Development of the CHORD Galaxy Search Strategy

by

Hans Hopkins

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Physics

Waterloo, Ontario, Canada, 2024

© Hans Hopkins 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis presents my contribution to the CHORD galaxies science case. I helped develop an algorithm that can automatically pick out galaxies from CHORD driftscan data. The method used is a matched filter, and it acts on spatial data and spectral data.

On the spatial side, it searches for point sources. Because CHORD is a highly redundant interferometer, it suffers from spatial aliasing. I present a program that is able to predict the severity of spatial aliasing. It predicts that integrating over periods of time is required to break the alias degeneracy, and that "dithering" CHORD (rotating it by a couple degrees) further helps in reducing the alias issue. I offer a framework for estimating the probability of spatial alias confusion.

On the frequency side, I present a method of running the matched filter quickly. CHORD frequency data undergoes a process called upchannelization, which would distort the shape of a galaxy profile. I show how this can be accounted for without incurring a large time-cost penalty.

Lastly, I discuss how a full matched filter program would be put together, and implications that my research has on selecting search parameters for a future CHORD galaxy survey.

Acknowledgements

I would like to thank:

- Dustin Lang for helping me with questions and bugs no matter how long they took to resolve, as well as detailed feedback about the writing and plots.
- Kendrick Smith for designing and proposing this project and meeting with me many times for guidance.
- James Taylor for his advice about structuring the thesis and for answering theory and survey design questions. I would also like to thank his Galaxy Cluster group for discussions about cluster physics.
- Will Percival and the Large Scale Structure group for helping with my understanding of cosmology. Percival also helped me regarding angular correlations early on in the project.
- The CHORD Galaxy group, including Adrian Liu, Kristine Spekkens, Akanksha Bij, Rebecca Ceppas de Castro, Olivia Pereira, and Chaitanya Khamar for our biweekly meetings, specifically for testing code, suggesting ideas, and improving my understanding of the subject.
- Jonathan Sievers for help with producing the predicted counts plot. He also taught me about matched filters, and so is responsible for my wanting to pursue this project in the first place.
- Seth Siegel and Rebecca Ceppas de Castro again for help with the M-mode mapmaker.
- Erik Schnetter for helping me with our local compute cluster and some C++ questions.
- Alice Mikon for helping with proofreading.
- Lisa Hopkins for her support.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Figures viii

List of Abbreviations x

List of Symbols xi

1 Introduction 1

1.1 Observing Galaxies with Radio Motivation 1

1.2 Science Questions 4

1.2.1 Gas Depletion and Accumulation Mechanisms 4

1.2.2 HI Mass Function 5

1.2.3 Optically Dark Galaxies 6

1.2.4 Stacking Analyses 7

1.3 Existing HI Galaxy Experiments 7

1.4 Neutral Hydrogen in Galaxies 9

1.5 CHORD Introduction 12

1.5.1 CHORD Sensitivity Analysis . . . 15
1.6 Method . . . 19
1.7 Previous Detection Algorithms . . . 20
1.8 General Theory of Matched Filtering . . . 21

2 Spatial Matched Filter **23**
2.1 Spatial Introduction . . . 23
2.2 Analytic Matched Filter (AMF) Intro . . . 27
2.3 Instantaneous Analysis . . . 28
2.4 Integrated Analysis . . . 33
2.4.1 Safety Parameter Sweep . . . 37
2.4.2 Alias Analysis . . . 37
2.5 Dithering . . . 47
2.6 Probability . . . 53
2.6.1 Case 1: 100% Detection Probability . . . 54
2.6.2 Case 2: Possibility of No Detection . . . 56

3 Frequency Matched Filter **60**
3.1 Frequency Axis Templates . . . 60
3.2 FFT Matched Filter . . . 65
3.3 Multiple Detections . . . 69
3.4 Upchannelization . . . 70

4 Discussion **76**
4.1 Combined Matched Filter Design . . . 76
4.1.1 Post Matched Filter . . . 78
4.2 Integration with the Rest of the CHORD Galaxies Team . . . 79
4.2.1 Choosing Parameters for a Survey . . . 80
4.3 Matching Galaxies with Optical Surveys . . . 82
4.4 Optimizing a Survey for Faint Galaxies . . . 85

5 Conclusion 87

References 89

APPENDICES 96

A Matched Filter Proof 97

B Omitting the Re() in the matched filter equation 99

C Closest Alias Locations in the Instantaneous Case 101

D Correlation Theorem Proof 103

List of Figures

1.1 The Schechter function that was found using 40% of the full ALFALFA catalogue . . . 6
1.2 THINGS measurements of NGC 6946 . . . 11
1.3 HI surface brightness profile of NGC 6946 . . . 12
1.4 The NGC 6946 rotation curve . . . 13
1.5 Predicted galaxy detection counts . . . 18

2.1 Instantaneous spatial case . . . 32
2.2 Alias difference in primary beam crossing . . . 35
2.3 Alias path deviation from curvature . . . 36
2.4 Alias wavelength dependence . . . 38
2.5 Verifying convergence with respect to time sampling rate . . . 39
2.6 Integrated spatial case . . . 40
2.7 Integrated spatial case near the North Pole . . . 41
2.8 Computing north alias declination . . . 42
2.9 Computing four nearest alias locations . . . 44
2.10 Alias strength dependence on the source's position in the primary beam . . . 45
2.11 Alias strength dependence on CHORD's declination . . . 46
2.12 High-correlation-coefficient alias in the integrated case . . . 48
2.13 Breaking alias degeneracy with dithering . . . 50
2.14 Optimal dithering angle . . . 51

2.15 Alias degeneracy breaking with multiple dithers 52
2.16 Illustration of the PDF that decides alias confusion 57
2.17 Probability of alias confusion as a function of CHORD declination 58
2.18 Probability of alias confusion as a function of source strength 59

3.1 An example disk or recessional velocities and their implied spectrum 62
3.2 The derived spectrum of a flat disk including dispersion velocity 63
3.3 Comparison of frequency template families 65
3.4 1-dimensional matched filter example 66
3.5 Testing the importance of matching the frequency template width to the source 67
3.6 Testing the importance of matching the frequency template shape to the source 68
3.7 Showing that the code can return multiple detection in the same spectrum 69
3.8 The effects of upchannelization on a galaxy profile 71
3.9 Showing that the effects of upchannelization changes depending on the profile's position 72
3.10 The matched filter accounting for upchannelization 75

4.1 Width histogram from the ALFALFA catalogue 81
4.2 Comparison to optical counts 83
4.3 Minimum mass threshold 84
4.4 Maximum magnitude threshold 85
4.5 Comparing high and low declination options for maximizing counts of faint objects 86

List of Abbreviations

ALFALFA Arecibo Legacy Fast ALFA Survey 8

AMF Analytic Matched Filter 26, 28

CHORD Canadian Hydrogen Observatory and Radio-transient Detector 12

DRA Digital Research Alliance of Canada 78, 82

FFT Fast Fourier Transform 20

HA Hour Angle 15, 36

HIPASS HI Parkes All Sky Survey 8

ICM Intracluster Medium 4

PDF Probability Distribution Function 54

RFI Radio Frequency Interference 15

SEFD System Equivalent Flux Density 7

SNR Signal to Noise Ratio 20

THINGS The HI Nearby Galaxy Survey 8

VLA Very Large Array 8

List of Symbols

HI Unionized hydrogen also called neutral hydrogen 2

L^* Luminosity Cutoff in Press-Schechter 5

L Luminosity 5

$\hat{e}_\rho$ ρth basis vector of the rectangular array 30

$\vec{u}_p$ A unit vector representing CHORD's pointing 24

σ Constant noise value 26

$R_{x'x}$ Correlation Coefficient. How much does the matched filter response with the x label correlate with the matched filter response with the x' label. 22

Σ The covariance matrix of the PDF in the probability section 54

N Covariance matrix 22

θ Declination. This has the dual purpose as an index of dithering angles. 47

$d(\vec{u})$ The dirty map. A map showing the instrument response at the $\vec{u}$ sky position 14

$\vec{x}_i$ The ith dish position in the array in length units 14

L Dish separation in length units. In general, the two directions of the rectangular array can have two different separations L_1, L_2. 25

η_A Aperture Efficiency. Multiplicative parameter that accounts for the effective dish area 8

μ The expectation value of the PDF in the probability section 54

t_{int} Integration time 17

M_x The matched filter indexed by parameter vector x. Sometimes the index is dropped. 22

m_i The number of dishes in the ith axis of the rectangular array. There is in total $m_1 m_2$ dishes. 27

ϕ^* Overall count in Press-Schechter 5

B^2 The primary beam 24

P Probability 54

R Response Matrix 70

ω Earth's rotational velocity in rads/time 33

$\mathcal{R}(\phi)$ The rotation matrix with respect to Earth's spin axis by angle ϕ 33

σ_T The sigma threshold to count as a detection 17

$\vec{u}$ A unit vector pointing at the sky 14

$\vec{u}_s$ A unit vector representing the position of a point source 24

$t(x)$ A template vector used in the matched filter equation. Sometimes written t_x. When indexed by i, $t_i(x)$. 22

τ Time in units of time, but equivalently used as an index over the time axis 33

U Upchannelization parameter 15

V_{ij} 2-dimensional visibility array (or visibilities) indexed by i and j 14

λ Wavelength 13

Chapter 1

Introduction

1.1 Observing Galaxies with Radio Motivation

In this thesis, I present predictive and software development work for the CHORD galaxy search science case. CHORD will produce a catalogue of radio spectra of galaxies that a pipeline will detect. This thesis explains my contributions to the pipeline, as well as forecasting that I have done about detection quality. The introduction explains the CHORD galaxies science objectives.

Structure evolution is well understood in the linear regime on scales larger than galactic superclusters. On scales smaller than this, baryonic effects such as feedback and cooling have a large enough impact that large scale structure simulations disagree with each other about measurables such as gas content of galaxies or number of dwarf galaxies in halos. Observations of galaxy characteristics, particularly the histories of those characteristics, will help constrain models of galaxy, group, and cluster evolution. A big point of interest is the scenario where baryons and dark matter do not trace each other. This can indicate nontrivial merging history, pressure stripping, or feedback, which are all processes that require special attention in structure formation models. Dark matter content can be measured with gravitational lensing, and baryonic matter emits radiation that can be measured directly. This thesis is about the process of efficiently identifying the 21 cm line from neutral hydrogen inside galaxies in order to take advantage of a statistically large catalogue of galaxy gas content.

Two major questions about galaxy evolution are how do galaxy populations statistically change over time and how does the environment impact galaxy evolution. An example of

the former is the Butcher-Oemler effect. The populations of galaxy clusters tend to be more blue and star forming in the past (about $z = 0.3 - 0.5$) compared to today. Their shapes also change. This indicates that galaxy populations evolve with redshift. It is interesting to track the gas fraction of galaxies up to higher redshift in order to statistically measure quenching over time. There is also evidence that the local environment has an impact on the HI content of galaxies. Galaxies in groups and clusters are statistically distinguishable from the ones in the field. Galaxies in groups or clusters more often have morphologies that indicate collisions or tidal stripping [14]. It is difficult to capture all of these effects in simulations for a variety of reasons, especially when considering internal matter-depletion mechanisms like AGN and supernova feedback. Simulations need to be fine-tuned to incorporate small time or space scale effects and fix other free parameters, which would allow us to predict the luminosity function to more accuracy for example. We need observations to complete our understanding, and gas is a critical component of galaxies to simulate.

Baryons are often not considered in simulations for simplicity. The scale of mechanisms that affect baryons like supernova feedback are too small for the resolution of cluster-scale simulations. However, gas is critical to understanding galaxy evolution, not only because it traces impacts from the local environment, but also because high gas density causes star formation. There are known mechanisms of gas accumulation and depletion which will be explained in section 1.2.1. We would like to measure how strong these mechanisms are over time and what is their environmental dependence. Gas density and position can be measured by observing particular emission lines caused by atomic processes, chemical bonds, and macro electromagnetic processes. This thesis will focus on measuring gas using neutral hydrogen and its 21 cm emission line.

The interstellar medium in a typical galaxy contains ionized and unionized hydrogen (also called neutral hydrogen or HI), molecular hydrogen, helium, and metals at a much smaller proportion. Unionized hydrogen is an attractive choice to study in the context of galaxies for several reasons. First, it emits a characteristic 21 cm signal which is unlikely to be confused with other lines due to the abundance of hydrogen. This wavelength signal is advantageous for observation because it can pass through dust and Earth's atmosphere. Second, hydrogen outside of thick clouds is easily ionized because of much higher photon penetration distance, so hydrogen closely traces galaxies. Third, it correlates with ionized hydrogen, which is the most important gas for star formation. That second point is important because it helps put an upper limit on the completeness of an HI galaxy survey. The vast majority of galaxies emit a detectable HI signal, so we can expect that a high-sensitivity HI survey will show a representative sample of galaxies.

The 21 cm signal is due to the quantum properties of the spin-flip event of neutral

hydrogen. Neutral hydrogen contains a proton and an electron, which are both spin 1/2 particles. There are four possible linear combinations of spin alignments. 3 of those states are degenerate and have an energy higher than the ground state by 5.874 μeV [section 5.2 of [29]]. A emitted photon would have a wavelength of about 21 cm. The spin flip event is rare, but the 21 cm signal is visible from galaxies because of the large number of neutral hydrogen atoms. The brightness of the HI line is proportional to the number of atoms, so we know the total mass of HI gas in a galaxy (HI mass) if we measure its 21 cm signal.

There are two approaches to studying extragalactic 21 cm signals. One can either select a small sample of known galaxies and integrate them for a long time with a high-resolution telescope, or search over large solid angles of the sky to record every instance of a 21 cm spike. This thesis will be considering implementation strategy for the second approach and scientific questions that it answers. The first approach gives gas distribution maps of each galaxy, and thus their integrated spectra and rotation curves. It also allows direct searching for tidal disruptions and collisions by showing gas streams and gas clustering. This approach can only be applied to galaxies nearby enough to be imaged at high resolution. The second approach gives less information about each individual galaxy but has more statistical power. It does not require high resolution, so the entire galaxy's surface is integrated into a single pixel which makes detection easier. These surveys can go to a larger distance and include dimmer galaxies, so their sample size is larger. The only information recorded about each galaxy is its rough position and its integrated spectrum. The following section introduces science questions that this approach can effectively address.

Each galaxy's HI spectrum contains information that can we can use for science without outside experiments. First, a galaxy's internal gas velocity is reflected in the spectrum. Gas dispersion (usually Gaussian distributed) results in a broadening of the 21 cm spike because of Doppler shift. An overall side-on rotational velocity can be seen in a spectrum for the same reason. Measuring the amount of gas that is rotating compared to the gas dispersion allows you to guess the morphology of the galaxy. Spiral galaxies oriented the right way will have more gas with a high rotational velocity than an elliptical galaxy. Second, in the galaxy's frame, the average emission will be at 21 cm, so we can deduce its recessional velocity (equivalently its redshift) from the center of the peak. Thirdly, the total HI mass of the galaxy can be found by integrating over the spectrum for total HI brightness.

1.2 Science Questions

The HI content of galaxies is a useful probe into their evolution. This section explores more precise questions about galaxy evolution and how they can be answered by a large statistical survey of galaxy HI.

1.2.1 Gas Depletion and Accumulation Mechanisms

Galaxies can be found alone, in groups, or in clusters. The difference between groups and clusters is a soft split based on number density reflective of how number density qualitatively impacts galaxies' histories. Groups are in the range of about 100-1000 times denser than the field, and clusters are more dense than that cutoff. Groups and clusters are gravitationally bound, sometimes only temporarily in the case of groups. Nearby clusters like Virgo and Coma have been studied to learn about galaxy interactions. Galaxies in these clusters have been observed to have deficiencies in gas compared to galaxies in the field. We have found several mechanisms that could explain how gas in galaxies could be affected by group or cluster interactions.

One of the biggest difficulties of studying groups and clusters is determining relative positions and velocities. This issue is sidestepped by a mass statistical approach since we can use large numbers to counteract large uncertainties. Our approach provides no tools for estimating the difference in radial position between two gravitational bound galaxies, but an HI survey in combination with external position and velocity data could answer questions that require them. The following is a list of mechanisms that affect gas content that can be studied with the statistical approach.

Ram pressure stripping is the pressure of the Intracluster Medium (ICM) displacing cold gas that would otherwise remain gravitationally bound to a galaxy [12]. A disk-shaped galaxy that passes through the intracluster medium experiences a pressure that threatens to overcome the gravitational force binding its gas. The condition on the ICM density for ram pressure stripping to be effective is

$$\rho_{ICM} > \frac{2\pi G \Sigma_\star \Sigma_{ISM}}{V_\perp^2} \tag{1.1}$$

where ρ_{ICM} is the density of ICM, $\Sigma_\star$ is the mass density of the disk, Σ_{ISM} is the mass density of the interstellar gas, and $V_\perp$ is the galaxy's velocity perpendicular to its face [23]. The strength of ram pressure stripping is proportional to this velocity component.

This is difficult to measure, and our catalogue does not provide any cluster-scale velocity information. Ram pressure stripping is also proportional to the density of the ICM. Galaxy clusters are more dense near their center, so we expect the most ram pressure stripping to be observed near the center of clusters. Indeed, there is an abundance of gas-stripped dead galaxies documented in clusters. Our catalogue can be used to statistically measure ram pressure stripping. In the center of clusters, galaxies are more densely packed, so we expect an inverse correlation between the separation of galaxies and the amount of stripping.

Tidal stripping occurs when a massive objects passes by a galaxy with a low impact parameter. The side closest and farthest from the object experience different gravitational attractions, so the material inside the galaxy will separate. It can result in stellar streams, globular clusters, and gas tails. A basic estimation of the radius of a galaxy that is vulnerable to tidal stripping is

$$r_t = \left(\frac{m/M}{(3 + m/M)}\right)^{1/3} R \tag{1.2}$$

where r_t is the tidal radius, m is the mass of the galaxy, M is the mass of the object and R is the impact parameter [23]. Tidal stripping is more impactful with lower velocities, so small groups should be where the effect is most pronounced [14]. The Coma cluster has many chances of collisions, but Coma has a low fraction of spiral galaxies and their velocities are high, there is likely a low amount of tidal stripping. Our catalogue could be used to statistically investigate the environments where tidal stripping is most pronounced. The catalogue would include features like HI tails, but it could link amount of stripping and morphology of galaxies to their positions within groups and clusters.

1.2.2 HI Mass Function

The HI mass function relates the mass of HI gas in galaxies to their abundance. Galaxy formation models predict that galaxies with larger gas contents are more rare. Measuring the HI mass function to lower masses challenges formation models of smaller galaxies. One of the most commonplace galaxy abundance models is the Press-Schechter formalism

$$dn(L) = \phi \, dL = \phi^* \left(\frac{L}{L^*}\right)^{\alpha} \mathrm{e}^{-L/L^*} d\left(\frac{L}{L^*}\right)$$

written out with mass expressed as luminosity L. L^* is a luminosity cutoff and ϕ^* is an overall count. This formalism is piecewise composed of an exponential decrease in the number of galaxies with respect to mass followed by a sharp drop-off after some special

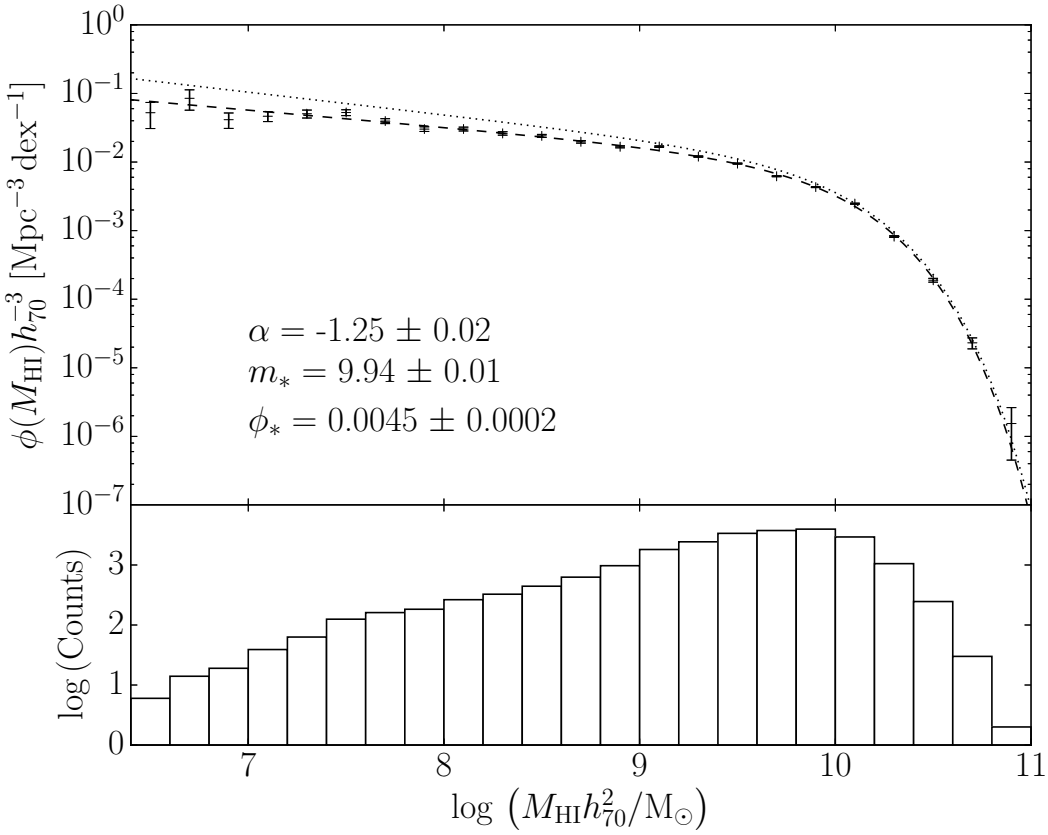


Figure 1.1: This is figure 2 in [18]. The dashed line represents the best fit Schechter function (with the parameters listed on the plot). The dotted line is the best fit Schechter function that was found using 40% of the full ALFALFA catalogue [16].

mass. This function was fit to the full ALFALFA catalogue to produce figure 1.1 [18]. Adding more samples alters the HI mass function prediction on the low mass side. The error bars on the first few points are large. Detecting more HI galaxies in this mass range and lower will help confirm or reject the Press-Schechter formalism for HI mass.

1.2.3 Optically Dark Galaxies

Team members working on the ALFALFA survey (introduced in section 1.3) tried to match entries in their HI galaxies catalogue to optical counterparts in the Palomar Digital Sky Survey or Sloan Digital Sky Survey [16]. It is known that as total galaxy mass decreases,

the fraction of stellar mass to gas mass drops below 1 [11]. The lower limit of this fraction is unknown. "Optically dark" objects emit almost no optical-wavelength light. Before ALFALFA, some of these objects were known, but they were all associated with some other optically bright object, for example as tidal debris fields. The ALFALFA team found 6.4% of their $\alpha.40$ catalogue could not be matched to an optical galaxy. 80% of those are suspected to be Milky Way clouds. Other groups have matched many of the remainder optically dark candidates to very faint extragalactic objects in optical surveys. In 2024, 50 candidates remained before a group matched 8 of them to galaxies in DECaLS [10]. It has been proposed that some of the remainders are "dark galaxies" which contain dark matter and optically invisible gas, but for some reason no star formation. If we find more HI sources with no optical counterparts, that would provide more evidence for the existence of such galaxies. It would also allow for better studying their probability of occurring and the environments where they occur. These galaxies would be more common as low-mass objects, so a survey that can detect fainter sources would be beneficial for this science case.

1.2.4 Stacking Analyses

Another use for an HI galaxy catalogue would be as a basis for a stacking analysis. A stacking analysis is used to reveal a weak trend or bias statistically by summing data vectors, usually with some weighting. An HI galaxy catalogue can be connected to data from another source, for example optical or x-ray, and then the emissions of each galaxy are stacked weighted by the HI mass recorded in the catalogue. This is a way to detect tidal stripping and morphological trauma as a function of HI content.

1.3 Existing HI Galaxy Experiments

21 cm astronomy has existed for almost a century, but it was originally limited to Milky Way observations. Starting in the 60s, large enough radio telescopes were built to be sensitive to extragalactic sources. Drift-scan telescopes collected data over large sky fractions and produced catalogues of galaxies, which have been improved upon with more and more ambitious surveys. The Very Large Array was completed in 1981, beginning array-based extragalactic radio astronomy which could leverage interferometry to produce high-resolution maps of hydrogen in nearby galaxies.

A useful metric to compare 21 cm telescopes is System Equivalent Flux Density (SEFD). This quantity estimates the sensitivity of the telescope, and later we can use it to predict

the number of sources that can be detected as a function of redshift. Radio astronomy often talks about the incoming signal in terms of temperature, which is the signal's power scaled by the Boltzmann constant k_b. This is to match the units of receiver noise which are already in temperature. SEFD is defined as the flux density of a radio source that adds 1 σ of temperature to the system temperature. The equation for it is

$$\text{SEFD} = \frac{2k_b T_{sys}}{\eta_A N_{dish} \pi r^2}.$$

This equation is general for both single-dish instruments and interferometers. T_{sys} is the system temperature, which is affected by the dish's structure and wiring, and environment like the sky and ground. This variable is more detailed than the comparisons that I am about to make require. N_{dish} is the number of dishes in the system, and r is the physical radius of each dish. η_A accounts for the effective area of the dish based on the dish's curvature. The important takeaway is that telescopes with more or larger dishes are more sensitive.

From 1997-2000, the HI Parkes All Sky Survey (HIPASS) [4] survey was carried out with the single-dish 64m Parkes Telescope. This created the first large catalogue of HI galaxies, containing 5000 galaxy detections. The Parkes Telescope could be moved in order to survey 30000 square degrees of the sky, but its relatively low sensitivity meant that it could not detect low mass galaxies.

The The HI Nearby Galaxy Survey (THINGS) [31] survey was conducted between 2003 and 2005 using the Very Large Array (VLA) [6]. The VLA is a 27 dish irregular array that has an effective baseline of 130m, which allows for high-resolution images. THINGS is not a large statistical survey, but rather a focused view of 34 nearby galaxies. It maps regions of HI gas inside each galaxy at a resolution of 6". It also has high spectral resolution which tells us the gas's radial velocity, which can be integrated into a 1D spectrum. THINGS shows that the spectra split into single peaks and two-horn profiles. Large galaxies like spirals with large edge-on rotational velocity have two-horned profiles because their gas is bimodally split into travelling away and toward us. Smaller elliptical galaxies tend to have Gaussian profiles because their gas is diffusion dominated.

