. Wayth, M. Waterson, R. L. Webster, A. R. Whitney, A. Williams, C. L. Williams, and J. S. B. Wyithe. Fast Holographic Deconvolution: A new technique for precision radio interferometry. *The Astrophysical Journal*, 759(1):17, 2012.

M. Tegmark. How to Make Maps from Cosmic Microwave Background Data without Losing Information. *The Astrophysical Journal Letters*, 480(2):L87, 1997.

A. R. Thompson, J. M. Moran, and G W. Swenson Jr. *Interferometry and Synthesis in Radio Astronomy*. Wiley, 2001.

M. E. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21/3:611–622, 1999.

J. A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 03(01n02):115–126, 2011.

J. A. Tropp and A. C. Gilbert. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

J. A. Tropp, J. N. Laska, Marco F. Duarte, J. Romberg, and R. G. Baraniuk. Beyond Nyquist: Efficient Sampling of Sparse Bandlimited Signals. *IEEE Transactions on Information Theory*, 56(1):520–544, 2010.

S. Wenger, M. Magnor, Y. Pihlström, S. Bhatnagar, and U. Rau. SparseRI: A Compressed Sensing Framework for Aperture Synthesis Imaging in Radio Astronomy. *Publications of the Astronomical Society of the Pacific*, 122(897):1367, 2010.

Y. Wiaux, L. Jacques, G. Puy, A. M. M. Scaife, and P. Vandergheynst. Compressed sensing imaging techniques for radio interferometry. *Monthly Notices of the Royal Astronomical Society*, 395(3):1733–1742, 2009a.

Y. Wiaux, G. Puy, Y. Boursier, and P. Vandergheynst. Spread spectrum for imaging techniques in radio interferometry. *Monthly Notices of the Royal Astronomical Society*, 400(2):1029–1038, 2009b.

N. Wiener. Generalized harmonic analysis. *Acta Mathematica*, 55:117–258, 1930.

S. J. Wijnholds and A.-J. van der Veen. Data driven model based least squares image reconstruction for radio astronomy. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2704–2707, 2011.

J. Yang and Y. Zhang. Alternating Direction Algorithms for l1-Problems in Compressive Sensing. *SIAM Journal on Scientific Computing*, 33(1):250–278, 2011.

S. Yatawatta. Adaptive weighting in radio interferometric imaging. *Monthly Notices of the Royal Astronomical Society*, 444(1):790–796, 2014.

F. Zernike. The concept of degree of coherence and its application to optical problems. *Physica*, 5(8):785–795, 1938.

Avenue du Tir-Fédéral 22
1024 Ecublens
Switzerland
☎ *+41-77 945 76 94*
✉ *vijay.kartik@gmail.com*
svijaykartik
vijaykartik

S. Vijay Kartik

Professional experience

Sep. 2013 – Present
EPFL (Ecole Polytechnique Fédérale de Lausanne), Lausanne, Switzerland.
Doctoral Assistant, Signal Processing Laboratory (LTS5)

- Researching dimensionality reduction methods for big data astronomy imaging
- Designing fast, distributed signal recovery from telescope data
- Achieved 30x size reduction on datasets with $\approx$ 50 million data points
- Developed online imaging algorithm, leading to 40% reduction in RAM usage
- Expertise in: high performance computing, compressed sensing, convex optimization

Jan. 2011 – Aug. 2013
CERN (European Organization for Nuclear Research), Geneva, Switzerland.
Research Fellow, LHCb Computing

- Benchmarked the ARM architecture for large-scale compute performance under low-power; ported a 3.6 MLOC High Energy Physics (HEP) C++ software stack - first LHC experiment to test-run on ARM. Performance benchmarks and findings available at `http://goo.gl/cBrK4`
- Developed and maintained the triggered event-writing software of the data acquisition pipeline at LHCb, one of the four large LHC experiments at CERN
- Investigated distributed file systems and scalable performance in HEP computing environments
- Developed and extended software-based tools for high-speed data emulation and processing – currently deployed as a modified test-bench simulating 'online' data acquisition
- Member of the system administrators team in LHCb, managing a cluster of $\approx$ 5000 compute nodes