The Arecibo Legacy Fast ALFA Survey (ALFALFA) survey ran from 2005 to 2011 using the Arecibo single dish 300m telescope. It produced a catalogue of 31500 galaxies. This is the largest out of previous-generation surveys, so it will be the main point of comparison for this thesis. The Arecibo telescope has a large collecting area, so it was sensitive to faint HI sources. The ALFALFA team employed a source detection algorithm that is similar in concept to the one this thesis will present. Using the data from the ALFALFA catalogue,

they were able to produce a well sampled HI mass function in the mass range 10^6 to 10^{11} $M_\odot$ (see fig 1.1). The HI masses from ALFALFA have also been used in cluster gas deficiency analysis as comparisons from the field [24].

The following table summarizes existing (and one near future) catalogues of HI galaxies.

Name	Year	Telescope	Number of Galaxies	Number of dishes
HIPASS [4]	1997-2000	Parkes	5000	1
ALFALFA [15]	2005-2011	Arecibo	31500	1
FASHI [34]	2020-2023	FAST	41741	1
WALLABY [21]	Future	SKA	500000 (proj)	$>$ 130000

Name	Dish Size (m)	Hemisphere	Area (deg^2)
HIPASS [4]	64	Southern	30000
ALFALFA [15]	300	Northern	6920
FASHI [34]	500	Northern	7600
WALLABY [21]	Varying	Southern	-

1.4 Neutral Hydrogen in Galaxies

These previous HI experiments provide a holistic depiction of neutral hydrogen in galaxies on both an individual and statistical level. THINGS's high resolution maps of the hydrogen in nearby galaxies give case examples of neutral hydrogen density and velocity distribution in several galaxy classes. The rotation curves computed from the THINGS measurements are useful for modelling spectrum templates that will be used in chapter 3. Matching the detections from previous large HI catalogues to optical galaxy databases reveals which classes of galaxies are easier to detect in HI. This section will be an overview about what observations about neutral hydrogen in galaxies can tell us about a future survey. We will follow one example spiral galaxy, NGC 6946.

Figure 1.2 includes an integrated HI map of NGC 6946 measured by THINGS [31]. Comparing against the optical image that is also included shows that the neutral hydrogen traces the stellar shape. Figure 1.3 shows the HI density of NGC 6946 as a function of radius. It has lower HI brightness in the very center, some dense patches in the outer bulge region, and then the density falls off at the arms. De Blok et al. [8] analyzed surface brightness distributions from flat THINGS galaxies and found that an exponential disk

describes the outer disk. Occasionally another central component is also necessary. The exponential disk profile has the form

$$I(R) = \frac{L}{2\pi R_d^2} \exp(-R/R_d) \tag{1.3}$$

where L is the total luminosity and R_d is the exponential scale length [23]. Note that the distribution in figure 1.3 does not match the exponential disk, likely because it does not correct for inclination or extinction. The HI map also shows multiple HI holes in NGC 6946. The cause of these holes is likely regions with high supernova activity [5].

The bottom left panel in figure 1.2 shows the velocity field of NGC 6946. The lighter greyscale indicates approaching emission and the darker greyscale indicates receding emission. This shows that the velocity in the galaxy is dominated by an overall rotation. Figure 1.4 from de Blok et al. [8] shows three mostly agreeing rotation curves for NGC 6946. After an initial increasing segment at low radius, they level off, which is consistent with a constant-density-core dark matter halo. This rotation curve coupled with the HI density profile causes the shape of the spectrum profile templates that will be used in the search. (See section 3.1.) There is a well observed positive correlation between the rotational velocities of spiral galaxies and their luminosities. Since there is also a positive correlation between galaxies' stellar luminosity and HI gas, then this relationship can be extended to predict that galaxies with higher HI mass will also tend to have higher rotational velocities.

Also shown in figure 1.2 is the velocity dispersion of NGC 6946. If a galaxy is spinning edge-on, then the gas velocity will be dominated by the rotation component, but it also has a random dispersion component. Kamphuis and Sancisi [19] measure a 15 km/s dispersion in the inner region of NGC 6946 and 9 km/s in the outer region. In both the Milky Way and NGC 6946, a high velocity HI gas component has been detected, including high velocity clouds which move in contrast to the bulk motion of the gas. This gas contributes to wider wings in HI spectra.

NGC 6946 is a late-type luminous galaxy, which is just one class of galaxies present in an HI galaxy catalogue. Galaxies with magnitude around and above -18 are called dwarf galaxies [23]. They are the most numerous type, but are harder to detect because they are gas poor and have less rotational velocity to exploit. The surface brightness profiles of dwarf galaxies are exponential. Another way to sort galaxies is by the Hubble sequence. It divides galaxies into early-type (ellipticals) and late-type (barred and unbarred spirals) with lenticular galaxies in between. An HI galaxy survey is biased toward detecting gas-rich blue late-type galaxies over gas-poor red early type galaxies. The HI distribution of spiral galaxies is similar to the NGC 6946 example. Morganti et al. [25] performed deep

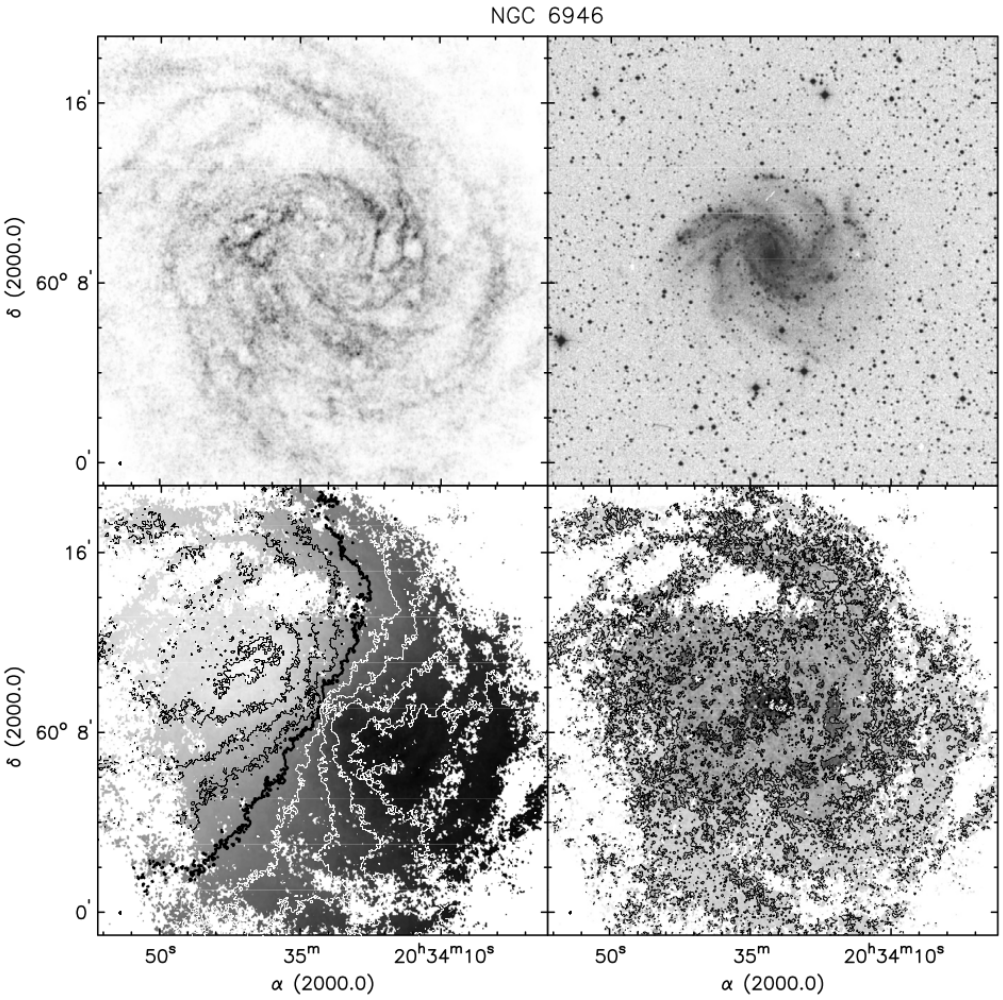


Figure 1.2: Figure 65 from Walter et al. [31]. Top left: Integrated HI map of NGC 6946. Top right: Optical image from Digitized Sky Survey (DSS). Bottom left: Velocity field with the color scale indicating approaching or receding velocity. Bottom right: Velocity dispersion map.

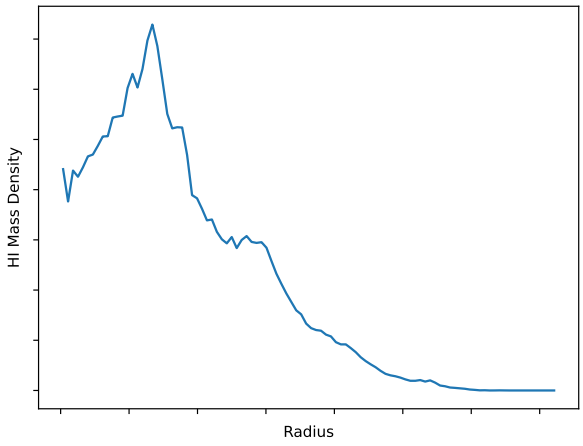


Figure 1.3: Plotting the surface brightness profile from the top left of figure 1.2. This plot does not include any corrections. Length units require a distance estimate to the galaxy and are not necessary for the illustration, so are omitted. The scales of both axes are linear.

observations on early type galaxies with the Westerbork Synthesis Radio Telescope and found that HI gas is often present, albeit fainter. They found that the HI distribution and kinematics in this class of galaxies is varied.

1.5 CHORD Introduction

The Canadian Hydrogen Observatory and Radio-transient Detector (CHORD) [30] is an upcoming experiment designed to run alongside its predecessor CHIME [3] starting in 2024 and ramping up to full functionality in 2025. It is a drift-scan maximally redundant rectangular-array interferometer. It will consist of 512 6-meter dishes arranged in a 22×24 grid with gaps. It is located at the Dominion Radio Astronomy Observatory near Penticton, B.C., which has a latitude of 49.322 degrees north. From that point, CHORD can be rotated up to 30 degrees north or south, but the rotation process is difficult and requires taking the telescope offline for long periods. Its frequency detection range is 300-1500 MHz, which the 21 cm line fits into at 1428 MHz. CHORD has several main science goals: hydrogen intensity mapping, Fast Radio Burst detection, pulsar observations, and

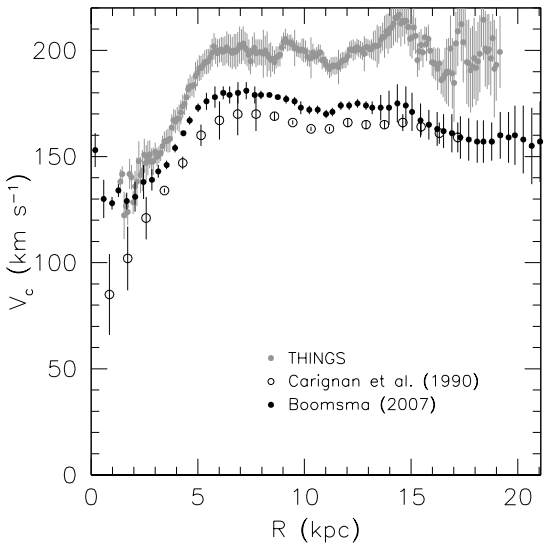


Figure 1.4: Figure 18 from de Blok et al. [8] This is the THINGS rotation curve for NGC 6946. Also plotted is other analysis of the same rotation curve.

galaxy neutral hydrogen measuring and cataloguing. The latter is the subject of this thesis. CHORD's specifications optimize it to cataloguing a large number of galaxies whose positions are known to within about 5 arcminutes, but are not spatially resolved.

CHORD is an interferometer. An interferometer is a type of telescope that measures the amplitude of electromagnetic waves as a function of time in multiple locations and uses the spatial and time information to reconstruct the direction of the wave. The first step of this is done by the correlator computer; it multiplies the voltages recorded by the i, j pairs of dishes to produce a power: [7]

$$P_{ij\ cos} = \frac{V_i V_j}{R} \cos\left(\frac{2\pi i c \tau_g}{\lambda}\right). \tag{1.4}$$

The resistance R is present for dimensional clarity, but in practice the voltages are scaled arbitrarily. These values are saved per wavelength λ. τ_g is some amount of time the

correlator waits before averaging. This information is combined with the sine version to create a quantity called the visibilities V_{ij}:

$$V_{ij} = |P_{ij}| \left(\cos\left(\frac{2\pi i c \tau_g}{\lambda}\right) - \sin\left(\frac{2\pi i c \tau_g}{\lambda}\right)\right) = |P_{ij}| \exp\left(-\frac{2\pi i c \tau_g}{\lambda}\right) \tag{1.5}$$

where c is the speed of light. This equation to compute the visibilities comes from the following definition, which writes the visibilities as a function of power from the sky $I(\vec{u})$ via the interferometer equation:

$$V_{jk} = \int_{\vec{u} \in sky} I(\vec{u}) \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j - \vec{x}_k) \cdot \vec{u}\right) \, d\Omega \tag{1.6}$$

where $\vec{x}_i$ is the position of dish i as a 3 vector parallel to the plane of the array. The last important quantity is another way to synthesize visibility information called the dirty map $d(\vec{u})$. It appears as a map on the sky as a function of direction $\vec{u}$:

$$d(\vec{u}) = \sum_j \sum_k V_{jk} \exp\left(-\frac{2\pi i}{\lambda}(\vec{x}_j - \vec{x}_k) \cdot \vec{u}\right) \tag{1.7}$$

This is the same as taking the Fourier transform of the power of each dish pair. The distances and angles between dish pairs are called baselines, the lengths of which determine important properties of the interferometer. The smallest baseline corresponds to the largest angular distances that can be probed at once. The largest baseline corresponds to the angular distance below which everything is smoothed by the resolution of the telescope. CHORD's shortest baseline is 6.3 m in the east-west direction and 8.5 m in the north-south direction, corresponding to a beam size of 1.9 degrees and 1.4 degrees at a wavelength of 21 cm. Its longest baseline is 139 m east-west and 204 m north-south, corresponding to a resolution of 5.2 arcmins and 3.5 arcmins at a wavelength of 21 cm respectively.

There are two different types of array interferometer layouts. Irregular arrays place dishes so that there are many different baselines between them. Regular arrays place all dishes on a repeating grid so that many baselines are identical. Irregular arrays allow you to measure a greater number of Fourier modes corresponding to the Fourier transforms of the greater number of baselines. CHORD is a maximally redundant array, laid out in a rectangular grid. This means it has the greatest number of repeated baselines for the number of antennas. The advantage is that repeated measurements reduce the noise, which improves detection probability. Our main goal is to detect the largest number of galaxies, so this array layout suits our science case.

CHORD is a transit telescope. Its dishes cannot be realigned in the Hour Angle (HA) direction. A transit telescope takes advantage of the Earth's rotation, recording data on a strip of the sky once per day. At a wavelength of 21 cm, CHORD is most sensitive within 1.4 degrees north and south of where it is pointing, so the nightly data product will be a strip about 3 degrees wide.

The CHORD array has a larger scale than previous radio astronomy experiments in part due to advancements in GPU technology. CHORD data will first be handled in compute huts on location right next to the dishes. The galaxies science case has its own compute pipeline that includes computing the visibilities and Radio Frequency Interference (RFI) excision. These steps require collecting data over varying time intervals, with the longest time interval being on the order of 30 seconds. After that, the data is sent in the form of a dirty map to an offsite compute node where my code can be run on it.

Since CHORD is still in the pre-first-light phase, the team is still making decisions regarding computing strategy as well as observation strategy. The decisions relevant to the galaxies science case are the cadences of various steps of the galaxies GPU pipeline, the upchannelization parameter (U) (see section 3.4), and which RFI flagging kernels to include. For observation strategy, CHORD can rotate north-south between declinations of +20 to +80, but rotating requires manually moving all 512 dishes which will take about a week. We can choose how long to keep CHORD pointed at a single strip of sky before moving it. Part of my project is to estimate how much moving CHORD will help with galaxy identification, and where it should be moved to.

CHORD's deployment will be gradual, with dishes being produced and turned on one at a time. An interferometer is technically functional with any number of dishes, so it was decided that once CHORD reached 64 (laid out as 6×11) dishes, it will start collecting data. This milestone is expected to be reached in early 2025. This proto-CHORD is called Pathfinder, and it can be used to test the galaxy search software and our detection predictions. Pathfinder will have much weaker resolution and collecting area, but new dishes will be installed and immediately start contributing to the observations. All of the software that I have written is general enough to be able to assume Pathfinder's 6×11 layout as well as CHORD's 22x24 layout.

1.5.1 CHORD Sensitivity Analysis

Many of the science questions can be better answered with an instrument that can detect either more or fainter galaxies. In this subsection, I will check whether CHORD will present

a significant improvement in sensitivity and try to estimate the expected galaxy count after CHORD has been running for multiple years.

I wrote a Python script that estimates the number density of HI objects using the ALFALFA Schechter function.[1]

The Schechter luminosity function is:

$$dn(L) = \phi \, dL = \phi^* \left(\frac{L}{L^*}\right)^{\alpha} \mathrm{e}^{-L/L^*} d\left(\frac{L}{L^*}\right)$$

From this ALFALFA paper [18], a good fit was $\alpha = -1.25$, $\phi^* = 4.5 \times 10^{-3}$ Mpc^{-3}, $\log(M_*/M_\odot) = 9.94 \implies M_* = 10^{9.94} M_\odot$. The standard formula to relate the HI mass and flux density (not accounting for HI self-absorption) is

$$\frac{M_{21}}{M_\odot} = 2.356 \times 10^5 D_{Mpc}^2 S_{21} \Delta V \text{ (Jy km/s Mpc}^2)^{-1}$$

And the definition of spectral luminosity is

$$L_\nu = 4\pi d^2 S_\nu$$

Combining these equations gives

$$L_{21} = 4\pi \frac{M_{HI}}{(2.356 \times 10^5 \text{ M}_\odot)(\Delta V)} \text{ Jy km/s Mpc}^2 \tag{1.8}$$

ΔV is the bandwidth of ALFALFA in km/s. The bandwidth in Hz is 2.4414×10^4. They can be converted with $\Delta V = \frac{c^2}{\lambda_0 \Delta \nu}$ to give 58.514 km/s.

Plugging in M_* from the ALFALFA paper gives $L^* = 7.939 \times 10^3$ Jy Mpc^2.

Integrating over luminosity will give us the number density of galaxies. The luminosity integral needs to start at some cutoff $L_{cut}(z)$ which corresponds to the luminosity of a galaxy that we can no longer detect at that redshift. To find this cutoff, we need CHORD's SEFD.

"SEFD is the system equivalent flux density (Jy), defined as the flux density of a radio source that doubles the system temperature. Lower values of the SEFD indicate more

[1] The code is hosted here https://github.com/hanshopkins/CHORD_21cm_galaxies_counts.

sensitive performance." as stated in [7] (ch 3.7.6 Sensitivity). It has units of Janskys. It is related to the thermal noise via:

$$\Delta I_m = \frac{SEFD}{\eta_c \sqrt{n_{pol} t_{int} \Delta\nu}}$$

where the number of polarizations n_{pol} is 2, η_c is the correlator efficiency assumed to be 1, t_{int} is the total integration time over a spot, and $\Delta\nu$ is the bandwidth. This paper [22] puts CHORD's SEFD at 12 Jy with the formula

$$\text{SEFD} = \frac{2kT_{sys}}{\eta_A N A_{phys}}$$

Where T_{sys} for CHORD is 30K, η_A is the aperture efficiency ≈ 0.5 for CHORD, N is the number of dishes (512 and 64 for full CHORD and Pathfinder), and A_{phys} is the physical area of the dishes (disks with 6m diameter).

We'll define the luminosity cutoff L_{cut} to be the luminosity such that $F/\sigma_T < \Delta I_m$, where σ_T is the minimum number of sigmas that you want for a detection. In terms of luminosity, this is

$$L_{cut} = F_{cut}(4\pi D^2) = \frac{\sigma_T (SEFD)}{\sqrt{n_{pol} t_{int} \Delta\nu}}(4\pi D^2)$$

Where D is the comoving distance. The goal is to find the number of detections for a redshift bin at redshift z with width dz. In the thin bin approximation, its number of galaxies is $n(\text{Vol(z)} - \text{Vol(z} - \text{dz}))$. This is the number of galaxies that we're expecting in 4π steradians, so we multiply by the fraction of the sky that we're expecting to cover.

The last step to predict the galaxy counts is to input an example scanning plan. The following plot was made assuming CHORD scanned 10000 deg^2 over the course of 24000 hours which is almost 3 years. t_{int} was calculated as the amount of time that the primary beam spends on average over a source over the 3 year duration. σ_T is usually taken to be 5.

Running the code yields figure 1.5. Integrating under the predicted ALFALFA curve gives 39,304 galaxies, and the final published catalogue has 31,502 galaxies, which indicates that this rough estimation is reasonable. Integrating the Pathfinder curve (which assumes that Pathfinder stays at 64 dishes and only gets around to a third of the sky that full CHORD will observe) gives 40,802 galaxies, and integrating the full CHORD curve gives 2,684,859 galaxies. This is about a two orders of magnitude improvement in detections,

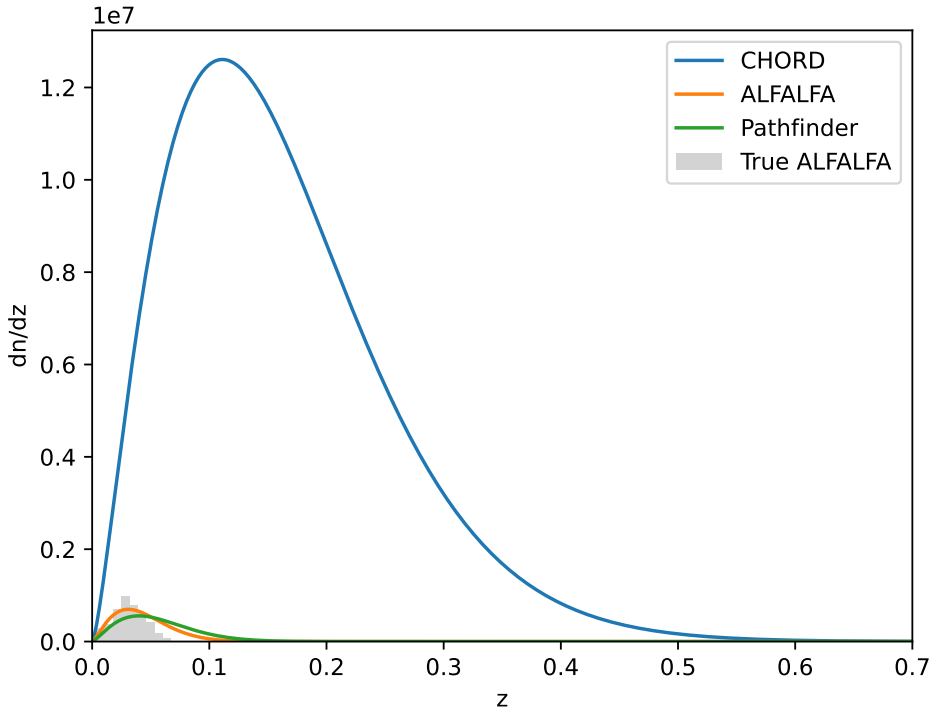


Figure 1.5: Following the method outlined in this subsection, the number of galaxies detected as a function of redshift is predicted for full CHORD, Pathfinder, and ALFALFA as a check. For easy comparison, a histogram of the published ALFALFA catalogue is included in grey.

which is significant for the science questions in the previous section. As a disclaimer, this estimation does not take into account RFI, which will block large frequency intervals. The exact problematic frequencies are being investigated. A major source of RFI that affected ALFALFA was the FAA radar at the San Juan airport [15], which was centered at 1350 MHz or about $z = 0.05$. This might help explain the discrepancy between my counts estimation and the actual ALFALFA histogram past $z = 0.05$.

1.6 Method

As seen in the science questions section, the most helpful quality of a large galaxy HI catalogue to improve is the detection threshold. It allows for detecting farther or fainter galaxies. We saw in the CHORD section that CHORD is capable of delivering this lower detection threshold. The rest of the thesis is about my work on the software side to minimize the detection threshold while also balancing for computational cost.

The CHORD galaxies team has members from McGill, Queens, and Waterloo. The short-term goal of the team is to predict our hardware and software ability to detect galaxies. This includes predicting the problems caused by spatial and frequency aliasing. We accomplish this by adapting simulation and data processing software from CHIME. We must also come up with technical and observing requirements to pass on to other CHORD teams. The galaxies team's long-term objective is to produce a large catalogue of galaxy spectra. We will also fit profiles to the spectra to try to extract redshifts and profile parameters. The detection and spectrum isolation pipeline will involve existing software developed for the intensity mapping science case as well as our own matched filtering software.

My personal role in the galaxies team is developing the matched filter algorithm that returns a list of detections and predicting the performance of this algorithm. This task is split into two parts. First, I wrote a one-dimensional matched filter that acts on galaxy spectra. Second, I wrote a filter that acts on the CHORD visibilities for a fixed frequency to spatially pick out point sources. Finally, the full matched filter combines these two parts according to the matched filter formula. For the frequency step, I also wrote software that analytically predicts the result of the matched filter picking up a point source, including spatial aliasing. This allows me to test the effects of changing observing parameters like CHORD's declination. The result is code that can be used to make recommendations for CHORD's design and observing strategy and that can eventually find galaxies when CHORD data starts coming in.

The spatial part of the matched filter is a weighting operation that acts on the dirty map of a single frequency. It was designed to take advantage of existing code that computes the visibilities for Intensity Mapping purposes. The most important consideration for the spatial part is aliasing, which takes the form of a regular grid of candidate locations of a source. My analytic prediction software of the spatial matched filter calculates curvature effects and the primary beam's passage over the sky to see how much alias locations are ruled out. Aliases are much easier to distinguish if data from separate CHORD runs at different declinations are combined. My analytic matched filter can calculate the result of dithering CHORD by arbitrary amounts several times and combining the data from all runs. I convert the correlation coefficients calculated analytically into likelihoods of alias confusion. This allows me to directly predict the consequences of picking a scan strategy that increases or decreases aliasing.

For the frequency part of the matched filter, the core design is a 1-dimensional Fast Fourier Transform (FFT)-based matched filter. This design is usually for finding a signal in time-series data, like LIGO's search for black hole mergers. The matched filter equation suggests that it should be an n^2 algorithm, but the FFT implements it at $n \log n$, requiring only the assumption that the template does not change with respect to its position in the vector. This assumption is violated in CHORD's case because of the effects of upchannelization, so an alteration needs to be made to the matched filter algorithm to preserve its speed. In this section I will also discuss the choice of templates to balance between detection rate and computational time.

The desired end product is a catalogue listing all detected galaxies, their spectra, and fitting parameters like their redshifts, widths, and shapes. In discussion section 4.1, I will explain different options for combining the frequency and spatial parts into a single program. Adding more steps to a matched filter usually increases its ability to match with low Signal to Noise Ratio (SNR) targets at the cost of multiplicatively increasing computation time. This detection step is followed up by flagging already-detected galaxies and a more comprehensive fitting step in order to fill the catalogue.

1.7 Previous Detection Algorithms

The galaxy search pipeline that ALFALFA used [26] is also split into spatial and frequency steps, but the spatial component is handled differently. "Data cubes" with 2 spatial dimensions 2.4 degrees across and 1024 frequency channels are set up. Cubes for the two polarizations are stored separately. First, one dimensional matched filters (see section 1.8) are run for each spatial pixel. They are run using several templates from a mono-parameter

family. The designers found that thinner galaxies were better approximated by Gaussians, and the wider functions were more like a top-hat shape. Also, wider galaxies were more likely to have horns. The designers took advantage of this to create a family of profiles that morphs from a thin Gaussian to a combination of top-hat and Hermite functions, which gives a wide peak with horns shape. This family only having one variable parameter greatly reduces the number of profiles that needed to be checked compared to a higher dimensional parameter space. If the 1D profile search passes a detection threshold, then those voxels as well as surrounding voxels in the data cube are flagged as a detection in order to avoid counting multiple detections for the same galaxy. The flags are then filtered by a series of criteria known about galaxies, for example the signal is expected to be unpolarized, so any candidates that appear too different in the two polarization data cubes are rejected. After this process, the positive spectra are isolated so that more costly fitting algorithms can be run on them later. This overall design is very similar to CHORD's, with the main difference being extra care to avoid false detections caused by spatial or frequency aliasing.

The pipeline developed for WALLABY (for SKA) is called SOFIA [27][33]. Its relevant steps are its filtering and source finding, and it has multiple options for both steps. For filtering, it has the option of convolving with a 3D kernel, which is computationally alike to a matched filter. It includes another option called 2D–1D wavelet de-noising, which appears to be similar to a high pass filter. After the filtering step, one of 3 source-finding algorithms is used. The first flags detections for any regions of the data cube that pass above a threshold. This is the planned method for the CHORD search. The other two methods operate at multiple scales, which is advantageous for a high-resolution instrument, but it is not as useful for CHORD. The second source-finding method is called Smooth + Clip (S+C) which smooths using a 3D kernel multiple times at different scales. For each scale, any voxels above a brightness threshold are flagged, and the output is a union of these masks. This method helps make sure the masking scale matched the detected galaxy scale. The third source-finding method is called CNHI, which acts on a single 1D spectrum, or a small combination of spectra from nearby pixels. It uses the Kuiper test to check if that pixel's spectrum is consistent with random noise, and flags the ones that aren't. SOFIA will use similar strategies to the ones I will describe in this thesis, but only the simple threshold source-finding algorithm should be necessary for CHORD.

1.8 General Theory of Matched Filtering

This section will introduce matched filtering in full generality. This is the technique that will be used for making detections, and the specific implementations and added assumptions

will be given in later sections relative to this base.

Matched filtering is a search method that is useful when you have a precise template that needs to be picked out from noisy data. Given a vector d that may contain a linearly scaled version of a template vector $t(x)$, plus zero-mean Gaussian noise of known variance (and covariance), the matched filter M_x is defined as

$$M_x = \frac{t_x^\dagger N^{-1} d}{\sqrt{t_x^\dagger N^{-1} t_x}}. \tag{1.9}$$

x is a set of possibly continuous index labels. If M_x is large, then the label x "matches" the data. N is the covariance matrix. Dagger is conjugate transpose.

The matched filter gives a signal to noise ratio for the hypothesis comparison between the data being produced by random correlated Gaussian noise and random correlated Gaussian noise plus a scaled version of the expected template. See the proof in appendix A. The numerator checks the closeness between the template and the data. The denominator normalizes the result to be in signal-to-noise units. In practice, this fraction is computed for a large number of templates covering the parameter space of the template model, and the best matching x label is returned.

The matched filter is a linear filter, meaning that it is a linear operation on the data vector. The matched filter is the optimal linear filter assuming additive stochastic noise, which means that any other linear filter would achieve a lower maximum at the correct x value.

A quantity that we are interested in is the correlation coefficient $R_{x'x}$:

$$R_{x'x} = \frac{M_x(d)}{M_{x'}(d)} = \frac{t_x^\dagger N^{-1} t_{x'}}{\sqrt{t_x^\dagger N^{-1} t_x}\sqrt{t_{x'}^\dagger N^{-1} t_{x'}}} \tag{1.10}$$

This is normalizing the matched filter so that the largest value (corresponding to the case where the template matches the data exactly, i.e. $t_{x'} = d$) is 1. It can be interpreted as the correlation between the matched filter picking up the signal at the correct location x' and another location x. This will be useful for interpreting problematic locations (aliases) that the matched filter may pick up as the true location.

Chapter 2

Spatial Matched Filter

This chapter is about the portion of the matched filter that searches for the spatial location of the galaxies using array interferometry. This section will not be considering the effects of multiple frequency channels, which is a good approximation for many smaller galaxies. I will introduce beam synthesis, and then use that to present my analytical approximation program of the final spatial matched filter. I'll present predictions that this program makes about spatial aliasing.

2.1 Spatial Introduction

Most galaxies in the sky will have an extent smaller than CHORD's ∼5 arcminute resolution at 21 cm. In addition, many galaxies will occupy only a few of CHORD's about 40 km/s frequency channels. Searching for galaxies will require different methods than CHORD's hydrogen intensity mapping which relies on the assumption of slow spatial variation as well as different methods than CHORD's fast radio burst search which can take advantage of a predictable sweep across many frequency channels. Likewise, previous galaxy HI observations have been done with single-dish telescopes or irregular arrays, so CHORD's highly regular array aliasing problem is new. CHORD will have the advantage of more effective scanning time, which results in a fainter detection limit, so it is worth trying to overcome this problem.

The spatial part of the search relies on synthesizing the signals from all 512 dishes into a single map. As an redundant-baseline interferometer, CHORD correlates the arriving radio waves between dishes. The phase offsets between the electric fields measured by the

dishes correspond to sky directions that the wave originates from. For each pair of dishes with the same orientation and separation this is a repeated measurement with independent noise, so a regular array decreases the noise relative to the signal. The CHORD correlator will compute these correlations, called complex visibilities, for each pair of dishes. For one dish at coordinate j in the array and a second dish at coordinate k, their complex visibility V_{jk} is equation 1.6:

$$V_{jk} = \int_{\vec{u} \in sky} I(\vec{u}) \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j - \vec{x}_k) \cdot \vec{u}\right) \, d\Omega$$

where $\vec{x}_i$ is the ith position in the array represented as a 3 vector and λ is some fixed wavelength. In this section, we'll be interested in the case where $I(\vec{u})$ is a delta function representing a single point source at $\vec{u}_s$. The integral evaluates to

$$V_{jk} = I_s \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j - \vec{x}_k) \cdot \vec{u}_s\right) \tag{2.1}$$

"Beamforming" is synthesizing the visibility data into a more human-readable form that depends on spatial coordinates $\vec{u}$ using equation 1.7. There are several common ways of combining all the dish pairs, and the method that optimizes this repeated measurement advantage is called the natural weighting, which treats each dish pair equally, resulting in the smallest and most common baselines being most important. Another popular option is to fix the importance of all baselines by upweighting the less common ones, for example the huge diagonal across the entire array. For this analysis, I will proceed with the former option. If we apply it onto the single source example, we get

$$d(\vec{u}) = \sum_j \sum_k I_s \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j - \vec{x}_k) \cdot (-\vec{u} + \vec{u}_s)\right)$$

This is a map of the sky without any post-processing or cleaning, so it is called the "dirty map". It is important to note that these calculations are done before the start of the galaxy search, so we can take advantage of them without an extra time cost.

I_s is a product of the brightness of the source and a scaling factor that represents how far away the source is from the center of the dishes; the farther away it is, the less power the dishes will pick up. This factor is the "primary beam" function B^2 that determines the signal loss of a source off-axis from the pointing of the dish $\vec{u}_p$, $B\left(|\vec{u}_s - \vec{u}_p|\right)^2$. This is a decreasing function of the angular difference. This factor sets the width of the strip of

sky that CHORD can cover in a scan. A model that I will be using for CHORD's primary beam function is an Airy beam:

$$B^2(\alpha) = \left(\frac{J_1(\tilde{\alpha})}{\tilde{\alpha}}\right)^2; \quad \tilde{\alpha} = \frac{\pi D \sin(\alpha)}{\lambda} \tag{2.2}$$

where D is dish size (6m for CHORD).

The equation for the dirty map shows a central problem of searching for point sources with a redundant-baseline interferometer, which is that d will have the same shape if $\vec{u}_s$ is replaced with other specific vectors. This is to say arbitrary pairs of $\vec{u}$s will have the same relative values, so the function d will only differ by an overall scaling factor. The multiplicative factor is affected by B, but redundantly with the brightness of the source. This phenomenon is called aliasing. If $d(\vec{u})$ is plotted, bright spots will appear not only around $\vec{u}_s$, but also around those other vectors. The bright spots around the other vectors are called aliases. I will explain in more detail later how to locate aliases, but the following is an example of one alias for understanding.

To demonstrate, consider a pair of two east-west neighboring dishes 1 and 2 that are pointing straight up, and there is a single source in the sky, so equation 1.7 is valid. Defining the unit vectors $\hat{v}_{EW}$, $\hat{v}_{NS}$ and $\hat{v}_{up}$ as the array coordinates in the east-west, north-south, and up directions respectively. In this example, $\vec{x}_2 - \vec{x}_1 = L\hat{v}_{EW}$ where L is the dish separation. It's easy to see that $d(\vec{u}_s)$ evaluates to I_s. There is another position east of CHORD's zenith $\vec{u}_a$ which will produce the same value. Set $\vec{u}_a = \vec{u}_s - \frac{\lambda}{L}\hat{v}_{EW}$. Inserting this u_a into the expression for d gives $2\pi i$ inside the exponential, so the expression also evaluates to I_s. I_s can either be the product of a galaxy intensity $\overline{I}_s$ and $B(|\vec{u}_s - \vec{u}_p|)^2$ or a brighter intensity $\overline{I}_a$ and $B(|\vec{u}_a - \vec{u}_p|)^2$. An observer at this stage has no way of knowing which of these two situations is true.

The problems aliases can cause are threefold. First, after detecting a galaxy in the visibility matrix, an algorithm might infer that the galaxy is at an aliased position rather than the true source position. It will be shown later (sec 2.4) that the perfect alias degeneracy can be broken, but some conditions will cause aliased sources to have correlation coefficients high enough that an unlucky noise fluctuation might cause an aliased position to appear brighter than the correct position. Second, a dim source in the alias position of a bright source might be drowned out by the contribution from the bright source. If galaxies are relatively sparse in the spatial-and-frequency space, then this problem will be rare. Thirdly, there might be false positive detections because of aliases from sources outside of the scanning area. CHORD is a drift-scan telescope, so we will observe strips of the sky at a time. If a source is outside of the strip but close enough that its aliases are inside of the strip, those aliases might be identified as real sources.

The alias positions depends on the frequency of the source, so for broadband targets like fast radio bursts, there is no alias degeneracy. This is a problem specifically for the galaxy search, which have narrow-band targets. The main purpose of this section is to investigate the conditions where the the aliasing degeneracy can be broken in the narrow-band target case. I will present a tool that can simulate the output of a matched filter running on CHORD data as it drifts across a single point source. The tool is called the Analytic Matched Filter (AMF). This will allow us to quantify the relative signal of the true source location and alias positions. This tool performs a similar calculation to the M-mode mapmaker (see section 4.2) for CHORD. The tool is much faster than the M-mode mapmaker because it takes advantage of the assumption of a single point source, no noise, and a rectangular array, allowing for an easier analysis of aliasing in different situations.

The target of the matched filter, and so the desired template by construction, is a point source emitter. This is easier to model than an extended source, and the vast majority of galaxies that we are trying to find will be smaller than CHORD's resolution. If the matched filter uses CHORD visibilities as the data, then the template should be equation 2.1:

$$t_{ujk} = B\left(|\vec{u} - \vec{u}_p|\right)^2 \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j - \vec{x}_k)\cdot\vec{u}\right) \tag{2.3}$$

This template is the visibilities of an emitter located at $\vec{u}$. The results of a matched filter do not depend on the overall scale of the template, so the source brightness is omitted. Only the primary beam part of I_s is kept in order to match the definition of the dirty map. We can plug this into the matched filter equation (eqn 1.9)

$$M_u = \frac{\mathrm{tr}({t_u}^\dagger N^{-1} V)}{\sqrt{\mathrm{tr}(t_u^\dagger N^{-1} t_u)}}. \tag{2.4}$$

N in this case is the covariance matrix of the visibilities, set by the noise of the dishes. The matrix operations are done over the j, k dish indices.[1] The numerator of this equation is performed by the galaxies pipeline using existing code from the intensity mapping science case. The only computation that remains is the normalization.

In the not-unrealistic assumption that the noise does not vary between dishes or dish pairs, the noise matrix becomes a constant σ times the identity $\mathbb{I}$, and the denominator

[1]The template t_u is also a complex matrix. Equation 2.4 is the same as equation 1.9 with the vectors wrapped into matrices. This matched filter equation will always be real because visibility matrices are Hermitian. Proof is in appendix B. Otherwise, we would have to take the real part of the numerator and denominator.

reduces to

$$\begin{aligned}
\text{denominator} &= \frac{1}{\sqrt{\sigma}}\sqrt{\mathrm{tr}(t_u^\dagger \mathbb{I} t_u)} \\
&= \frac{1}{\sqrt{\sigma}}\left(\sum_{j,k} B\left(|\vec{u}-\vec{u}_p|\right)^2 \exp\left(\frac{-2\pi i}{\lambda}(\vec{x}_j-\vec{x}_k)\cdot\vec{u}\right) B\left(|\vec{u}-\vec{u}_p|\right)^2 \right. \\
&\qquad\qquad \left. \times \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j-\vec{x}_k)\cdot\vec{u}\right)\right)^{1/2} \\
&= \frac{1}{\sqrt{\sigma}} B\left(|\vec{u}-\vec{u}_p|\right)^2 \sqrt{\sum_{j=1}^{m_1 m_2}\sum_{k=1}^{m_1 m_2} 1} \\
&= \frac{m_1 m_2}{\sqrt{\sigma}} B\left(|\vec{u}-\vec{u}_p|\right)^2
\end{aligned}$$

where m_i is the number of dishes on the ith axis, so the sums over the dish indices go to $m_1 m_2$, the total number of dishes.

The whole matched filter equation becomes

$$M_u = \frac{\mathrm{tr}(t_u{}^\dagger V)}{m_1 m_2 \sqrt{\sigma} B\left(|\vec{u}-\vec{u}_p|\right)^2}. \tag{2.5}$$

This equation displays the main function of the spatial matched filter; it takes the dirty map or image of the sky that CHORD produces, and it divides out by the dimming of the primary beam so that the remaining brightness represents only the signal to noise ratio, not how far away the source is from the dish pointing. This has the side effect that aliases which are suppressed in the dirty map by being far away from the primary beam now appear bright. After the division step, the matched filter finds the highest SNR points and flags them as sources, so there is a risk of flagging aliases as sources.

2.2 Analytic Matched Filter (AMF) Intro

In the following three subsections, I will derive an explicit equation for the matched filter for increasingly general situations. The first will be about what a rectangular array instrument's matched filter would return after a single instant of time. The next will show what happens when the instrument is allowed to drift across the sky and we combine data from

multiple snapshots. The last subsection will be about the instrument drifting for several days and changing the angle that it points to between days, called "dithering". These equations will be the basis for the numerical tool called the AMF.[2]

The following assumptions were made in order to simplify the equations and reduce compute time:

- Noise is constant across all dish pairs. I.e. $N \longrightarrow \sigma\mathbb{I}$.
- The array is perfectly rectangular. There are no missing dishes in the array, and the dish positions are all perfectly aligned. For my tests, I modelled CHORD as a 22×24 array, which has 16 more dishes than the real CHORD to fill in some gaps.
- The primary beam function is an Airy beam.
- There is exactly one point source emitter in the sky and no radio frequency interference.

2.3 Instantaneous Analysis

For the instantaneous case, we need to run the matched filter on the visibilities of a single source at one time. That's what we were doing already in the intro, so we can use equation 2.3. The AMF is assuming constant noise, so we can also use equation 2.5. For the AMF, the visibility data will be exactly the visibilities caused by a single source at some input position $\vec{u}_s$, so it is also given by equation 2.3. That is to say, in this idealization, the data matches the template perfectly. Last, instead of computing the matched filter M_u, we will instead compute the correlation coefficient R_{uu_s} whose form is similar to eqn 2.4 but normalized relative to the true template. Combining these equations gives the correlation

[2]The code for the AMF is hosted at https://github.com/hanshopkins/analytical_CHORD_HI_Galaxy_mf.

coefficient of the instantaneous case.

$$\begin{aligned}
R_{uu_s} &= \frac{\mathrm{tr}(t_u{}^\dagger V)}{m_1^2 m_2^2 B\left(|\vec{u}-\vec{u}_p|\right)^2 B\left(|\vec{u}_s-\vec{u}_p|\right)^2} \\
&= \frac{1}{m_1^2 m_2^2 B\left(|\vec{u}-\vec{u}_p|\right)^2 B\left(|\vec{u}_s-\vec{u}_p|\right)^2}\left(\sum_{j,k=1}^{m_1 m_2} B\left(|\vec{u}-\vec{u}_p|\right)^2 \exp\left(-\frac{2\pi i}{\lambda}(\vec{x}_j-\vec{x}_k)\cdot\vec{u}\right)\right. \\
&\qquad \left.\times\, B\left(|\vec{u}_s-\vec{u}_p|\right)^2 \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j-\vec{x}_k)\cdot\vec{u}_s\right)\right) \\
&= \frac{1}{m_1^2 m_2^2}\sum_{j=1}^{m_1 m_2}\sum_{k=1}^{m_1 m_2}\exp\left(\frac{2\pi i}{\lambda}(\vec{x}_j-\vec{x}_k)\cdot(-\vec{u}+\vec{u}_s)\right)
\end{aligned}$$

This computation would involve summing $m_1^2 m_2^2$ terms, which is time consuming. With the assumptions made about a rectangular array, there is an algebra trick that we can use in order to simplify this expression into a quicker to compute form. The following is the proof of this useful fact.

Useful fact:

$$\begin{aligned}
&\sin\Big(\frac{c}{2}\Big)\sum_{j=1}^{m}\exp\big(ic(\tfrac{-m-1}{2}+j)\big) \\
&\quad= \frac{1}{2i}\left(e^{ic/2}-e^{-ic/2}\right)\sum_{j=1}^{m}\exp\big(ic(\tfrac{-m-1}{2}+j)\big) \\
&\quad= \frac{1}{2i}\sum_{j=1}^{m}\exp\big(ic(-\tfrac{m}{2}+j)\big)-\exp\big(ic(-\tfrac{m}{2}+j-1)\big) \\
&\quad= \frac{1}{2i}\left(\exp\big(ic\tfrac{m}{2}\big)+\sum_{j=1}^{m-1}\exp\big(ic(-\tfrac{m}{2}+j)\big)-\exp\big(-ic\tfrac{m}{2}\big)-\sum_{j=2}^{m}\exp\big(ic(-\tfrac{m}{2}+j-1)\big)\right) \\
&\quad= \frac{1}{2i}\left(\exp\big(ic\tfrac{m}{2}\big)-\exp\big(-ic\tfrac{m}{2}\big)+\sum_{j=1}^{m-1}\exp\big(ic(-\tfrac{m}{2}+j)\big)-\sum_{j=1}^{m-1}\exp\big(ic(-\tfrac{m}{2}+j)\big)\right) \\
&\quad= \sin\Big(\frac{cm}{2}\Big)
\end{aligned}$$

$$\implies \sum_{j=1}^{m}\exp\big(ic(\tfrac{-m-1}{2}+j)\big)=\frac{\sin\big(\frac{cm}{2}\big)}{\sin\big(\frac{c}{2}\big)} \tag{2.6}$$

We'll get the correlation coefficient R_{uu_s} into a form where we can take advantage of this fact. For the sake of the proof, I'll split up the dish indexing system. Instead of each dish being labelled by a single number j or k, it will instead be labelled with two numbers representing its row and column j_1 and j_2. With this labelling system, the correlation coefficient is written as

$$R_{uu_s} = \frac{1}{m_1^2 m_2^2} \sum_{j_1,k_1}^{m_1} \sum_{j_2,k_2}^{m_2} \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_{j_1 j_2} - \vec{x}_{k_1 k_2}) \cdot (-\vec{u} + \vec{u}_s)\right)$$

Define $\hat{e}_1$, $\hat{e}_2$ as the two basis vectors of the rectangular array with $\hat{e}_\rho$ as the general version. Setting the origin of the coordinates at the 0,0th dish, the dish positions can be written explicitly as

$$\vec{x}_{\ell_1 \ell_2} = \left(\frac{-m_1 - 1}{2} + \ell_1\right) L_1 \hat{e}_1 + \left(\frac{-m_2 - 1}{2} + \ell_2\right) L_2 \hat{e}_2$$

This allows us to expand the exponential sum:

$$\begin{aligned} R_{uu_s} = \frac{1}{m_1^2 m_2^2} \sum_{j_1,k_1}^{m_1} \sum_{j_2,k_2}^{m_2} \exp\Bigg(\frac{2\pi i}{\lambda}\Bigg(&\left(\frac{-m_1 - 1}{2} + j_1\right) L_1 \hat{e}_1 + \left(\frac{-m_2 - 1}{2} + j_2\right) L_2 \hat{e}_2 \\ &- \left(\frac{-m_1 - 1}{2} + k_1\right) L_1 \hat{e}_1 - \left(\frac{-m_2 - 1}{2} + k_2\right) L_2 \hat{e}_2\Bigg) \cdot (-\vec{u} + \vec{u}_s)\Bigg) \\ = \frac{1}{m_1^2 m_2^2} E(0,1) E(0,2) E(1,1) E(1,2) \end{aligned}$$

Where we introduce the shorthand

$$E(\alpha, \rho) = \sum_{\ell=1}^{m_\rho} \exp\left((-1)^\alpha \frac{2\pi i}{\lambda}\left(\frac{-m_\rho - 1}{2} + \ell\right) L_\rho \hat{e}_\rho \cdot (-\vec{u} + \vec{u}_s)\right)$$

The sum symbols can be distributed between the exponential factors because they do not share the summing index. Now we can take advantage of the algebra fact eqn 2.6 with $c = (-1)^\alpha \frac{2\pi L_\rho}{\lambda} \hat{e}_\rho \cdot (-\vec{u} + \vec{u}_s)$ to rewrite this as

$$\begin{aligned} E(\alpha, \rho) &= \frac{\sin\left((-1)^\alpha \frac{\pi m_\rho L_\rho}{\lambda} \hat{e}_\rho \cdot (-\vec{u} + \vec{u}_s)\right)}{\sin\left((-1)^\alpha \frac{\pi L_\rho}{\lambda} \hat{e}_\rho \cdot (-\vec{u} + \vec{u}_s)\right)} \\ &= \frac{\sin\left(\frac{\pi m_\rho L_\rho}{\lambda} \hat{e}_\rho \cdot (-\vec{u} + \vec{u}_s)\right)}{\sin\left(\frac{\pi L_\rho}{\lambda} \hat{e}_\rho \cdot (-\vec{u} + \vec{u}_s)\right)} \end{aligned}$$

This is the step where the regular array assumption is required. Note that the sum has been removed completely, and $E(\alpha, \rho)$ actually does not depend on α. Substituting this result back in gives

$$R_{uu_s} = \frac{1}{m_1^2 m_2^2} \frac{\sin\left(\frac{\pi m_1 L_1}{\lambda} \hat{e}_1 \cdot (-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_1}{\lambda} \hat{e}_1 \cdot (-\vec{u} + \vec{u}_s)\right)^2} \frac{\sin\left(\frac{\pi m_2 L_2}{\lambda} \hat{e}_2 \cdot (-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_2}{\lambda} \hat{e}_2 \cdot (-\vec{u} + \vec{u}_s)\right)^2} \tag{2.7}$$

This is much quicker to compute than the original form. It's quick enough that we can compute it for a large grid of $\vec{u}$ directions and make a tangent plane plot. The tangent plane plot is a useful projection method because it is defined everywhere on the sphere. To plot some function defined on a sphere, choosing some basis point on the sphere, each pixel on the tangent plane plot is the function value of the direction that intersects with the tangent plane. We can use this tangent plane plot to confirm that perfect aliases exist in the instantaneous case.

Before proceeding, there is one more detail about the baseline L_1: I present the baseline L_1 as the distance between dishes in the north-south direction. This is only true if the dishes are pointing straight upward. CHORD must be calibrated as though all antennae are in the same plane, that is to say it's as if the dishes are sitting on a wedge perpendicular to the direction that the dishes are pointing in. This has the effect that when CHORD tilts, its north-south baseline shrinks because in the plane that CHORD is pointing towards, the dishes are closer together.

In fig 2.1, aliases appear on a rectangular grid in tangent-plane coordinates corresponding to a grid of east-west and north-south vectors projected onto the sky. They are separated by about 1.5 degrees. The alias grid spacing can be derived on paper from equation 2.7. Since the sine is squared and m_ρ is an integer, the fraction will evaluate to m_ρ whenever

$$\frac{L_\rho}{\lambda} \hat{e}_\rho \cdot (-\vec{u} + \vec{u}_s)) \in \mathbb{Z}$$

This indicates that the alias separation is scaled by $\frac{\lambda}{L_\rho}$. This estimate is confirmed numerically in fig 2.8.