Oct. 2008 – Oct. 2010
ASUS (ASUSTeK Computer Inc.), Taipei, Taiwan.
Software R&D Engineer, Core Technology Center

- Designed speaker adaptation techniques for non-native English speech; achieved 15% performance improvement in base recognition rate
- Wrote speech firmware and control drivers for ARM Cortex A8 and NIOS II processors (Currently being used in a prototype robot at ASUS)
- Developed modified device drivers for new electrophoretic display technology as a member of the core team for the award-winning E-Book reader ASUS DV-950

Education

Sep. 2013 – Present
EPFL (Ecole Polytechnique Fédérale de Lausanne), Switzerland.
Ph.D. in big data imaging

Ph.D. thesis
Low-dimensional data embedding for scalable astronomical imaging in the SKA telescope era

Research themes: Large scale data analysis, convex optimization, compressed sensing, inverse problems, radio astronomy, Square Kilometre Array (SKA)

2003 – 2008
IIT Bombay (Indian Institute of Technology Bombay), India.
Dual Degree (Bachelor's and Master's) in Electrical Engineering

Graduated with First Class

2003 **Kendriya Vidyalaya (Central Board of Secondary Education)**, India.
Senior Secondary School Leaving Certificate (12th Grade)
Graduated with First Class, ranked first

Short-term research projects

Summer 2007 **Microsoft Research Summer School on ASR**, *IEEE Signal Processing Society*, Bangalore.

- Development and evaluation of projects involving audio I/O, HMM-based speech recognition
- ‘Silver’ position after evaluation of projects
- One out of only four teams to complete all projects successfully

Summer 2006 **Siemens Corporate Technology - India**, *Siemens Information Systems Ltd.*, Bangalore.

- Designed and developed a ‘Very Low Bit-Rate segmental speech coder’
- Achieved target bit-rates of 600 bps
- Resulting work and subsequent algorithms applied for patent by Siemens Corporate Technology

2006–2007 **Electronic Design Project**, *IIT Bombay*, Mumbai.

- One-year project with entire mechanical design, circuitry, control logic and software built bottom-up and developed in a team of four
- Built a capacitance-based liquid-drip sensor (using measurements in the picofarad (pF) range) for an intravenous (IV) drip rate meter

Selected publications

- **S. Vijay Kartik** *et al.*, “A Fourier dimensionality reduction model for big data interferometric imaging”, *Monthly Notices of the Royal Astronomical Society*, 2017.
- **S. Vijay Kartik** *et al.*, “Dimension embedding for big data in radio interferometry”, *Proc. Int. Biomedical and Astronomical Signal Processing Frontiers*, 2015.
- **S. Vijay Kartik** *et al.*, “Measurements of the LHCb software stack on the ARM architecture”, *J. Phys: Conf. Ser.*, 2013.
- **S. Vijay Kartik** and Niko Neufeld, “Comparative Investigation of Shared Filesystems for the LHCb Online Cluster”, *J. Phys: Conf. Ser.*, 2012.

Scholastic activities and achievements

2016 Won the Swiss Federal Railways (SBB) challenge at the Lausanne Hackathon

2007 Represented India at the “Innovate 2007” conference on technology, globalisation and leadership at Bangalore. One out of six students selected from India

2003 – 2008 Awarded government-sponsored merit scholarship for entire undergraduate course duration

2003 Ranked first in school-leaving board examinations. Also awarded ‘Certificates of Merit’ in Physics and English (Top 0.1% of *all* qualifying students from India)

Technical skills

Languages	C/C++, Python (≈ 10 years)
	MATLAB (≈ 5 years)
System/Tools	Linux (> 10 years), Git (≈ 5 years)

Languages

Native fluency	English, Hindi, Tamil
Advanced (C1)	French