For a more heuristic explanation, we expect the alias correlation coefficients to be 1 because a redundant-baseline array cannot distinguish brightness coming from those aliased locations from brightness of the true source. We expect sources farther away from the primary beam to produce a dimmer response, but this dimming factor is cancelled out by the denominator of the matched filter equation. This is so that potential sources that are

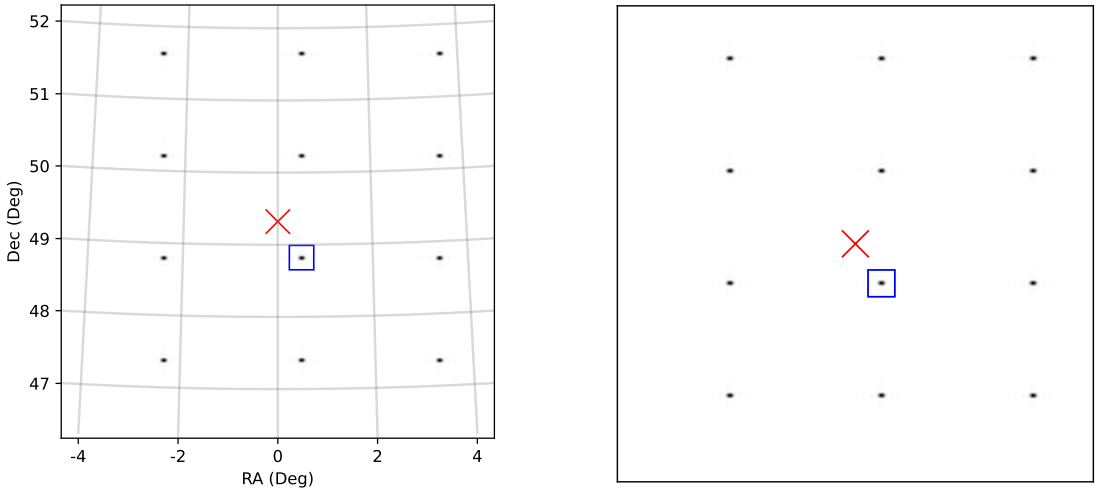


Figure 2.1: A tangent plane plot plotting equation 2.7, the correlation coefficient R_{uu_s}, with and without gridlines. The direction that CHORD is pointing in $\vec{u}_p$ is indicated with the ✕. It is pointed at its actual zenith. The source position u_s is indicated by the box ☐. The wavelength is set to 21 cm. A grid of locations yield $R_{uu_s} = 1$, which are perfect aliases.

farther away from the primary beam are weighted properly. In other words, the matched filter knows that the primary beam will attenuate sources far from CHORD's pointing direction, so it can simply compensate by assuming that the source is much brighter. It is important to note that the correlation coefficient map is a map of the response of the matched filter, and not a map of what the sources would look like on the sky. Sources at distant alias positions would have to be much brighter, but the matched filter already corrects for that attenuation.

It's clear that a search for galaxies using a matched filter on a single visibility matrix snapshot will suffer badly from alias indistinguishability. In the next section we will show that adding in information from other time samples can break this degeneracy.

2.4 Integrated Analysis

CHORD's galaxy pipeline has various stages which will operate at different cadences. The important cadences to know are the rate that the visibility matrices are taken in, which is on the order of milliseconds, and the rate that output visibility matrices are finally saved to disk, which is on the order of 30 seconds. This final step is so long because of RFI removal requirements as well as storage and bandwidth costs. The matched filter will run on multiple snapshots at once in order to take advantage of all the time the source spends in CHORD's field of view.

If we want to search for a signal in multiple snapshots at once, the matched filter template must be adapted to include a time index. This can be done by rotating all sky vectors relative to the direction CHORD is pointing at each point in time. We'll choose coordinates where CHORD is always stationary, and the sky is moving. That is to say $\vec{u}_p$ stays constant and other $\vec{u}$s do not. If $\mathcal{R}(\phi)$ is the rotation matrix with respect to Earth's spin axis, and ω is the rotational velocity of the Earth, then a template with a time index τ is

$$t_u = \exp\left(\frac{2\pi i}{\lambda}(\vec{x}_{j_1 j_2} - \vec{x}_{k_1 k_2}) \cdot \mathcal{R}(\omega\tau)\vec{u}\right) \; B\left(|\mathcal{R}(\omega\tau)\vec{u} - \vec{u}_p|\right)^2 \tag{2.8}$$

This template includes information about the increase and decrease of power as the source passes over the primary beam, as well as the path that the source travels on across the sky, which is a line of constant declination.

Placing this template into the matched filter equation (eqn 1.9) reveals that the nu-

merator and denominator do not cancel as much. In particular, the primary beam terms.

$$R_{uu_s} = \frac{\sum_\tau B\left(|\mathcal{R}(\omega\tau)\vec{u} - \vec{u}_p|\right)^2 B\left(|\mathcal{R}(\omega\tau)\vec{u}_s - \vec{u}_p|\right)^2}{\left(\sum_\tau m_1^2 m_2^2\right)\sqrt{\sum_\tau B\left(|\mathcal{R}(\omega\tau)\vec{u} - \vec{u}_p|\right)^4}\sqrt{\sum_\tau B\left(|\mathcal{R}(\omega\tau)\vec{u}_s - \vec{u}_p|\right)^4}} \times \frac{\sin\left(\frac{\pi m_1 L_1}{\lambda}\hat{e}_1 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_1}{\lambda}\hat{e}_1 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2} \frac{\sin\left(\frac{\pi m_2 L_2}{\lambda}\hat{e}_2 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_2}{\lambda}\hat{e}_2 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2} \quad (2.9)$$

For a single time step, there are always completely degenerate aliases. If the template includes multiple time steps, then there is an opportunity to break these degeneracies. Two effects contribute to this degeneracy breaking: one is based on how the primary beam affects the brightness of the alias as a function of time; the other is based on the location of the alias due to the curvature of the sky. These two effects can be seen in the two different places $\mathcal{R}(\omega\tau)$ is introduced in equation 2.9. I'll explain these effects in detail in the following paragraphs.

The first effect is that the true source and the alias template see different parts of the primary beam curve. This effect applies to all alias positions. The matched filter expects to see the strongest signal when the source passes over the primary beam; when the source is far from the primary beam, the template is suppressed. For aliased source positions, the matched filter expects the signal to peak at the wrong time, so although the individual time steps have the right shapes, their relative amplitudes are mis-matched with the observed signal. This mismatch is corrected for in the single time step case, but when there are multiple time steps being summed over, the mismatch causes the aliased position to lose signal.

This effect can also be seen in equation 2.9. Ignoring the sine fraction part, the correlation coefficient looks like the correlation between the alias's primary beam value and the source's primary beam value. It is 1 if $\vec{u} = \vec{u}_s$, and less than 1 otherwise. Figure 2.2 shows how this correlation would drop below 1 for two aliases: one to the east of the main source and one to the north.

Correlating the east alias B^2 to the true source B^2 would yield a low result since they peak at different times. This effect helps with aliases that have a different RA than the true source much more than sources that have the same RA. For example, this plot also shows the B^2 if the alias directly in the north. This correlation would be close to 1, but the north alias's fringes arrive too soon and are relatively higher. The second effect can be used in conjunction with this one to differentiate the north alias.

The second effect that helps to reduce aliasing is the difference in the alias travel path compared to the path of a real source. Real sources follow paths with constant declination,

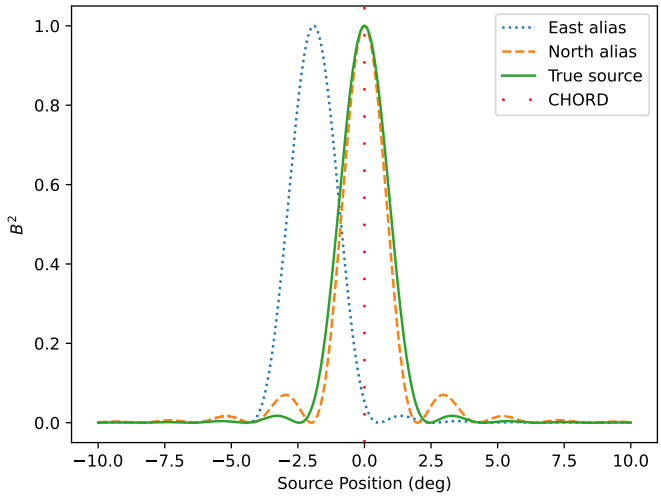


Figure 2.2: This plot shows the difference in response when aliases cross over the meridian. Since the source is moving across the sky at a constant rate, the x-axis is also time. When the real source is 2 degrees away from CHORD, the matched filter for the east alias is expecting a spike in brightness that isn't there. An alias with the same RA as the true source but north by one grid space will experience the spike at the correct time, but since the primary beam function is not separable, its fringes do not line up properly.

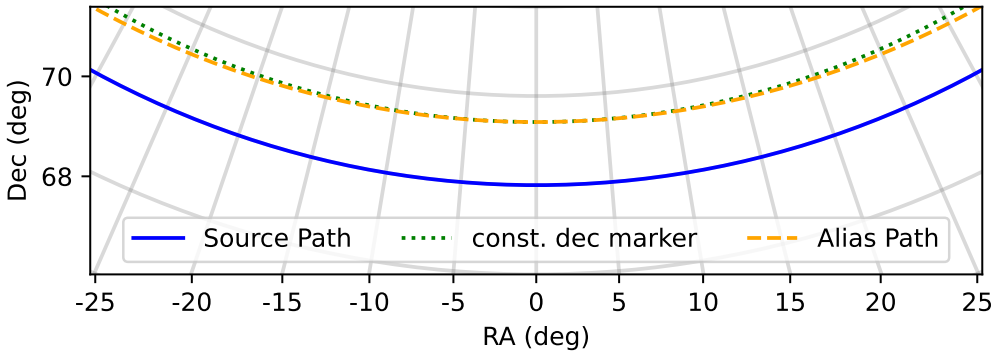


Figure 2.3: This plot shows the deviation of an alias's path to a path of constant declination. In this example, CHORD is pointed at a high 70 degree declination, and there is a source directly on top of it. The position of the first northern alias is computed at 0 RA. If the alias was a true source, it would follow a path of constant declination marked by the dotted line. Instead, in this tangent plane projection, it follows a path that is the source's path shifted upwards, shown by the dashed line.

but aliases are fixed relative to the position of the real source, so their path has the same curvature as the path of the real source. Considering an alias with a declination 2 degrees different than the source, the difference in curvature between these two paths is measurable. Templates that are placed where an alias is therefore do not follow the same path as the alias and so do not pick up all of the power of the alias at each time step. This effect is stronger at the poles and weaker at the equator since it relies on curvature. In the flat sky approximation, this effect is nonexistent. The difference in paths can be seen in figure 2.3.

This effect can be seen mathematically by comparing the analysis at the end of section 2.3 with equation 2.9. For any individual term in the sum, it is possible to find a $\vec{u}_a$ other than $\vec{u}_s$ which causes the sine fractions to evaluate to $m_1^2 m_2^2$. This $\vec{u}_a$ will depend on the relative positions of CHORD and the source, so a viable $\vec{u}_a$ for one time step will no longer work in other time steps. This causes the matched filter to drop below 1 after summing over multiple time steps.

Figure 2.3 shows that the difference in paths is very small, especially over CHORD's primary beam (a few degrees around HA=0), which is the the most important area. The difference in the matched filter is only noticeable in accumulation over many time steps. This effect also works in conjunction with the previous effect by adding to the distortion of the primary beam crossing curves that can be seen in figure 2.2.

2.4.1 Safety Parameter Sweep

The goal of this section will be to compute the correlation coefficient of aliases in different parts of parameter space in order to judge whether they will be a problem. In this section, I will show one example each for two parameters which are not interesting. First, I will demonstrate that the wavelength of the source does not have a qualitative impact on the subsequent analysis. I will also show that as long as the time sampling rate is sufficiently high, it does not impact the analysis at all.

The wavelength of the source is related to its redshift by $\lambda = (z + 1)(21 \text{ cm})$. The wavelength appears several times in 2.9, but to linear order its effect mostly cancels. In figure 1.5, we see that most detected galaxies will have a redshift of about 0.1, so their wavelength will only differ from 21 cm by about 10%. It can be shown numerically that this 10% wavelength difference will not have unpredictable impacts on the alias correlation coefficients.

Figure 2.4 shows that it's not necessary to thoroughly investigate the effects of redshift for the alias analysis. Over the redshifts we care about, a typical alias might change its strength by a couple percent. Even in the more extreme cases, there's no unpredictable fluctuations. The alias separations scale with wavelength λ, so it is expected that effects like curvature will be exacerbated with higher wavelength. All further plots are done with $\lambda = 21\text{cm}$.

The second parameter to get out of the way is the time separation between snapshots, which I call the cadence or $\Delta\tau$. In order to see the curvature effects of figure 2.3 or the primary beam correlation in figure 2.2, we need enough points to sample the curves. We can test numerically how often we need to sample one full rotation of the Earth to reach convergence.

Figure 2.5 shows that once a full day is split at least 500 times, the AMF has converged. CHORD's actual sampling rate is much higher than this requirement, even with upchannelization (see section 3.4), so this test only is only used to confirm that I am sampling finely enough.

2.4.2 Alias Analysis

To show what the matched filter would return taking into consideration time integration, we can remake figure 2.1, which shows a typical case. Figure 2.6 shows numerically that the aliases in the instantaneous case are no longer perfect when considering multiple time samples. The aliases still form nearly the same grid pattern with the same spacing, but

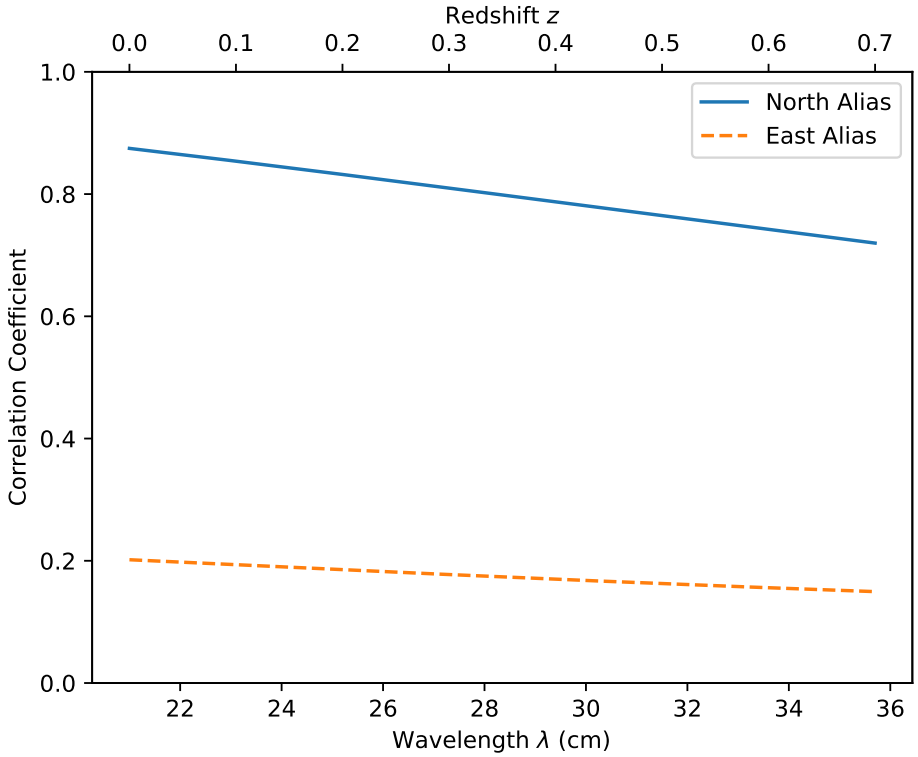


Figure 2.4: This is an example situation showing the stability of the alias correlation coefficients when varying the wavelength. For this example, I placed CHORD at zenith and a source directly overhead. The correlation coefficient for the first alias to the east and the first alias north are shown. The north alias's correlation coefficient dropped by 0.02 by redshift 0.1, and dropped by 0.15 by the very high redshift 0.7. The east alias dropped by 0.008 and 0.05 by redshift 0.1 and 0.7 respectively.

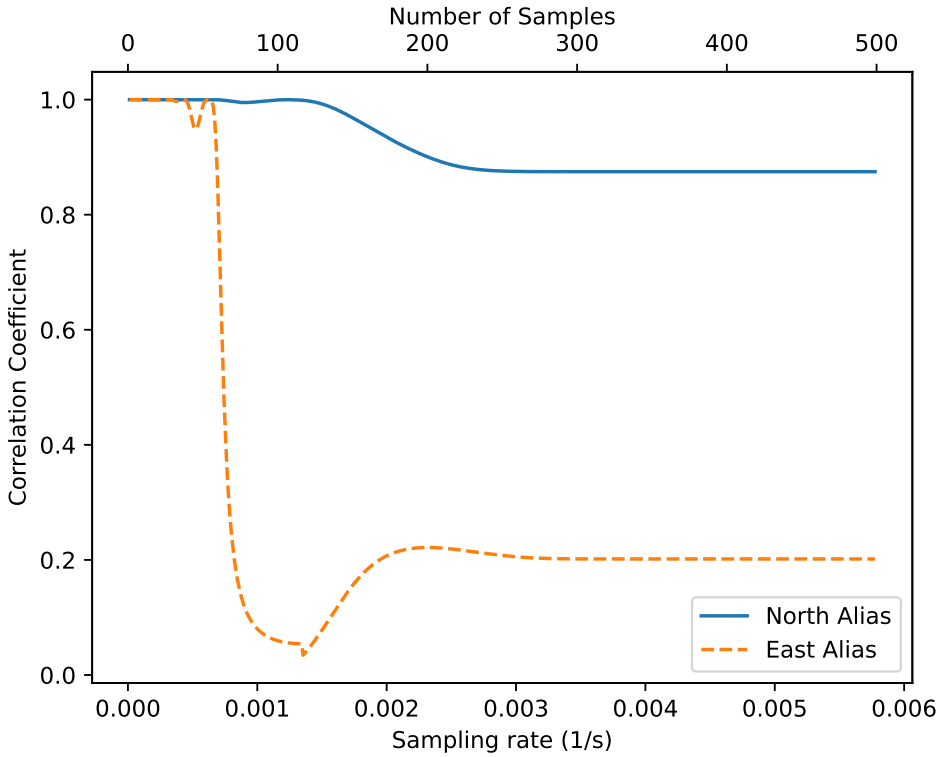


Figure 2.5: Checking that I am running the AMF over fine enough time samples to simulate CHORD's near-continuous sampling. For this example, I placed CHORD at zenith and a source directly overhead. The correlation coefficient for the first alias to the east and the first alias north are shown. Section 2.3 showed that in the case of only 1 time sample, both aliases have a correlation coefficient of 1.0. The effects discussed in section 2.4 break the degeneracy and decrease the correlation coefficients with enough time samples taken.

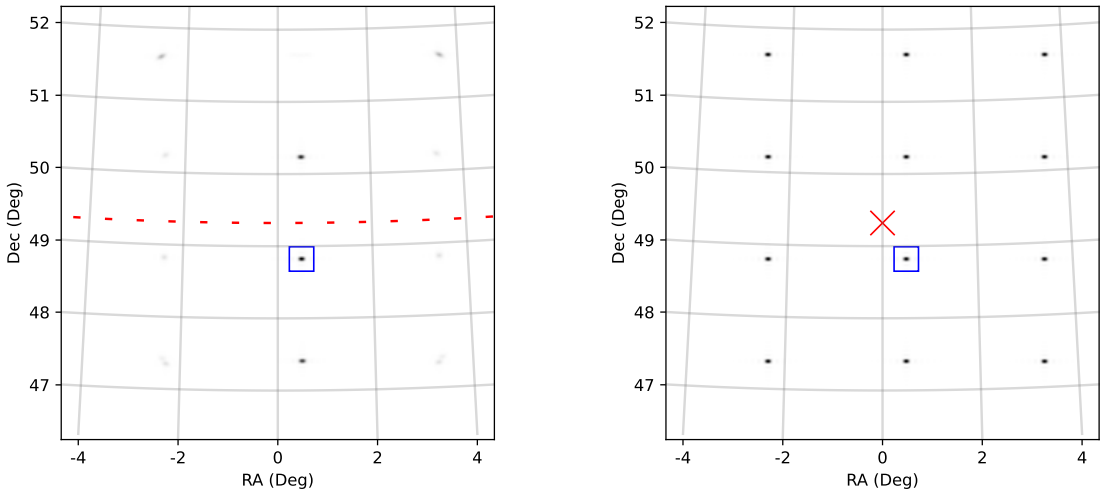


Figure 2.6: Left side: The correlation coefficients given by equation 2.9 for a typical declination and source position. In the integrated case, the sky and CHORD move relative to each other, so CHORD's declination is indicated with a dashed line. The position of the true source is indicated by the box □. Right side: a copy of figure 2.1, the instantaneous case, for comparison. The east-west aliases are largely eliminated, as are all but the first north-south aliases.

they are suppressed to various degrees. The two effects discussed in the introduction of this subsection can be seen. The east and west aliases are highly suppressed because of the primary beam crossing effect. The north and south aliases are affected less by the primary beam crossing effect, and the curvature effect is less significant. The aliases can also change shape, which can be seen with the two aliases in the upper corners. There is no longer a symmetry that guarantees that they all look identical.

We are interested in how this picture changes depending on CHORD's declination and the source position. Figure 2.7 shows an exaggerated scenario where CHORD is pointed only one degree away from the north pole, which will not happen in any real observation. Being so close to the north pole makes the effect of curvature much more pronounced. The alias paths deviate significantly from the path of a source that was in the alias's position, which causes the alias power to be spread out in rings. The alias power being spread out means that the likelihood of any individual alias position being misidentified as a true source is greatly decreased. Pointing more north makes alias disambiguation easier because

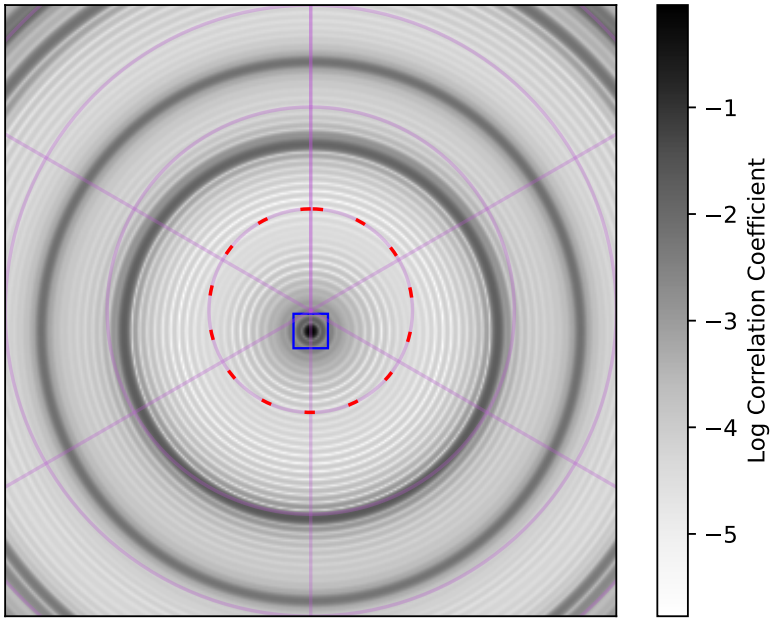


Figure 2.7: A map of the log base 10 correlation coefficients for the case of a source near the north pole. The position of the source is indicated by the box □. The purple declination gridlines are separated by 1 degree. CHORD's declination at 99 degrees is indicated by the dashed line.

the curvature effect is stronger. A ringing pattern can be seen around the source which is caused by the sinc-like shape of the primary beam.

I will now move away from showing maps to focus on the correlation coefficients of the nearest aliases to the source. The analytic matched filter prediction tool is useful because it can predict the correlation coefficients of aliases that might cause confusion. This section will be investigating the severity of nearby aliases depending on the scan strategy of CHORD. To do this, we need to estimate the positions of the nearby aliases well enough that we can compute their correlation coefficients without having to run the full grid of template positions. Figure 2.8 shows this being done successfully for an example north alias.

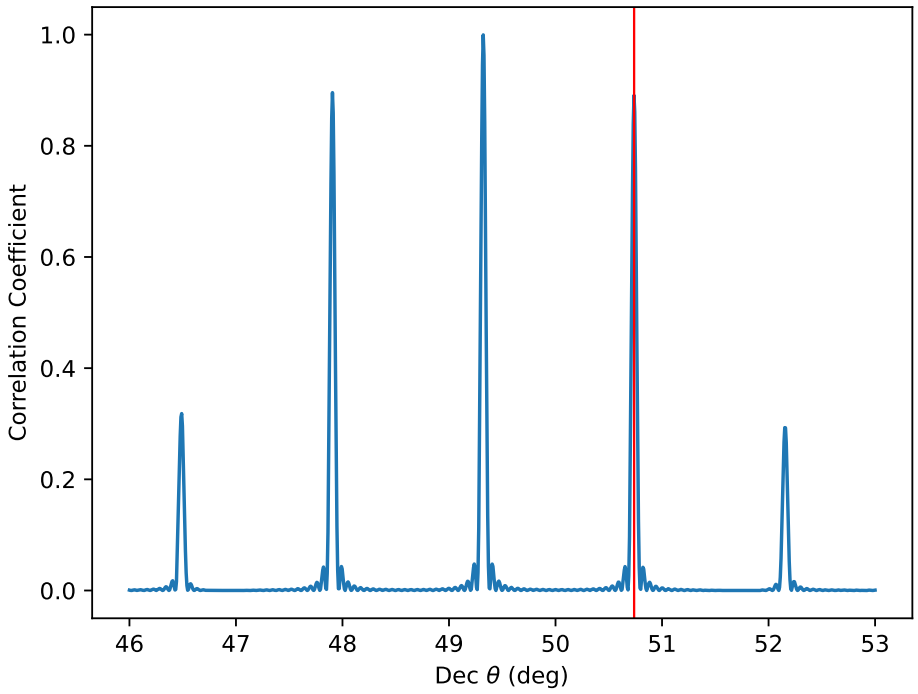


Figure 2.8: This plot shows how the analytic matched filter can be used to predict the locations of aliases. It shows the matched filter response on a one-dimensional north-south slice over CHORD's pointing with a true source directly above CHORD at declination 49.3. The straight vertical line is the estimation of the alias's position explained in section 2.3. It matches the desired position of the first northward alias.

The north and south aliases are easier to find because of a symmetry argument. If there is continuous time sampling, then there is complete east-west symmetry. The no-noise assumption implies that every time sample is weighted equally, so the only difference in swapping the east and west directions is a different ordering of the terms in the matched filter sum along the time index. This fact allows us to restrict the search for the north and south aliases to the same RA as the source. It cannot be slightly east or west without breaking symmetry. Once we know the approximate location of the alias, we can perform a 1-dimensional grid search to find the alias peak.

This fact cannot help with the search for the east and west aliases. For them, it is necessary to perform a small 2-dimensional peak search. For all aliases, the first step to finding them is to determine their exact location in the instantaneous case to use as a first guess. This can be done with the method in appendix C. The method can be used to approximate the locations of the east or west aliases, but it assumes a fixed position for CHORD, which is only valid in the instantaneous case. The integrated east-west aliases are then pinpointed by an iterative grid-based search. Figure 2.9 shows that the four nearest aliases can be located.

With the four most important aliases identified, we can now quantify the matched filter response of the aliases in different scenarios. The following plots explore the predictions depending on CHORD's positioning. The parameters being tested are CHORD's declination, the relative positions of the source and CHORD, and dithering. We explore these options by simulating the same base case, CHORD pointing at zenith with a source about half a degree away, and then vary one parameter at a time.

The first configuration tested is fixing CHORD at its zenith, and then moving the source away from it by an angular distance $\delta\theta$. The results are plotted in figure 2.10. As a general trend, alias disambiguation becomes easier the farther away it is because the primary beam crossing effect is stronger. The curves are not monotonic because of the non-monotonic shape of the primary beam. Alias disambiguation being easier if the source is farther away does not mean that we should aim CHORD such that the targets are on the very edge of the strip. CHORD's sensitivity drops off near the edge of the strip, so targets should still be placed in the center of the strip in order to optimize detection probability. The exception would be if the target's detection is guaranteed and the only priority is minimizing alias confusion.

The next result in figure 2.11 is about the direction that CHORD is aimed in. CHORD can point anywhere from a Dec of +20 degrees to +80 degrees. Because curvature is more important near the poles compared to near the equator, the hypothesis is that alias disambiguation should be easier as CHORD is pointed more northward. The east alias

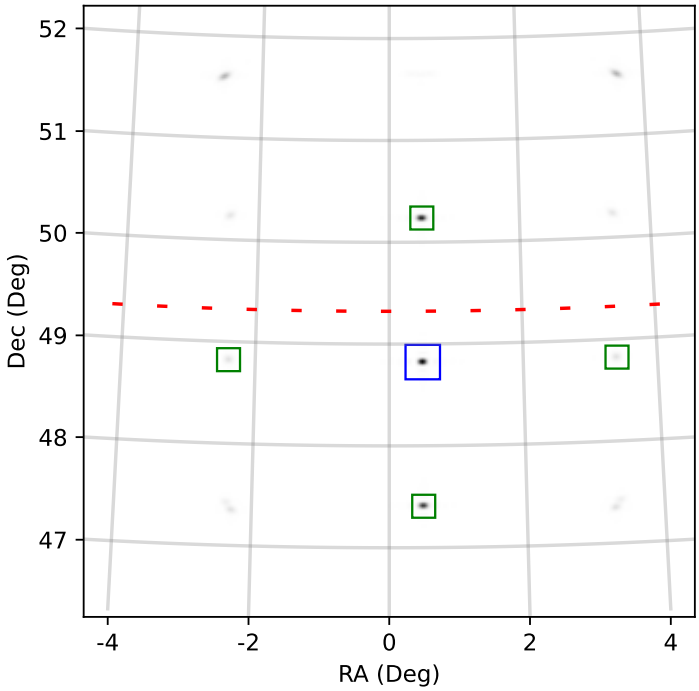


Figure 2.9: The four nearby aliases are automatically found, and their peak correlation coefficients are returned. This is the same situation as plot 2.6, but with the four nearby aliases marked with boxes □.

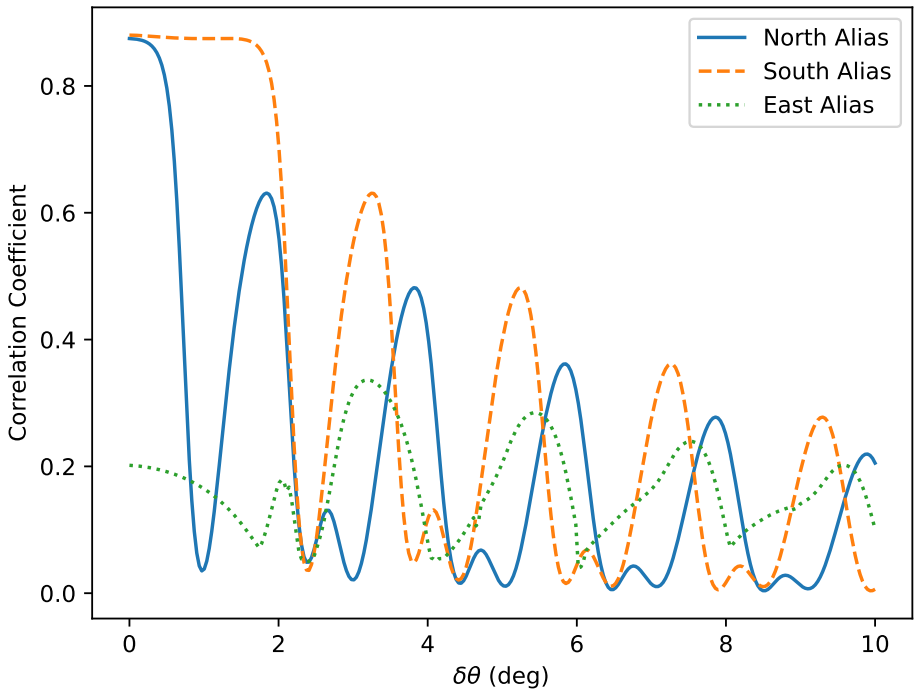


Figure 2.10: This plot shows how the severity of aliases depends on the source's position within the observed strip of sky. CHORD is most sensitive to a strip about 3 degrees north and south of the direction that it is pointing in. The north, south, and east alias positions are computed, and their correlation coefficients are plotted.

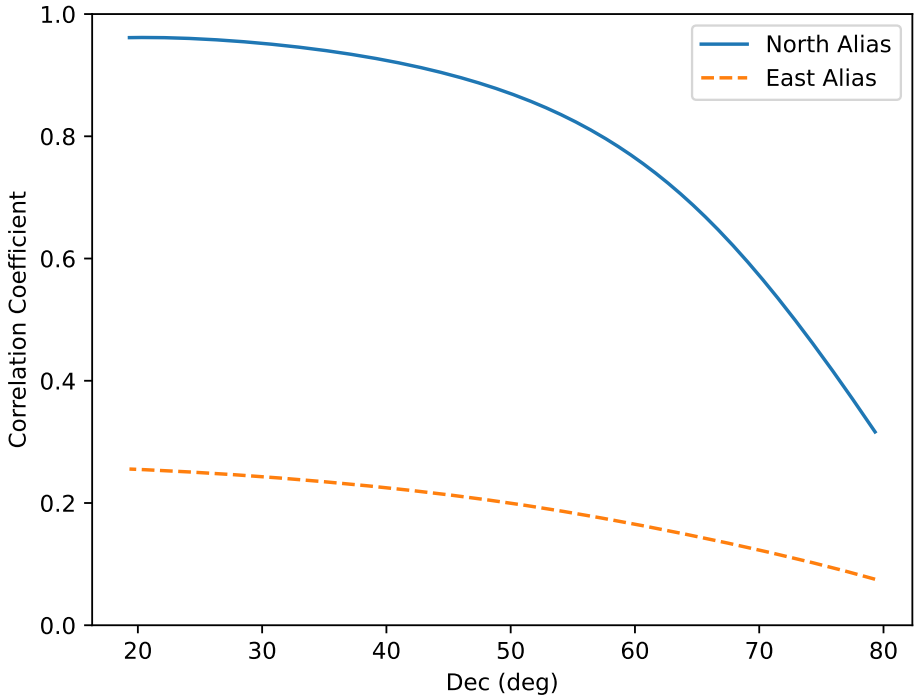


Figure 2.11: The relationship between alias disambiguation and the direction that CHORD is pointing in.

correlation coefficient is reduced a lot by the primary beam crossing effect regardless of declination. The north and south aliases can be disambiguated once curvature is important, but at low declinations in the flat sky regime, they are almost indistinguishable from the true source.

This relationship is key to optimizing the CHORD scan strategy so that we detect the largest number of galaxies. Scanning at high declination decreases the solid angle scanned each night, but also decreases the correlation coefficient of aliases of detected sources and increases the sensitivity, since the exposure time (signal collected) per unit area is larger. A lower correlation coefficient decreases the probability of a detection[3], but also decreases the probability of alias confusion as a trade-off. A more in-depth explanation of the probability of each outcome is in section 2.6. A more in-depth discussion about

[3]Usually. There is actually a global maximum that occurs close to but below $R = 1$.

considerations for survey design is in the discussion section 4.2.1.

We have seen that integrating over time provides mechanisms to reduce the alias correlation coefficients below 1. However, there exists a limiting case where an alias correlation coefficient approaches 1 because neither the primary beam crossing effect nor the curvature effect apply. This is bad because galaxies in this corner of parameter space will be subject to high alias confusion probabilities.

To produce this scenario, CHORD must be pointed at the equator in order to enter the flat sky regime. Then, the source must be placed a distance below CHORD such that its northern alias is the same distance above CHORD. In this scenario, the alias's and the source's primary beam response will be identical since their path through the primary beam is the same except reflected across the equator. Figure 2.12 illustrates the situation.

Figure 2.12 proves that an alias with a correlation coefficient of 1 is possible in the integrated case. CHORD will never actually point at the equator, but when it points at low declinations, then galaxies which are about this angular distance from CHORD will come with aliases that are nearly indistinguishable. This is the limit of using information inferred from pointing CHORD in a single direction. If we combine data from pointing CHORD in multiple directions, then this degeneracy can also be broken.

2.5 Dithering

CHORD's pointing can be changed in the altitude (Dec) direction but not in the azimuth (hour angle) direction; it always points along the meridian. If it is repointed by a small amount multiple times, this is called dithering, and this has applications for the galaxy search. The process of repointing must be done manually for all 512 dishes, and it is estimated that this will take about a week. Furthermore, changing direction might be counterproductive for some of CHORD's other science cases. This option is costly, but it has enormous power for alias disambiguation.

Dithering enters the matched filter equation as another dimension of the template and visibility arrays. The arrays depend on CHORD's pointing $\vec{u}_p$, which is equivalent to its declination θ. The dot product between the template and the visibility array can be extended to also sum over the dithering angles θ. The correlation coefficient equation can

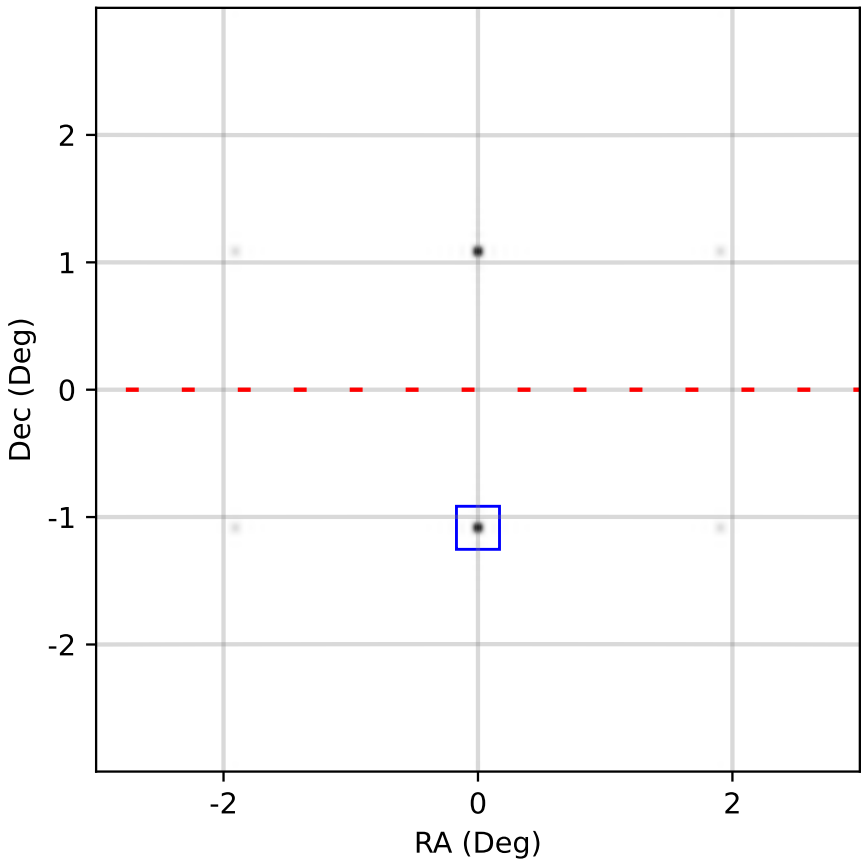


Figure 2.12: This setup features an alias with a correlation coefficient of 1, even in the integrated case. The dashed line indicates CHORD's declination, and the box □ indicates the position of the source.

be generalized to

$$R_{uu_s} = \frac{\sum_{\tau,\theta} B\left(|\mathcal{R}(\omega\tau)\vec{u} - \vec{u}_p(\theta)|\right)^2 B\left(|\mathcal{R}(\omega\tau)\vec{u}_s - \vec{u}_p(\theta)|\right)^2}{\left(\sum_{\tau,\theta} m_1^2 m_2^2\right)\sqrt{\sum_{\tau,\theta} B\left(|\mathcal{R}(\omega\tau)\vec{u} - \vec{u}_p(\theta)|\right)^4}\sqrt{\sum_{\tau,\theta} B\left(|\mathcal{R}(\omega\tau)\vec{u}_s - \vec{u}_p(\theta)|\right)^4}} \frac{\sin\left(\frac{\pi m_1 L_1}{\lambda}\hat{e}_1 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_1}{\lambda}\hat{e}_1 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2} \frac{\sin\left(\frac{\pi m_2 L_2}{\lambda}\hat{e}_2 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_2}{\lambda}\hat{e}_2 \cdot R(\omega\tau)(-\vec{u} + \vec{u}_s)\right)^2} \quad (2.10)$$

This is implemented into the analytic matched filter as summing over multiple complete Earth rotations, each with CHORD pointed at a different direction. If the plot of the worst-case scenario (fig 2.12) is remade with dithering, then the previous alias with correlation coefficient 1 would have a correlation coefficient less than 1 (fig 2.13). This is because, as can be seen in figure 2.10, the alias correlation coefficients change depending on the relative positions of the source and CHORD. A near-degenerate alias in the case of one relative position will not be nearly degenerate when CHORD is pointed somewhere else.

Dithering introduces a new controllable $\delta\theta_d$, which is the angular distance that CHORD is repointed between dithers. If $\delta\theta_d$ is too low, then the dither is almost a repeated measurement and no new information is added. If $\delta\theta_d$ is too high, then CHORD at its new pointing will not pick up any signal from the source and cannot help. Therefore, there must be some optimal $\delta\theta_d$ to improve alias disambiguation. Figure 2.14 demonstrates that for the best alias disambiguation, CHORD should be repointed to the locations of the aliases. This location depends on the position of the source, so figure 2.14 does not have general applicability. In the common case where the source is close to where CHORD is pointing, then the optimal angle becomes about 2 degrees. From an alias disambiguation perspective, ideally, the entire sky would be covered in 2 degree dithered strips. The advantages of this are likely outweighed by the time costs of repointing the array.

The final plot will be a comparison between the instantaneous case figure 2.1 and the same scenario but with time integration and dithering in order to show how both sources of new information allow the true source to be easily picked out by the matched filter. This is figure 2.15.

If there was no noise in the real data, as is the case with the AMF, then aliasing would no longer be a problem. After the matched filter is run, picking the highest value would always be the position of the real source. This is not the case when there are random noise draws. In the next section, we will translate the correlation coefficients that we have found into probabilities of alias confusion.

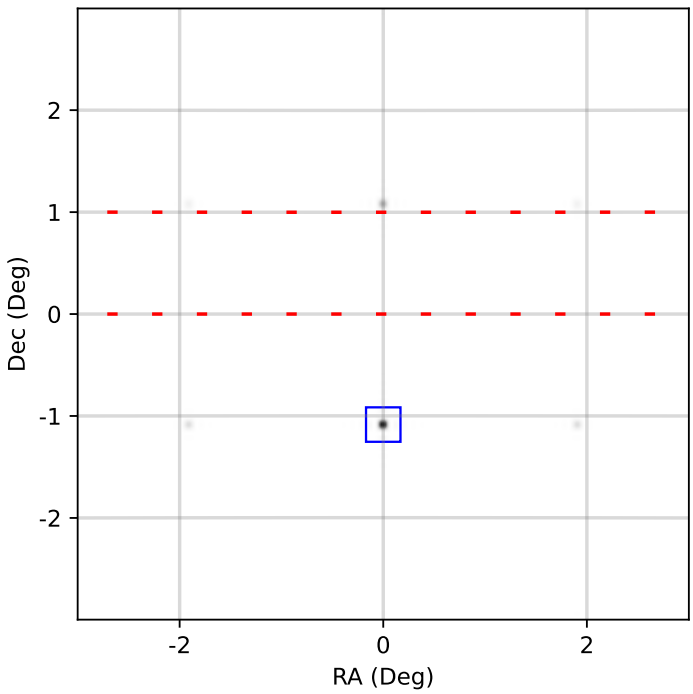


Figure 2.13: A recreation of figure 2.12 except with a single dither upward by 1 degree. The dashed lines show the two declinations which CHORD is pointed at. In figure 2.12, the northward alias has a correlation coefficient of 1, but with dithering it is reduced to below 1.

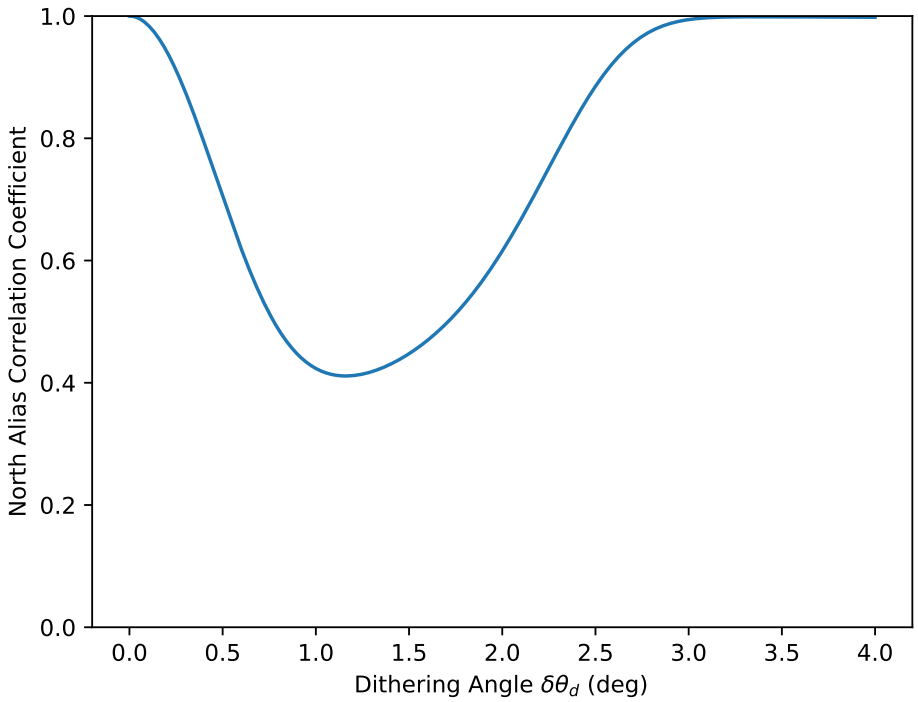


Figure 2.14: This plot shows how the northern alias's correlation coefficient changes depending on how much CHORD is rotated between dithers (i.e. how much the two dotted lines are separated by in fig 2.13). The setup is the same "worst case" setup as in figure 2.12. The minimum is at $\delta\theta_d = 1.16$.

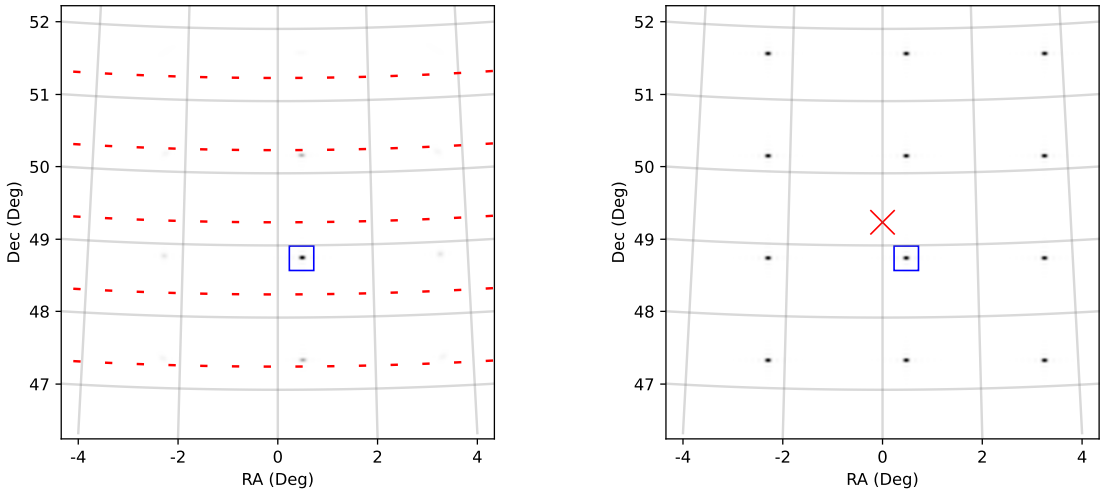


Figure 2.15: Left side: The correlation coefficients given by equation 2.10 for a typical declination and source position. CHORD is dithered four times, and the dithering angles are indicated with dashed lines. The position of the true source is indicated by the box □. Right side: a copy of figure 2.1, the instantaneous case, for comparison.

2.6 Probability

The output of the analytic matched filter is the correlation coefficient of a point on the sky. That is, how much is the output of the matched filter at that point would correlate with the output of the matched filter pointed on the true source. The analytic matched filter tells us that there is a regular grid of correlation peaks separated by about two degrees. The matched filter will return the position of the brightest peak as the position of a galaxy and the height of that peak in units of SNR. For each source, there are three possible outcomes: first, that the matched filter output never exceeds the SNR threshold; second, that random noise causes one of the aliases to have larger SNR than the true source location and thus appear as the more likely location for the source; and third, that the matched filter correctly chooses the correct source location. We would like to determine the probability of each of these outcomes given precomputed correlation coefficients.

The grid of aliases theoretically spans the entire sphere. However, we only need to consider some these aliases as possible alternative options. The degeneracy that the aliases represent in the first place is between the source brightness and the primary beam attenuation. If a source is farther away from where CHORD is pointing, it appears dimmer. The amount of dimming is given by the primary beam equation (eqn 2.2)

$$B^2(\alpha) = \frac{J_1(\tilde{\alpha})}{\tilde{\alpha}}; \quad \tilde{\alpha} = \frac{\pi d \alpha}{\lambda}$$

Where α is the angle away from where CHORD is pointing. If the true source is directly over CHORD, and an alias candidate is λ/L radians away, then a source in the alias location must be $\left(\frac{\pi d}{L J_1(\frac{\pi d}{L})}\right)^2 \approx 3.6$ times brighter in order to compensate. A source two grid-distances away must be 165 times brighter. It's unlikely that a two-distance away alias would present a problem, since the true source would have to be 165 times brighter than the detection threshold for us to consider the possibility that they are aliased from two spaces away. We can safely ignore these cases since the very brightest sources are already known. For the rest of the probability analysis, I will only consider the four nearest grid spaces.

This probability can be computed analytically in a special case, or numerically generally. The special case is when outcome # 1 is extremely unlikely, i.e. the target source is so bright that there is negligible probability that it is not detected. In this case, we only need to determine the likelihood that an alias location randomly becomes higher than the true location.

2.6.1 Case 1: 100% Detection Probability

Let the array $M(u) = \frac{t_u^\dagger N^{-1} d}{t_u^\dagger N^{-1} t_u}$ be the results of a matched filter run on noisy data. We want to find, given $u_1, u_2, \ldots, u_N$, the probability that one of $M(u_1), M(u_2), \ldots, M(u_N)$ is larger than $M(u_0)$, which is the matched filter value at the desired signal location. Say that you know you detect the actual signal at n_σ detection strength. Assume that R_{ij} is known, which is the correlation coefficient between u_i and u_j.

First, we need to model the Probability Distribution Function (PDF) $g(M(u_0), M(u_1), \ldots, M(u_N))$. We are assuming Gaussian noise, so the PDF will be a Gaussian $g(\mu, \Sigma)$. The expectation value μ is given by n_σ and the correlation coefficients.

$$\mu = \begin{pmatrix} 1 \\ R_{01} \\ \vdots \\ R_{0N} \end{pmatrix} n_\sigma \tag{2.11}$$

The diagonal of the covariance matrix Σ is all 1s. This is because the matched filter variables are using units of sigmas. The first row/column is how much the noise for each other u_i is correlated with the noise at u_0. This is just $R_{01}, R_{02}, \ldots, R_{0N}$. In fact, the covariance matrix is the correlation coefficient R matrix.

$$\Sigma = \begin{pmatrix} 1 & R_{01} & R_{02} & \ldots & R_{0N} \\ R_{10} & 1 & R_{12} & \ldots & R_{1N} \\ R_{20} & R_{21} & 1 & \ldots & R_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{N0} & R_{N1} & R_{N2} & \ldots & 1 \end{pmatrix} \tag{2.12}$$

And the PDF is

$$g(M(u_0), M(u_1), \ldots, M(u_N)) = (2\pi)^{-N/2} \sqrt{\det(\Sigma^{-1})} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right) \tag{2.13}$$

where x is the vector $(M(u_0), M(u_1), \ldots, M(u_N))$.

Second, the probability P of no confusion is the probability of $(M(u_1) < M(u_0))$ and

$(M(u_2) < M(u_0))$ and ... and $(M(u_N) < M(u_0))$, which is the integral

$$P = \int_{-\infty}^{M_0} \dots \int_{-\infty}^{M_0} \int_{-\infty}^{M_0} \int_{-\infty}^{\infty} g(x)\ \mathrm{d}M_0\ \mathrm{d}M_1\ \mathrm{d}M_2\ \dots\ \mathrm{d}M_N$$

$$P = (2\pi)^{-(N+1)/2}\sqrt{\det(\Sigma^{-1})} \int_{-\infty}^{M_0} \dots \int_{-\infty}^{M_0} \int_{-\infty}^{M_0} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \mathrm{d}...$$

Σ^{-1} is diagonalizable because Σ is symmetric. We can write $\Sigma^{-1} = A^{-1}DA = A^TDA$ since A is orthogonal (implied by Σ being symmetric).

$$P = (2\pi)^{-(N+1)/2}\sqrt{\det(\Sigma^{-1})} \int_{-\infty}^{M_0} \dots \int_{-\infty}^{M_0} \int_{-\infty}^{M_0} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(A(x-\mu))^TDA(x-\mu)\right) \mathrm{d}...$$

Now we change the integral into coordinates $y = Ax$. A is a rotation matrix, so the determinant of A is 1, meaning that I can change the differential form from $dM_0\ dM_1\ ...\ dM_N$ into $dy_1\ dy_2\ ...\ dy_N$. In doing this, we rotated the Gaussian to align with the axes, so that means the bounds of integration now go from 0 to ∞ except for the first variable, which is still $-\infty$ to ∞.

$$\begin{aligned}
P &= (2\pi)^{-(N+1)/2}\sqrt{\det(\Sigma^{-1})} \int_{-\infty}^{0} \dots \int_{-\infty}^{0} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(y-A\mu)^TD(y-A\mu)\right) \mathrm{d}y_0\ \mathrm{d}y_1\ ...\ \mathrm{d}y_N \\
&= (2\pi)^{-(N+1)/2}\sqrt{\det(\Sigma^{-1})} \left(\int_{-\infty}^{\infty} \exp\left(-\frac{D_0}{2}(y_0 - (A\mu)_0)^2\right) dy_0 \right. \\
&\qquad \left. \times \prod_{i=1}^{N} \int_{-\infty}^{0} \exp\left(-\frac{D_i}{2}(y_i - (A\mu)_i)^2\right) dy_i \right) \\
&= (2\pi)^{-(N+1)/2}\sqrt{\det(\Sigma^{-1})} \left(\frac{\sqrt{2\pi}}{\sqrt{D_0}}\right) \prod_{i=1}^{N} \frac{\sqrt{\pi}\left(\mathrm{erfc}\left(\frac{(A\mu)_i\sqrt{D_i}}{\sqrt{2}}\right)\right)}{\sqrt{2D_i}}
\end{aligned}$$

Where D_i is the ith diagonal element of D. We can make the substitution $\det(\Sigma^{-1}) = \det(A^{-1}DA^{-1}) = \det\{(D)\} = \prod_{i=0}^{N} D_i$.

$$P = \frac{1}{2^N} \prod_{i=1}^{N} \mathrm{erfc}\left(\frac{(A\mu)_i\sqrt{D_i}}{\sqrt{2}}\right)$$

2.6.2 Case 2: Possibility of No Detection

In this case, a third region of the PDF must be considered, so the bounds of integration change. In this situation, there is a minimum threshold σ_T (usually 5 sigma) that the matched filter qualifies as a detection. The region of no detection is the quadrant of the PDF where every axis is below σ_T.

$$P_{\text{no detect}} = \int_{-\infty}^{\sigma_T} \cdots \int_{-\infty}^{\sigma_T} \int_{-\infty}^{\sigma_T} PDF(M)\ dM_1 dM_2 \cdots dM_N$$

The region of detection without confusion becomes the no-confusion region in the last section minus the no detection region. The boundaries of this region are no longer perpendicular, so the Gaussian cannot be rotated to make the integral easier. This is shown visually in figure 2.16. The integral is instead evaluated numerically with the Monte Carlo method.

The conversion between correlation coefficients and the probabilities of these outcomes can be applied to the plots in section 2.4.2. This allows us to quantify the effects that the various parameters will have on the final catalogue. For example, figure 2.17 shows the fraction of galaxies that will be missed or confused as a function of CHORD's declination.

Figure 2.17 mirrors figure 2.11. At low declinations, the correlation coefficients of the north and south aliases are high, so the matched filter has a high probability of confusing one of the aliases for the true position. On the other hand, more aliases that are near the detection threshold allow for more draws for a chance to count as a detection. Even though those draws are correlated, the plots shows that this effect decreases the probability of no detection by about 3%. The region between the orange and blue lines is the outcome where the matched filter makes a detection, but reports the wrong position because of confusion. This region is less significant near the north pole for reasons discussed in section 2.4.

Another axis to consider is the strength of the source. Figure 2.18 fixes the alias correlation coefficients and the detection threshold and varies the source strength. The extremes of this plot are predictable. As the source strength approaches 0σ, it will not overcome noise and the detection probability is 0. As the source strength approaches infinity, the probability of the source dropping below the noise level approaches 0, and the probability of alias confusion is also low as long as the correlation coefficient is less than 1. Even for very high correlation coefficients, the fraction that is uncorrelated is unlikely to make up for the drop in response value for high source strengths because the mean of the Gaussian is far away from the confusion region.

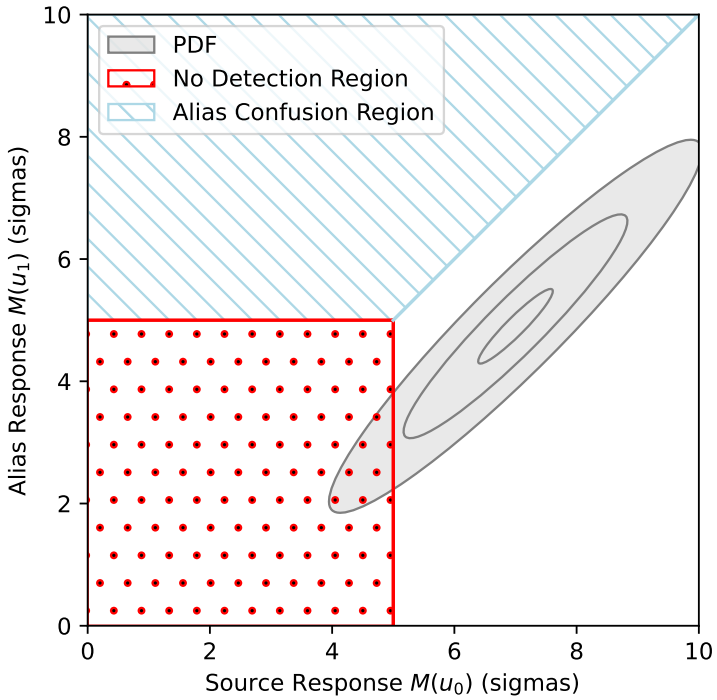


Figure 2.16: An illustration of the probabilities of the three different outcomes if only one alias is being considered. The axes are the results that the matched filter randomly draws for the source at u_0 and an alias at u_1 in units of sigmas. The source's strength is 7 sigma, and the correlation coefficient is 0.7. A higher correlation coefficient moves the Gaussian upward, but the contours would become tighter. The "No Detection" region is the case where both the true source and the alias draw below the 5σ threshold. The "Alias Confusion" region is when the alias draws higher than the true source.

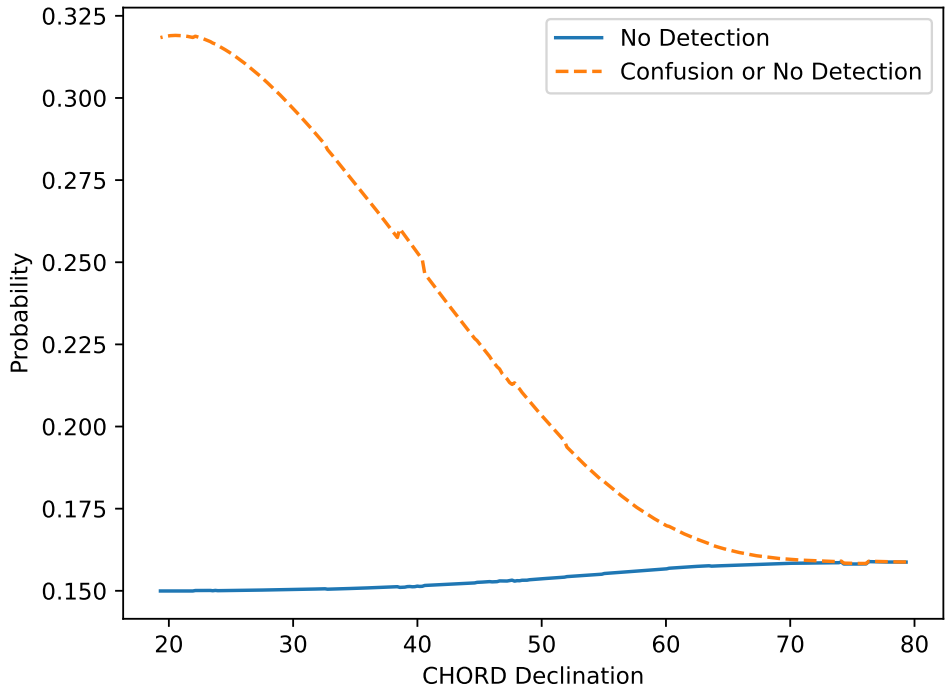


Figure 2.17: A prediction of the fraction of missed and confused galaxies depending on where we point CHORD. The source is directly over where CHORD is pointing. The detection threshold is 5σ and the source strength is 6σ.

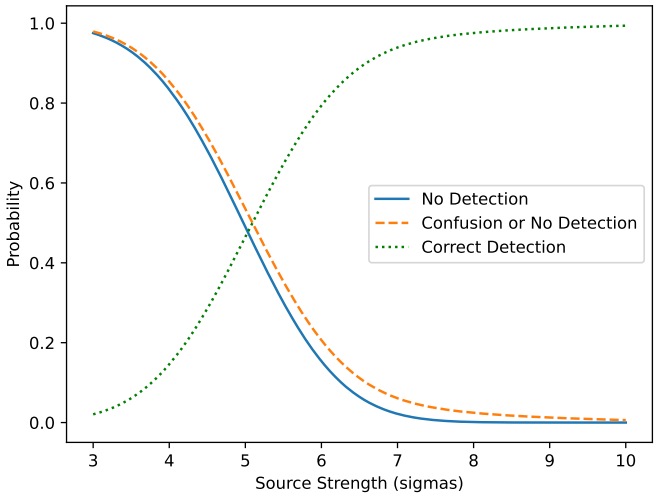


Figure 2.18: A prediction of the fraction of missed and confused galaxies depending on the strength of the source. CHORD is pointed at zenith and the source is offset by 0.3 degrees. The detection threshold is 5σ.

These probabilities allow us to predict what percentage of the final catalogue will be problematic. The catalogue can record cases where alias confusion is likely so that the end user can make an informed decision about which galaxies to consider. These predictions can also help inform where to point CHORD in order to minimize alias confusion.

Chapter 3

Frequency Matched Filter

The method for the frequency axis search largely follows [26]. It consists of running a matched filter on the frequency channels of a single spatial pixel using a family of templates. The implementation for CHORD will differ by the upchannelization step and combining the result from separate pixels after the frequency step.

The goal of the frequency axis search is to match a pre-computed gas velocity dispersion template (t) to frequency axis data from a dirty map outputted from CHORD. The secondary goal is to minimize the number of flops through design choices because this search will be run over months worth of data. Subsection 3.1 will explain how these templates are chosen. Subsection 3.2 will explain the fast implementation of the matched filter. Subsection 3.3 demonstrates that the algorithm can handle multiple sources in the same pixel. Subsection 3.4 will explain how the frequency aliasing from CHORD's upchannelization pipeline will be accounted for efficiently.

It should be noted that decisions are made during this step to optimize the probability of detections and correct detections. This step might not necessarily produce useful results about the shape and position of the galaxy profiles. There is an additional step (explained in subsection 4.1.1) dedicated to expensive fitting for parameters that could be published in a catalogue.

3.1 Frequency Axis Templates

Gas in galaxies is cooler than the ICM, but it still has a random velocity dispersion which is measurable with HI observing instruments. A fraction of the total gas in galaxies emits

21 cm radiation which is shifted from the velocity Doppler effect as well as cosmological expansion redshift. These two effects are visible as a width and position of a spike in power on the frequency axis because the concentration of HI gas in galaxies is much higher than the concentration outside of galaxies. The THINGS survey [31] measured the velocity profile of the HI gas of 34 nearby galaxies. It found that a typical galaxy has an HI velocity profile width of tens to hundreds of km/s. A discussion of how well we need to sample this profile is given in subsection 3.4; this section is only concerned with the shape of the profile. THINGS also confirmed that overall rotation of edge-on galaxies will manifest as a bimodal splitting of the spike because of the extra correlated rotational component of the velocity. Any family of templates should account for possibility of "horns" on the profile.

A model of the HI density curve and the rotation curve of a galaxy are required to predict its observable spectrum. Figures 1.3 and 1.4 in the introduction are examples of each. All HI gas emits at the same frequency (1428 MHz), so a galaxy's spectrum is caused entirely by Doppler shift, and the x-axis of the spectrum can use velocity units. We want to derive the shape of the spectrum that CHORD will output when observing a flat, circularly symmetric, rotating galaxy.

Let the inclination of the galaxy be i. Take a patch of gas with polar coordinates (R, ϕ) where $\phi = 0$ is facing toward the observer. The relationship between gas's rotational velocity V_{rot} and the observable recessional velocity V is

$$V(R, \phi) = \sin(i) \sin(\phi) V_{rot}(R) \tag{3.1}$$

The total integrated flux from HI gas for a particular frequency $F(V)\mathrm{d}V$ for a flat galaxy is

$$F(V)\mathrm{d}V = \frac{M(V)}{(2.356 \times 10^5 \ \mathrm{M}_\odot) D_{\mathrm{Mpc}}^2} \ \text{Jy km/s}$$

where D_{Mpc} is the distance to the galaxy in Mpc, and $M(V)$ is the total integrated HI mass which has the recessional velocity V. The flux that CHORD will output will be binned by some width ΔV, so the integrated mass integral is

$$M(V) = \int_{\{(R,\phi) \ | \ V \leq V'(R,\phi) < V + \Delta V\}} \Sigma(R) \ \mathrm{d}\Omega \tag{3.2}$$

where $\mathrm{d}\Omega$ is a disk area element and Σ is the mass distribution function of the disk.

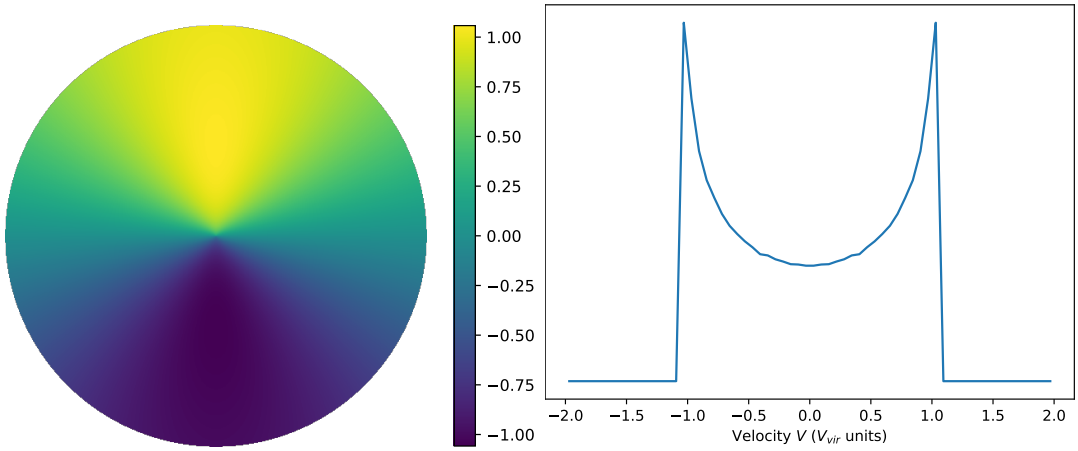


Figure 3.1: Left: The recessional velocities V on a disk in units of V_{vir}. The observer is in the positive x direction. Right: The implied spectrum given an exponential disk density.

Take for example a disk of HI gas with inclination $i = 0$, an exponential disk density profile given by equation 1.3, and the rotation curve from and NFW dark matter halo [23] given by

$$V_{rot}(R) = \sqrt{\frac{1}{x}\frac{\ln(1+cx) - cx/(1+cx)}{\ln(1+c) - c/(1+c)}}V_{vir} \qquad \text{where } x = \frac{R}{R_{vir}}.$$

The recessional velocity $V(R, \phi)$ is plotted in figure 3.1. To get the spectrum, the mass density must be integrated along the iso-velocity strips. The resulting spectrum is also plotted in figure 3.1.

The last step is to take into account the dispersion velocity component. This can be done by convolving by a Gaussian representing random normal dispersion. The width of this Gaussian is expected to be on the order of 10 km/s, much less than the rotational velocities of galaxies on the order of about 100 km/s. This results in a spectrum like the one in figure 3.2. This is the template that we want to use to search for galaxies on the frequency axis. In practice, we will be using a standard family of templates that resembles this shape.

There is a balance between the number of templates chosen and computational cost. A family of templates adds another index to the matched filter equation, so the total number of flops will be increased by the number of templates chosen. On the other hand, a larger variety of templates increases the space of galaxies which will pass the matched filter detection threshold. The strategy for choosing a family of templates will attempt

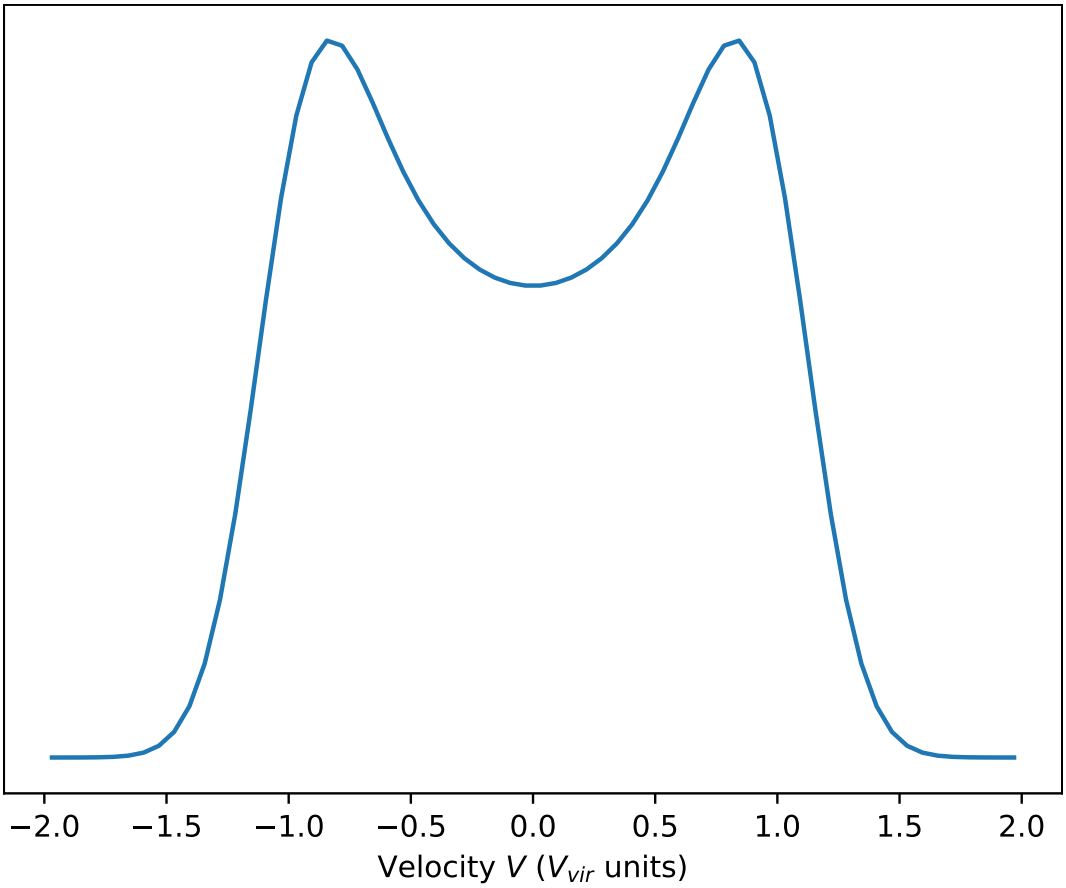


Figure 3.2: The HI spectrum of a flat, circularly symmetric rotating disk galaxy. This is just one example shape; the horns and the width will vary as the chosen mass and rotational velocity changes. The convolution step makes the horns less sharp, as can be seen by comparing with figure 3.1.

to optimize the range of size and shape of detectable galaxies for the lowest number of templates.

I considered two strategies for generating gas velocity dispersion templates. One is a single-parameter family generated from symmetric Hermitian functions used by the ALFALFA galaxy search [26]. The other is a multi-parameter family called busy functions which is used more commonly for current applications [32].

The ALFALFA matched filter strategy is motivated by reducing computational cost and the observation that Gaussians only well-approximate the distribution of galaxies with thin profiles. Galaxies will have wide profiles if they have edge-on rotational velocity, which then gives the profile horns and makes it flatter. Wider distributions are better approximated by a boxcar function. A one-parameter family is proposed that smoothly transitions a thin Gaussian into a wide bump with horns. A one-parameter family greatly reduces the number of templates. This strategy minimizes the number of missed galaxies since the number of galaxies that do not follow the rotation-width correlation is low.

The family is defined smoothly piecewise depending on a width parameter[1]. The three piecewise subsets are shown in the left side of figure 3.3. The thin profiles are Gaussians, which evolve to have horns and then a deep trough as the profiles become wider.

Westmeier et al. [32] criticizes this family because the horns are not as steep as the horns found in observations. They propose a function named the "busy function" which fixes that without sacrificing smoothness which is helpful for fitting algorithms. It consists of error functions multiplied by a positive-curvature polynomial to create the dip. This emulates the sharp horns observed on real spectra e.g. HIPASS [20]. The downside is that this family depends on three parameters. The asymmetric version which allows for variable dip shape requires 6 parameters. These would provide much better matching templates, which would boost SNR, but at the cost of a much larger search parameter space. Busy function templates are defined as

$$B_0(\xi) = (\mathrm{erf}(b(w+\xi))+1)(\mathrm{erf}(b(w-\xi))+1)(c\xi^2+1). \tag{3.3}$$

The three parameters are w, the half-width of the profile, b, which controls how flat the peak is, and c which controls the depth of the trough. Some examples of busy function templates are given on the right side of figure 3.3.

Now that the templates are chosen we can see how equation 1.9 will be computed. Define the template value $t_\ell[\kappa]$ as the κth element of a template that is centered at element ℓ. The numerator of the matched filter equation is the 1-dimensional convolution between the template centered at position ℓ and the data vector.

[1]The definition in [26] is incorrect. $a_{N,i}$ is missing a division sign.

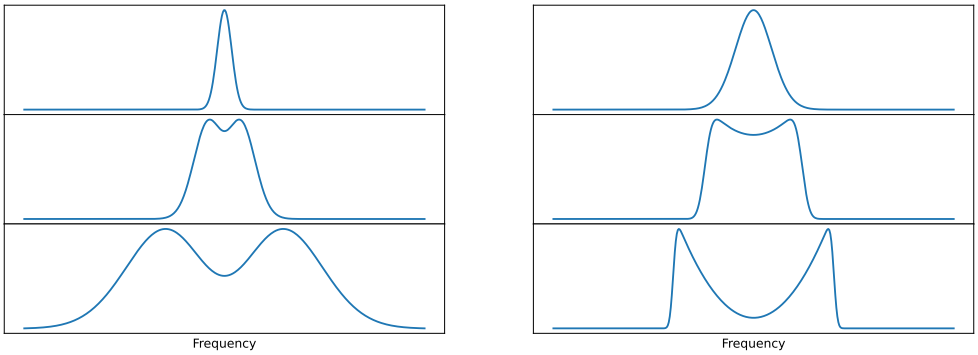


Figure 3.3: Left: Three of the templates from the family defined by [26] for the ALFALFA matched filter. Top left: A thin Gaussian. Bottom left: A linear combination of a Gaussian and the second Hermite Polynomial to give the distribution horns. Center left: A transition between the two. Right: Three of the templates from the family defined by [32] called busy functions. Top right: Opposing error functions with no dip. Bottom right: Farther apart error functions with a quadratic dip. Center right: A transition between the two.

3.2 FFT Matched Filter

A matched filter with a template that only depends on the array position by shifting by that amount (i.e. $t_\ell[\kappa] = t_0[\kappa - \ell]$) can be rewritten as a convolution times a data-independent prefactor. To see this, take the matched filter equation where d is a 1-dimensional array of length n. Then we have

$$\begin{aligned} M_\ell &= \frac{t_\ell^\dagger N^{-1} d}{\sqrt{t_\ell^\dagger N^{-1} t_\ell}} \\ &= \frac{1}{\sqrt{t_\ell^\dagger N^{-1} t_\ell}} \sum_{\kappa=0}^{n} t_0[\kappa - \ell](N^{-1}d)[\kappa] \\ &= \frac{(t_0 * (N^{-1}d))[\ell]}{\sqrt{t_\ell^\dagger N^{-1} t_\ell}} \end{aligned}$$

This is advantageous because of the convolution theorem (see appendix D). This allows

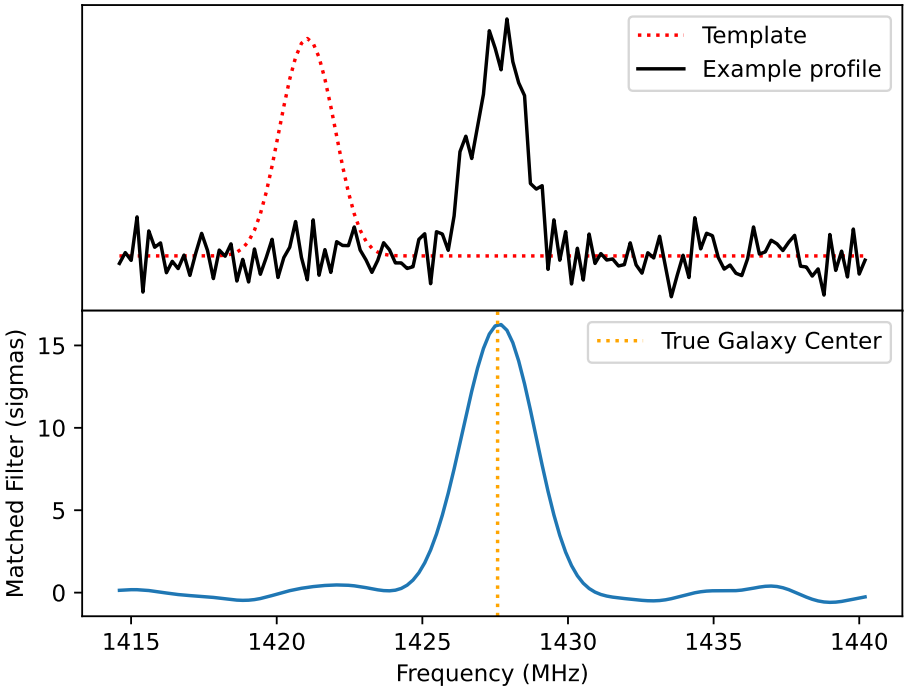


Figure 3.4: Top: An example source with randomly generated Gaussian noise. Also shown is a template centered at the wrong frequency. Bottom: The result of running the matched filter (equation 3.4). The peak aligns with the center of the example source.

us to write the above as

$$M_\ell = \frac{n \ \text{idft}(\overline{\text{dft}(t_0)}\text{dft}(N^{-1}d))}{\sqrt{t_\ell^\dagger N^{-1} t_\ell}} \tag{3.4}$$

We can compute this particular matched filter with 3 discrete Fourier transforms. The Fast Fourier Transform is order $n\log(n)$, which is an improvement over the naive n^2 computation time.[2] This algorithm can be seen in action by placing the desired template on a background with normally-distributed noise. The maximum of the vector that is returned is the location where the data vector best matches the input template, as per the proof in appendix A.

[2]Note that 0 padding is required so that the Fourier transforms do not treat the vector as periodic.

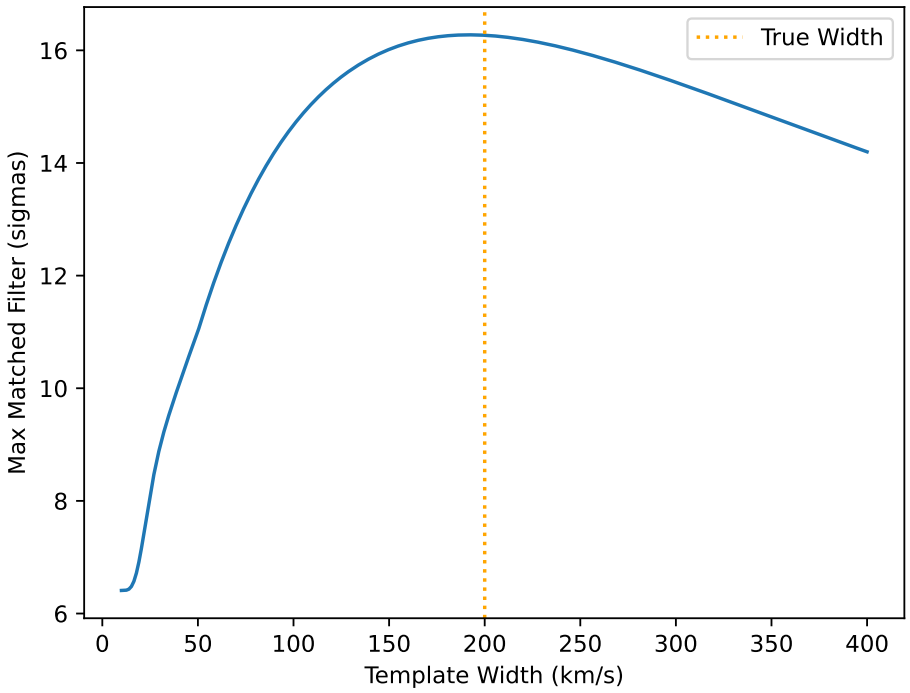


Figure 3.5: Trying matched filters on the same signal as in figure 3.4 with varying width templates.

Figure 3.4 shows the problem setup and output. From this noisy spectrum, we want to automatically find the template position that would have a high chance of producing it. The highest point of the output of the matched filter is the best guess for this position, in units of SNR.

In this example, I ran the search using the exact template that was used to produce the example spectrum. In reality, we will be running a finite number of pre-selected templates that we hope will cover the parameter space of galaxies in the sky. In order to choose which templates we will use, we need to know how detrimental the guess being wrong is. For example, if we search using a template that is too wide or thin, figure 3.5 shows that the strength of the detection decreases significantly. In this example, a template that is off by 100 km/s decreases the detection by 2 sigma. There's more discussion on choosing template widths in discussion section 4.2.1.

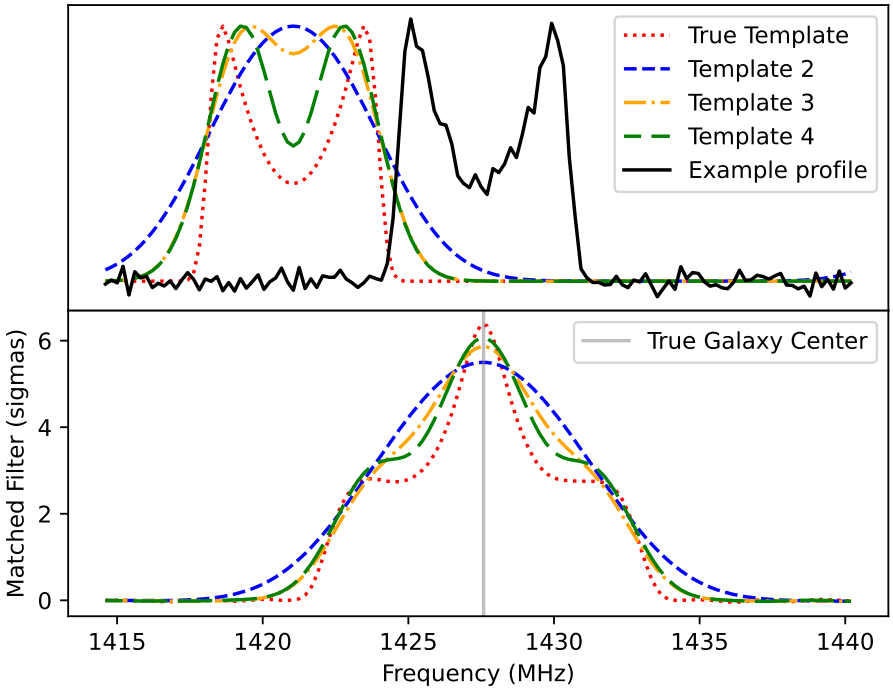


Figure 3.6: The matched filter is ran on a spectrum using 4 different templates of roughly equal width. Top: The spectrum and the 4 templates overlayed. Bottom: The results of the matched filter, with the plotting style corresponding to the template that was used.

Galaxy spectra will not only have different widths, but also different overall shapes, and matching those shapes is also critical to optimizing the detection strength. In particular, we want to take advantage of the distinct horn shape of galaxies that have them in order to pick faint ones out from noise. To do that, the matched filter needs to be run using templates that do and do not have horns. Figure 3.6 shows the performance of four different templates on an example profile generated using the first template which has strong horns. The four templates are controlled to have similar widths. The true template has a stronger detection than the other templates. In particular, it beats the template without horns by about 1 sigma. This higher sensitivity is valuable, but each extra template that is tried linearly increases the computational cost.

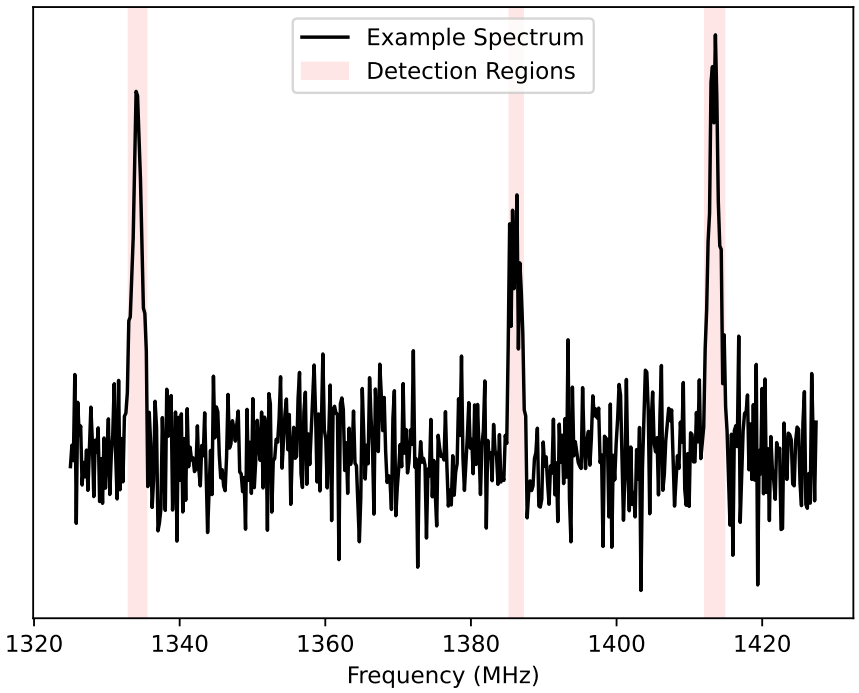


Figure 3.7: This plot shows the matched filter acting on a spectrum with three sources and Gaussian noise. The algorithm returns intervals where the matched filter is above a 5σ threshold which are marked by the transparent rectangles.

3.3 Multiple Detections

The size of the CHORD beam is 5 arcminutes at redshift 0 and 10% higher at redshift 0.1. It becomes increasingly likely that multiple galaxies will be in the same line of sight, so the matched filter needs to be able to return multiple detections. In matched filter space, the case of multiple detections looks like multiple peaks of varying heights. Ignoring worrying about frequency aliasing, the challenge is to decide, in the case of multiple peaks, whether they are separate objects or random noise spikes of a single object. Galaxy profiles are mostly monomodal, so the strategy is to fix some SNR threshold and to flag any bumps above this threshold as a separate detection. Figure 3.7 shows the algorithm navigating multiple peaks correctly.

3.4 Upchannelization

CHORD will operate in the 300 MHz to 1500 MHz frequency range. This will be divided into 6000 channels, which corresponds to a per channel resolution of 0.2 MHz or 42 km/s. A typical galaxy has a profile width in the tens to hundreds of km/s range. For galaxies with a low profile width, this default resolution might result in a galaxy width being less than a resolution element. If we want to leverage profile shape information, then a higher resolution is required. Upchannelization provides a means of trading time resolution for frequency resolution. There are multiple possible implementations of upchannelizations. The one I will be considering is outlined in an internal document written by Kendrick Smith [28].

Upchannelization will be done in the galaxies pipeline of the real-time processing of the CHORD data. All implementations of upchannelization work by taking a linear combination of the frequency data for multiple time steps in order to produce a single higher-resolution spectrum. It multiplies the resolution by the upchannelization factor U. We say CHORD has a coarse resolution of 0.2 MHz, and a fine resolution of $0.2/U$ MHz. The fine channels are evenly spaced within the larger coarse channels. This linear combination has the side effect of moving power from the true signal (i.e. the signal with infinite resolution) around to other places within the same coarse channel. For galaxies, this takes the form of moving signal away from the peak and into smaller spikes near the coarse channel boundaries. This movement of power is predicable, and we can record it in a Response Matrix R.

Given that we want upchannelization to increase resolution to take advantage of different profile shapes, we must investigate the side effects of upchannelization. The spikes that upchannelization tends to cause have a similar shape to Gaussians. The spikes' brightnesses are proportional to the brightness of the original source, which means that they might be confused for thin, dim sources. This is called frequency aliasing, and the additional spikes are called upchannelization artifacts. Some code by Olivia Pereira[3], inspired by the upchannelization document by Kendrick Smith, simulates the upchannelization response on a given profile.

Figure 3.8 shows how much distortion upchannelization causes. The template on the right, which consists of a single peak, gets split up into two peaks. Naively running a matched filter without accounting for upchannelization may return two detections when there should only be one. Additionally, it would return the center of the galaxy at the

[3]The code is hosted here: https://github.com/orpereira/CHORD_Research/tree/main/Channelization

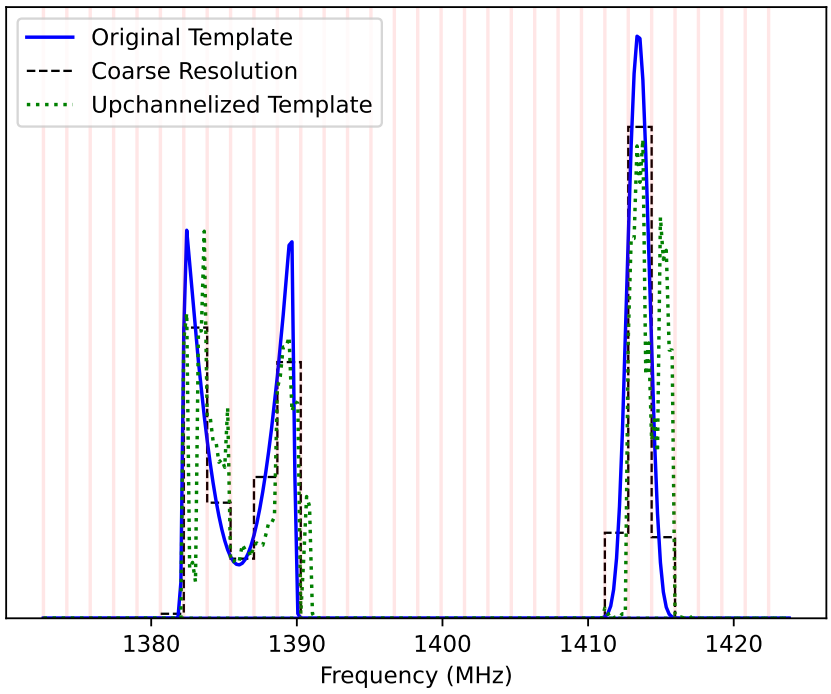


Figure 3.8: A figure showing the effects of upchannelization with a factor $U = 8$. The vertical transparent lines are the boundaries of the coarse channels. Solid: The original infinite-resolution template. Dashed: How the template would appear if not upchannelized at all, with only one step per coarse channel. Dotted: How the template appears after being upchannelized.

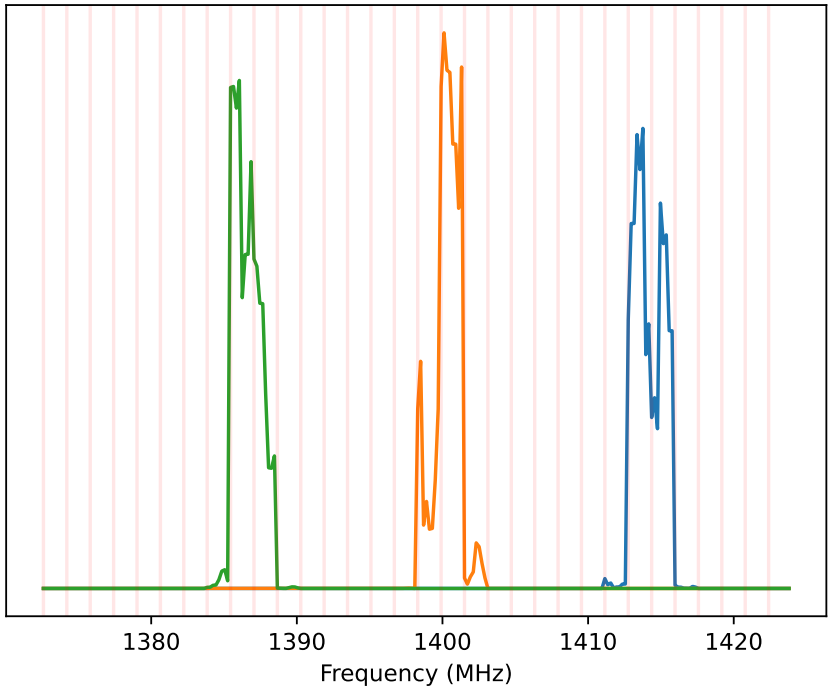


Figure 3.9: The same template centered in three different positions is affected differently by upchannelization. The vertical transparent lines show the positions of the coarse channel boundaries.

wrong frequency because the power is moved. The template on the left is wider, and upchannelization causes many spikes. We saw before how matching templates is important for optimizing the detection strength, so we want to use upchannelized templates when searching through upchannelized CHORD data.

An important consideration is that the power shifts from upchannelization are applied differently depending on the template's position with respect to the coarse frequency channels. Therefore, the FFT implementation strategy which relies on the assumption that the template stays static can no longer be used. Figure 3.9 shows how a template's shape after upchannelization changes chaotically depending on its position. Specifically, it's the position of the template relative to the coarse channel boundaries that affects its shape after upchannelization.

If we want to use templates that accurately capture upchannelization artifacts and implement the matched filter with $\mathcal{O}(n \log n)$ speed, then we need to do some rebracketing for the matched filter equation.

Define $\tilde{t}$ as the upchannelized template and t as the template with high resolution but no upchannelization artifacts. The relationship between the original template and the upchannelized template is

$$\tilde{t} = Rt$$

To take advantage of the correlation theorem, the template needs to stay the same shape with respect to its position. This is not the case for the upchannelized template, which changes when the center of t changes. This problem can be avoided with some re-bracketing.

The matched filter equation without noise is

$$\begin{aligned} M &= \frac{\tilde{t}^T d}{\sqrt{\tilde{t}^T \tilde{t}}} \\ &= \frac{(t^T R^T) d}{\sqrt{(t^T R^T)(Rt)}} \\ &= \frac{t^T (R^T d)}{\sqrt{t^T (R^T R) t}} \end{aligned}$$

This is almost the same as a matched filter with t as the template instead of $\tilde{t}$. The only difference is the $R^T R$ in the denominator. $R^T R$ is a diagonal matrix, so this operation is a one-dimensional vector multiplication. Also, it is an operation in the denominator, so it must only be performed one time per template. The equation is more complicated if there is noise:

$$M = \frac{t^T (R^T N^{-1} d)}{\sqrt{t^T (R^T N^{-1} R) t}}$$

But the correlation theorem can still be used.

$$M_\ell = \frac{n \text{idft}(\overline{\text{dft}(t_0)} \text{dft}(R^T N^{-1} d))[\ell]}{\sqrt{t_\ell^T (R^T N^{-1} R) t_\ell}} \tag{3.5}$$

This allows us to account for upchannelization without significantly increasing computational cost. Figure 3.10 shows that ignoring upchannelization on an example upchannelized profile reduces the effectiveness of the matched filter. The correlation of the

upchannelized template with the upchannelized profile is higher than for the original template. Furthermore, the matched filter using the original template would return a source center that is offset by about a MHz because the galaxy's signal's mean value gets shifted by upchannelization, and this is only accounted for by the upchannelized template.

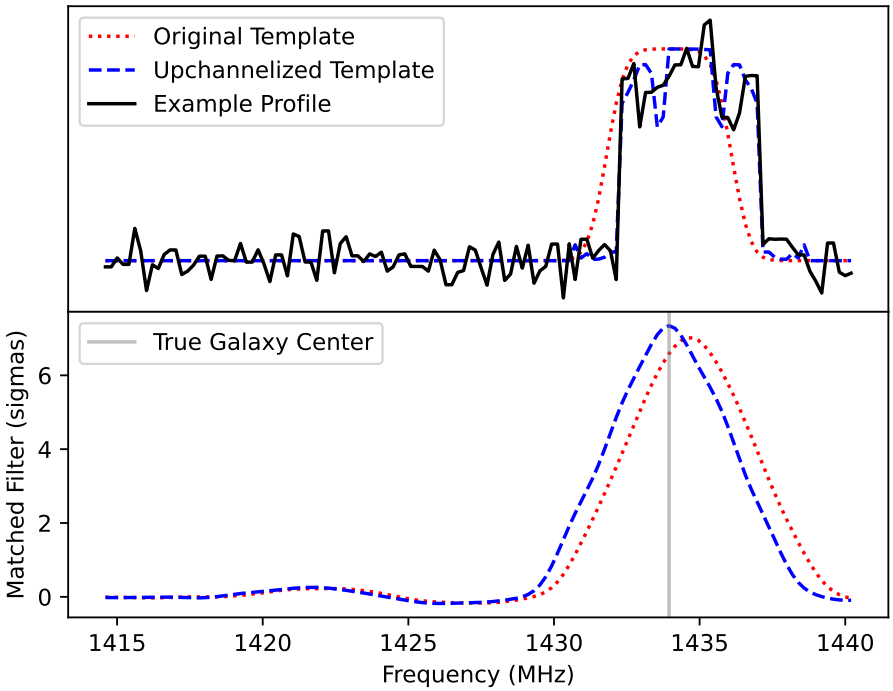


Figure 3.10: A plot comparing running the matched filter with and without accounting for upchannelization. For this example, $U = 8$. A spectrum with random noise is produced using the dotted template and then upchannelized. The two curves on the bottom are the results of the matched filter accounting for upchannelization (dashed) and not accounting for it (dotted). The dashed peak lines up with the original source's position.

Chapter 4

Discussion

4.1 Combined Matched Filter Design

In sections 2 and 3, I explained how the matched filter will be set up to run on the first two and third dimensions of the data respectively. In both sections, I assumed the other dimension(s) were held constant. For the full matched filter, we want to take advantage of galaxy signal that is shared across the spatial and frequency axes. There is different machinery to compute the two parts of the matched filter. Now, I'll gather the two main equations (eqns 2.4 and 3.5) from both sections to show how they can be combined.

$$M_{spatial} = \frac{\mathrm{tr}(t_u{}^\dagger N^{-1} V)}{\sqrt{\mathrm{tr}(t_u^\dagger N^{-1} t_u)}} \qquad M_{frequency} = \frac{n\mathrm{idft}(\overline{\mathrm{dft}(t_0)}\mathrm{dft}(R^T N^{-1} d))[\ell]}{\sqrt{t_\ell^T (R^T N^{-1} R) t_\ell}}$$

To explain how these equations combine, we need to be explicit about which indices are summed over and which are indices of the output. That is, in the equation

$$M_x = \frac{\sum_{i,i'} t_i^*(x) N_{ii'} V_{i'}}{\sqrt{\sum_{i,i'} t_i^*(x) N_{ii'} t_{i'}(x)}}$$

which axes or parameters go into x and which go into i_1, i_2. The ones that go into x are:

- The angular coordinate $\vec{u}$
- The frequency profile-center coordinate ℓ

- The busy template parameters w, b, c.

The ones that go into i, i' are:

- The dish coordinates j, k
- The frequency index κ
- Polarization index p
- Dithering angle index θ
- Time index τ

I will proceed in this section ignoring time integration and dithering for simplicity. The full matched filter equation can actually be written as

$$M_{\vec{u},\ell,w,b,c} = \frac{\sum_{p,p'} \sum_{\kappa,\kappa'} \sum_{k,k'} \sum_{j,j'} t^*_{j,k,\kappa,p}(\vec{u},\ell,w,b,c) N_{j,k,\kappa,p,j'k',\kappa',p'} V_{j',k',\kappa',p'}}{\sqrt{\sum_{p,p'} \sum_{\kappa,\kappa'} \sum_{k,k'} \sum_{j,j'} t^*_{j,k,\kappa,p}(\vec{u},\ell,w,b,c) N_{ii'} t_{j',k',\kappa',p'}(\vec{u},\ell,w,b,c)}}$$

This equation contains everything that the computer must process to return the matched filter array. The denominator is data-independent, so it should cost negligible compute time. Under the assumption that all templates are point sources spatially (any radio source with extent larger than 5 arcminutes is probably already known about) and that the frequency part of the template never depends on which pixel that it's in, then the sums in the numerator are separable. This is because the template can be split into $\tilde{t}_\kappa(\ell,w,b,c) t_{j,k}(\vec{u})$. This is using the notation that $\tilde{t}$ is the upchannelized frequency template. Note the p index on the template can be dropped because the templates are never polarized.

$$\text{numerator} = \sum_{p,p'} \sum_{\kappa,\kappa'} \tilde{t}_\kappa(\ell,w,b,c) \sum_{k,k'} \sum_{j,j'} t^*_{j,k}(\vec{u}) N_{j,k,\kappa,p,j'k',\kappa',p'} V_{j',k',\kappa',p'}$$

We'll take advantage of the fact that the dirty map, defined in equation 1.7 is equal to

$$d_{\vec{u},\kappa',p'} = \sum_{k,k'} \sum_{j,j'} t^*_{j,k}(\vec{u}) N_{j,k,j'k'} V_{j',k',\kappa',p'}$$

and is already computed on site, so does not add any additional cost to us.

$$\begin{aligned}
\text{numerator} &= \sum_{p,p'} \sum_{\kappa,\kappa'} \tilde{t}_\kappa(\ell, w, b, c) N_{\kappa,p,\kappa',p'} d_{\vec{u},\kappa',p'} \\
&= \sum_{p,p'} \sum_{\kappa} t_\kappa(\ell, w, b, c) \sum_{\kappa''} \sum_{\kappa'} R_{\kappa,\kappa''} N_{\kappa'',p,\kappa',p'} d_{\vec{u},\kappa',p'} \\
&= \sum_{p,p'} n \ \text{idft}(\overline{\text{dft}(t(0, w, b, c))} \text{dft}(R N_{p,p'} d_{\vec{u},p'}))[\ell] \qquad (4.1)
\end{aligned}$$

This is the equation that would be computed on Digital Research Alliance of Canada (DRA) computers in practice.

4.1.1 Post Matched Filter

After the matched filter is run, the next step is source finding. The outline of the strategy is a simple SNR threshold. Adjacent voxels that surpass the SNR threshold are grouped and flagged as a single detection. This strategy is complicated by the presence of aliases. It's possible that the matched filter will return a peak for the true source and a peak for an alias that are both above the SNR threshold. Whichever peak is higher will be saved at the galaxy's position, but it's important that all suspected alias positions are also flagged as the same detection in order to avoid multiple detections of the same galaxy. Therefore, the peaks would have to be flagged in order of brightest-to-dimmest so that no alias locations are flagged before flagging the true source.

Instead of flagging and ignoring suspected aliases, another option is to predict their signal strength and then subtract that off from the matched filter result. This strategy is called CLEAN, and its advantage is that it can recover sources that are covered by an alias from a brighter source. This can be risky because if the matched filter response of an alias is underestimated, then the algorithm won't subtract 100% of the alias and the remainder will appear like a dim galaxy. This would limit the sensitivity depth of the galaxy search. An alias randomly covering a dimmer galaxy would require the alias to land on the same pixel as well as occupy the same frequency channels. It will be the galaxy team's decision as to whether that will happen often enough to make it worth utilizing CLEAN.

There is an additional possible treatment for the polarization axis at the end of the matched filter step. The HI signal from galaxies is unpolarized, which can be used as a check before the polarization is summed over. If the matched filter response for one

polarization differs too greatly from the response for the other polarization for a would-be detection, then the detection can be discarded.

After the source finding step, each galaxy profile should undergo a more expensive profile-fitting step. The matched filter will return the profile's center up to the upchannelized resolution, and it will also return which profile template shape is the best match for the galaxy out of a small number of pre-selected templates. These can be used as initial guesses for a fitting algorithm, which would be too costly to run on the full spectrum for every single pixel. Running a fit only on spectra already flagged on containing galaxies is more reasonable and would make the final catalogue more useful. Also, this fitting step could be used as another rejection step if the fit fails. It's unlikely that spatial fitting is necessary in most cases because of CHORD's low resolution.

An additional anti-alias check can be made on spectrally-wide sources at the end of the matched filter step. As shown in section 2.3, the spatial positions of the aliases depends on the wavelength λ. For a sufficiently wide source, the position of its alias might change by a measurable amount between its highest or lowest frequency. If this change is detected, then the alias can be flagged as such, which reduces the probability of confusion because of random noise.

4.2 Integration with the Rest of the CHORD Galaxies Team

Several times in this thesis I make analytic or theoretical predictions about the success rate or alias confusion rate of the algorithm when it is applied to real data in the future. These predictions are useful in order to make informed decisions about whether to prioritize sensitivity or completeness. In this thesis, I showed figures which demonstrate changing one parameter at a time in an otherwise fixed scenario. These figures are meant to exemplify the effects of these parameters, but, to more precisely optimize the parameter choice, multiple parameters should be explored at once. It would be useful to test my matched filter code on realistic simulations while allowing all parameters to change for more concrete predictions.

These realistic simulations are being developed by the rest of the CHORD galaxies team. They use population models to randomly generate a sky with realistically sized and spaced galaxies. Using this simulation pipeline, we could see the ratio of detections to simulated galaxies using the matched filter. This would be another way of estimating the total number of galaxies that CHORD will detect. These simulations would also give us the probability of aliases covering up dim galaxies.

The simulation pipeline uses software called the CHORD M-mode Mapmaker. This takes in a sky and simulates the dirty map that CHORD would produce. The dirty map would be used in equation 4.1 at the end of section 4.1. As of the time of writing, there is a known discrepancy between my AMF and the M-mode mapmaker, which should in theory be performing the same computation in different ways. This might suggest that there is a bug in the AMF, so we are investigating the discrepancy. The M-mode mapmaker is currently under construction. There is a known issue where asking to simulate an array as large as CHORD causes a crash because of memory issues.

4.2.1 Choosing Parameters for a Survey

A theme of this thesis has been the impact of scan strategy or computational parameters on the effectiveness of the matched filter. CHORD's scan strategy will likely be chosen to match the needs of the other main science cases, but the galaxies team will have full control over the template parameters for the frequency matched filter. In this subsubsection, I will briefly discuss the considerations that would go into choosing parameters for a survey considering the 21 cm galaxies science case only.

Which declination to point CHORD at is largely a question of total counts versus completeness. If a drift telescope points toward the equator, the observed strip of sky will have the largest area. If it points closer to the poles, the observed strip will be smaller, which means each point of the sky stays in the primary beam for longer. Since sensitivity goes like $\sqrt{t_{int}}$ and the number of counts goes like area, optimizing the area is generally better for higher counts, but pointing at high Dec would give more sensitivity to find lower mass galaxies. CHORD's effective North-South resolution is lower if it points to high Dec. The probability of alias confusion is higher if it points to low Dec (see figure 2.17). For some science questions, like analyzing ram pressure stripping, it might also be desirable to aim at a particular cluster like Coma. All of these considerations must be balanced depending on the goal of the survey.

Dithering (moving CHORD by a couple of degrees) would always help reduce aliasing issues, but it comes at a time cost. Every time CHORD is moved, every dish must be set manually which will take on the order of a week, during which time CHORD is not observing anything. Whether dithering is good also depends on the area versus integration time balance. An extension question for this project would be: for a given time interval, how often should CHORD be dithered, and by how many degrees in order to optimize the number of counts and minimize aliasing. This could also be done using the simulation pipeline.

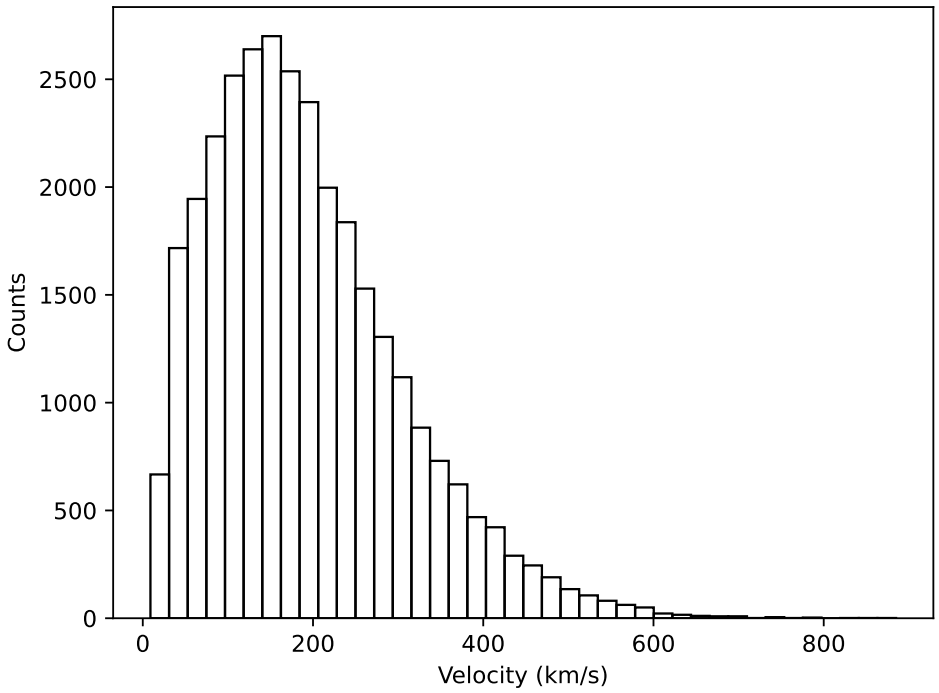


Figure 4.1: A histogram of the 50% widths (W_{50}) from the ALFALFA catalogue [15]. The median is 174 km/s.

On the frequency side, we must choose how many profile templates to use in the search. Each new template linearly increases the total time cost of the search, so there is pressure to minimize this number. We want to choose templates that are close in shape and width to as many real galaxies as possible, so that the correlation between the template and the real profile is high. The template that produces the best match with a galaxy can be used to guess that galaxy's shape parameters, but there will be a fitting step after the matched filter, so template parameters should not be oversampled for that purpose. Templates should only be chosen to optimize the number of detections.

Figure 4.1 shows data that could be used to inform the choice of the width parameter. This histogram can serve as a population model that we want to match with our distribution of templates. To get the highest counts, we might want to have templates with widths close to the peak to get the most close matches. If we want to make sure that we detect

galaxies on the extreme ends, then we also need templates with extreme widths. Note that CHORD is expected to see smaller mass galaxies which likely have smaller rotational velocities and dispersions, so this distribution would be shifted down a bit.

There is a strong correlation between the template width and the shape. Galaxies with high widths are probably rotating side-on to us, meaning that their distributions will probably have horns. It is this observation that led the ALFALFA matched filter to utilize a single parameter family of templates, where the wider ones got horns at the same time. This one dimensional space of templates costs much less to fully sample than the three-dimensional busy functions. Another extension of this project could be using an existing HI spectra catalogue in order to find the optimal 1 parameter family of busy functions to match with the most galaxies. On the other hand, science cases that are interested in extreme galaxies would want all possible shapes to be sampled.

One of the main computational parameters we get to choose is the upchannelization parameter U. Olivia Pereira has done work determining how upchannelization affects the rate of detections. If the resolution is too low, then we can't take advantage of galaxies' horns in order to pick them out. This is balanced by the fact that higher upchannelization increases the bandwidth of sending the CHORD data to the DRA computers. One option is to pick U as just high enough that the dip in horned profiles is visible. Section 3.4 showed the complications that frequency aliasing can cause for the search, so another option is to forgo upchannelization entirely at the risk of losing a bit of sensitivity. The worst option is probably in the middle, an upchannelization parameter of $U = 2$ or $U = 4$, which does not improve the resolution very much but brings all the aliasing problems.

4.3 Matching Galaxies with Optical Surveys

For many of the science questions, it is necessary to have optical counterparts to the HI detections. For example, an optical counterpart is necessary to measure gas stripping with respect to stellar mass or to find optically dark galaxies. The ALFALFA team has a process to identify optical counterparts. Its success depends on how much the observing footprint overlaps. ALFALFA's spring sky overlap with SDSS is large, and they matched 83% of galaxies with an SDSS spectroscopy entry [16]. 199 ALFALFA objects were classified as dark, but subsequent efforts discarded or matched many of those [10]. We would like to know if we can achieve a similarly high rate of matching with existing optical catalogues. I will compare the expected CHORD detections to the SDSS DR8 Main Galaxy Sample [2] and the DESI EDR "1% survey" [9].

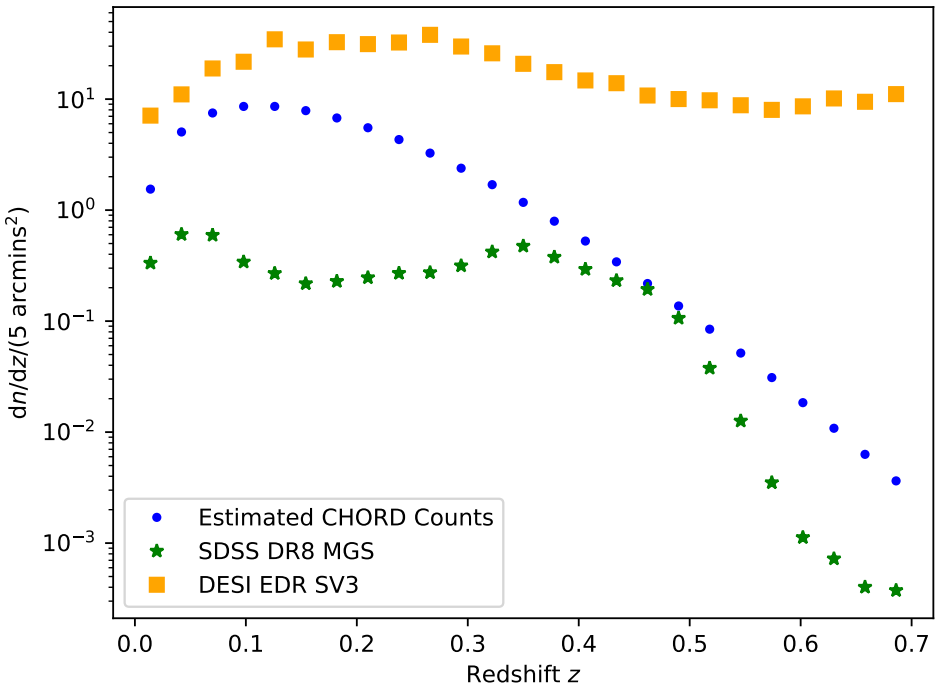


Figure 4.2: Comparing the expected counts of CHORD to the counts of SDSS DR8 MGS and DESI EDR SV3 surveys. The DESI counts are strictly higher than the expected counts of CHORD.

In subsection 1.5.1, I estimated that CHORD will detect 2,684,859 galaxies over a 10,000 deg^2 footprint. The SDSS DR8 Main Galaxy Sample has 221,046 entries and a footprint of 9274 deg^2. The DESI EDR 1% Survey has 736,620 entries and a footprint of 197 deg^2. Entries with warnings were discarded. One pixel of the CHORD survey will have a resolution of about 5 arcminutes^2. For comparison, we normalize the number of counts by the inverse of the footprint. Figure 4.2 shows the counts per CHORD pixel from the three surveys. For ideal matching, the optical survey should have more counts than CHORD. If the optical survey has too few, then it might not include the optical counterparts. If the optical survey has too many, then we can choose the galaxy that is most likely to have a high HI concentration as the most likely counterpart.

Figure 1.5 is focused on the number of galaxies that will be detected, but the more important quantity for some science cases is the minimum mass threshold that CHORD

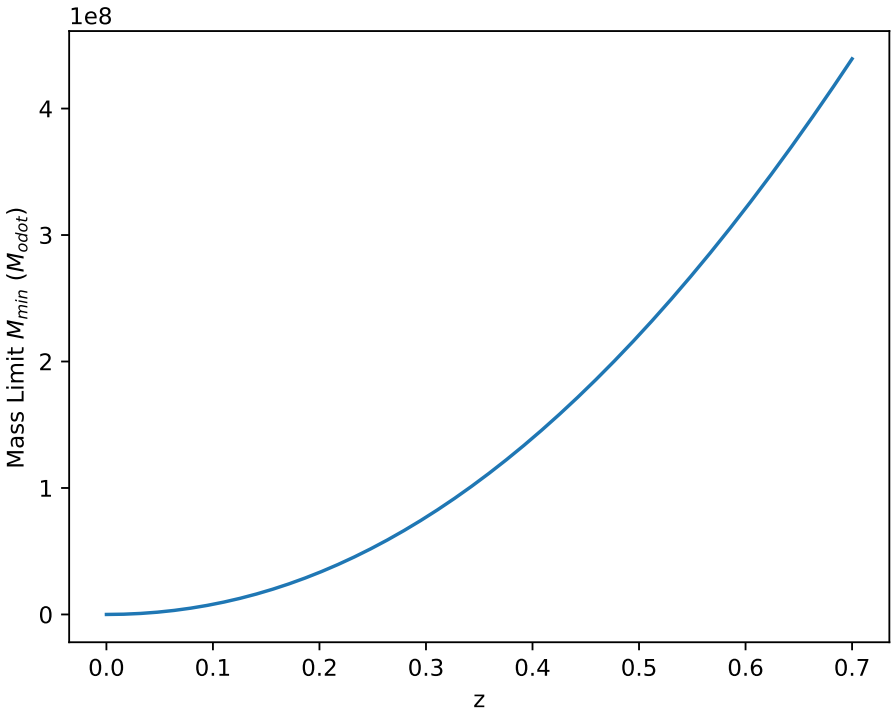


Figure 4.3: The minimum mass galaxy that is expected to meet a 5σ detection.

can observe as a function of redshift. Equation 1.8 can be inverted to give the lowest mass that will be detected at a σ_T threshold as a function of redshift z. This curve is plotted in figure 4.3.

Huang et al. [17] fitted a correlation between total integrated HI mass and stellar mass of a galaxy. It changes depending on the mass of the galaxy. Equation 4.2 is valid for low mass galaxies, and it can be used as an estimate for the stellar mass minimum detection threshold.

$$\log(M_{HI}) = 0.712 \log(M_*) + 3.117 \tag{4.2}$$

The magnitude limit for the SDSS Main Galaxy Sample is 17.77 [1]. The magnitude limit for the DESI Bright Galaxy Survey that I am comparing with is 19.5 [13]. The stellar mass minimum detection threshold can be converted into a magnitude maximum threshold. Figure 4.4 shows that CHORD will be able to detect objects with much lower optical brightness than the optical surveys. This is because the optical brightness decreases quicker

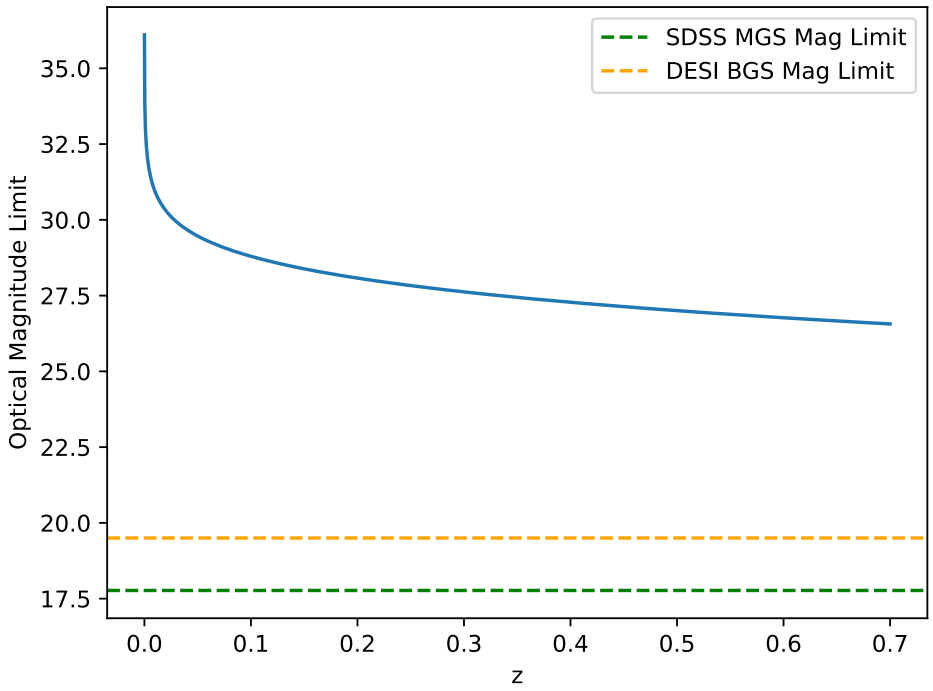


Figure 4.4: Comparing the maximum magnitude galaxy that CHORD is expected to detect as a function of redshift to the magnitude limits of the optical surveys.

than the gas brightness for low mass galaxies. There will be galaxies in the catalogue that cannot be matched with SDSS or DESI.

4.4 Optimizing a Survey for Faint Galaxies

An important class of targets for science is the low mass galaxies. They require high sensitivity to detect. In order to best measure the low end of the HI mass function, for example, we want the largest number of detections possible. Controlling the declination of CHORD allows us to make a choice between a larger area (low declination) and a longer integration time on a smaller area (high declination). Which option optimizes the number of detections of low mass galaxies depends on the shape of the HI luminosity function. Just like in section 1.5.1, we can extrapolate ALFALFA's HI mass function to compute the

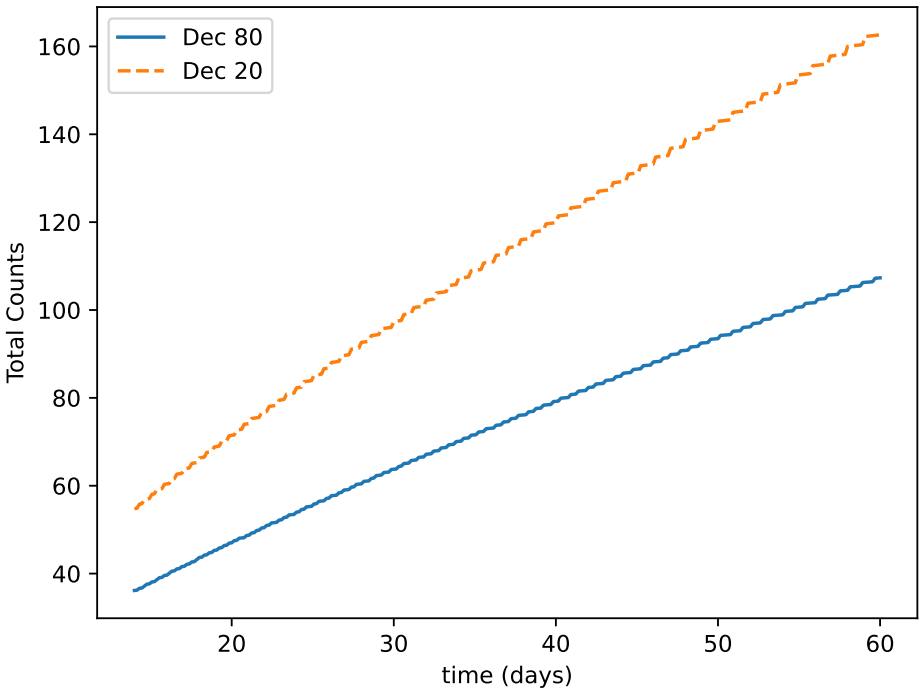


Figure 4.5: Comparing high and low declination options for maximizing counts of faint objects. This is only counting detections of objects with lower mass than 10^7 $M_\odot$. Both scenarios leave CHORD pointing at one declination for the entire period without moving it. The lower declination option (which leads to higher area and lower integration time) produces more counts.

number of detections in both scenarios. Figure 4.5 makes this comparison, and the option of covering a wider area wins.

Chapter 5

Conclusion

This thesis showed development work and performance predictions for the matched filter that will be used in CHORD's galaxy search. A matched filter was chosen as a fast option for automatically flagging galaxy locations and redshifts. The matched filter equation is

$$M_x(d) = \frac{t_x^\dagger N^{-1} d}{\sqrt{t_x^\dagger N^{-1} t_x}}. \tag{5.1}$$

where d is data from CHORD, N is a covariance matrix, and t is a template for a galaxy with x parameters. It has units of sigmas. It will act on a 3D data cube. The spatial and frequency axes were analysed separately.

For the spatial side, the matched filter will act on a visibility matrix which yields the matched filter equation 2.4.

$$M_{spatial} = \frac{\mathrm{tr}({t_u}^\dagger N^{-1} V)}{\sqrt{\mathrm{tr}(t_u^\dagger N^{-1} t_u)}}$$

This allows us to take advantage of the dirty map, which is computed in real time. Because CHORD is a highly redundant array interferometer, this equation will be affected by aliasing. This means the signal from any real source could be explained by a source in multiple sky positions. This is problematic for detecting dim galaxies and deducing the locations of galaxies we do detect. I developed code that quickly estimates the response of the spatial matched filter depending on CHORD's position, the sources position, the size of the array, the wavelength of the source, and the rate of time sampling. This code showed that when only considering one time sample, there is a grid of perfect aliases. Considering

multiple time samples at once breaks these degenerate aliases, but the code showed that there are still situations where aliases approach being perfect. These situations are broken with dithering, which means repointing CHORD by a couple of degrees. I was able to estimate the probability of alias confusion through Monte Carlo given arbitrary severity of aliases.

For the frequency side, the matched filter acts on a spectrum of a single pixel, which yields the matched filter equation 3.5.

$$M_{frequency} = \frac{n\text{idft}(\overline{\text{dft}(t_0)}\text{dft}(R^T N^{-1} d))[\ell]}{\sqrt{t_\ell^T (R^T N^{-1} R) t_\ell}}$$

Where t is a 1-dimensional template for a galaxy profile, and R is the upchannelization response matrix. Following the example of other galaxy searches, we will choose a family of templates that include Gaussian-like single peaks as well as peaks with horns. Using the correlation theorem and fast Fourier transforms, the matched filter operation can be greatly sped up. Upchannelization is useful for increasing the resolution of the data, but it will impact the galaxy search unless it is accounted for, which requires a modification of the matched filter formula.

Through a sensitivity analysis, I estimate that CHORD will detect 2.7 million galaxies over a three year period, which is two orders of magnitude greater than current HI galaxy catalogues. This increased sensitivity will help investigate the gas content of small galaxies and galaxies with higher redshift.

References

[1] Legacy survey target selection. https://www.sdss4.org/dr17/algorithms/legacy_target_selection/#MainGalaxySample. Accessed: 2024-09-22.

[2] Hiroaki Aihara, Carlos Allende Prieto, Deokkeun An, Scott F. Anderson, Éric Aubourg, Eduardo Balbinot, Timothy C. Beers, Andreas A. Berlind, Steven J. Bickerton, Dmitry Bizyaev, Michael R. Blanton, John J. Bochanski, Adam S. Bolton, Jo Bovy, W. N. Brandt, J. Brinkmann, Peter J. Brown, Joel R. Brownstein, Nicolas G. Busca, Heather Campbell, Michael A. Carr, Yanmei Chen, Cristina Chiappini, Johan Comparat, Natalia Connolly, Marina Cortes, Rupert A. C. Croft, Antonio J. Cuesta, Luiz N. da Costa, James R. A. Davenport, Kyle Dawson, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Edward M. Edmondson, Daniel J. Eisenstein, Stephanie Escoffier, Massimiliano Esposito, Michael L. Evans, Xiaohui Fan, Bruno Femenía Castellá, Andreu Font-Ribera, Peter M. Frinchaboy, Jian Ge, Bruce A. Gillespie, G. Gilmore, Jonay I. González Hernández, J. Richard Gott, Andrew Gould, Eva K. Grebel, James E. Gunn, Jean-Christophe Hamilton, Paul Harding, David W. Harris, Suzanne L. Hawley, Frederick R. Hearty, Shirley Ho, David W. Hogg, Jon A. Holtzman, Klaus Honscheid, Naohisa Inada, Inese I. Ivans, Linhua Jiang, Jennifer A. Johnson, Cathy Jordan, Wendell P. Jordan, Eyal A. Kazin, David Kirkby, Mark A. Klaene, G. R. Knapp, Jean-Paul Kneib, C. S. Kochanek, Lars Koesterke, Juna A. Kollmeier, Richard G. Kron, Hubert Lampeitl, Dustin Lang, Jean-Marc Le Goff, Young Sun Lee, Yen-Ting Lin, Daniel C. Long, Craig P. Loomis, Sara Lucatello, Britt Lundgren, Robert H. Lupton, Zhibo Ma, Nicholas MacDonald, Suvrath Mahadevan, Marcio A. G. Maia, Martin Makler, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Claudia Maraston, Daniel Margala, Karen L. Masters, Cameron K. McBride, Peregrine M. McGehee, Ian D. McGreer, Brice Ménard, Jordi Miralda-Escudé, Heather L. Morrison, F. Mullally, Demitri Muna, Jeffrey A. Munn, Hitoshi Murayama, Adam D. Myers, Tracy Naugle, Angelo Fausti Neto, Duy Cuong Nguyen, Robert C. Nichol, Robert W. O'Connell, Ricardo L. C. Ogando, Matthew D. Olm-

stead, Daniel J. Oravetz, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, Parul Pandey, Isabelle Pâris, Will J. Percival, Patrick Petitjean, Robert Pfaffenberger, Janine Pforr, Stefanie Phleps, Christophe Pichon, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Beatriz H. F. Ramos, Céline Reylé, James Rich, Gordon T. Richards, Hans-Walter Rix, Annie C. Robin, Helio J. Rocha-Pinto, Constance M. Rockosi, Natalie A. Roe, Emmanuel Rollinde, Ashley J. Ross, Nicholas P. Ross, Bruno M. Rossetto, Ariel G. Sánchez, Conor Sayres, David J. Schlegel, Katharine J. Schlesinger, Sarah J. Schmidt, Donald P. Schneider, Erin Sheldon, Yiping Shu, Jennifer Simmerer, Audrey E. Simmons, Thirupathi Sivarani, Stephanie A. Snedden, Jennifer S. Sobeck, Matthias Steinmetz, Michael A. Strauss, Alexander S. Szalay, Masayuki Tanaka, Aniruddha R. Thakar, Daniel Thomas, Jeremy L. Tinker, Benjamin M. Tofflemire, Rita Tojeiro, Christy A. Tremonti, Jan Vandenberg, M. Vargas Magaña, Licia Verde, Nicole P. Vogt, David A. Wake, Ji Wang, Benjamin A. Weaver, David H. Weinberg, Martin White, Simon D. M. White, Brian Yanny, Naoki Yasuda, Christophe Yeche, and Idit Zehavi. The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III. *Astrophysical Journal, Supplement*, 193(2):29, April 2011.

[3] Mandana Amiri, Kevin Bandura, Anja Boskovic, Tianyue Chen, Jean-Franç ois Cliche, Meiling Deng, Nolan Denman, Matt Dobbs, Mateus Fandino, Simon Foreman, Mark Halpern, David Hanna, Alex S. Hill, Gary Hinshaw, Carolin Höfer, Joseph Kania, Peter Klages, T. L. Landecker, Joshua MacEachern, Kiyoshi Masui, Juan Mena-Parra, Nikola Milutinovic, Arash Mirhosseini, Laura Newburgh, Rick Nitsche, Anna Ordog, Ue-Li Pen, Tristan Pinsonneault-Marotte, Ava Polzin, Alex Reda, Andre Renard, J. Richard Shaw, Seth R. Siegel, Saurabh Singh, Rick Smegal, Ian Tretyakov, Kwinten Van Gassen, Keith Vanderlinde, Haochen Wang, Donald V. Wiebe, James S. Willis, and Dallas Wulf. An overview of CHIME, the canadian hydrogen intensity mapping experiment. *The Astrophysical Journal Supplement Series*, 261(2):29, jul 2022.

[4] D. G. Barnes, L. Staveley-Smith, W. J. G. de Blok, T. Oosterloo, I. M. Stewart, A. E. Wright, G. D. Banks, R. Bhathal, P. J. Boyce, M. R. Calabretta, M. J. Disney, M. J. Drinkwater, R. D. Ekers, K. C. Freeman, B. K. Gibson, A. J. Green, R. F. Haynes, P. te Lintel Hekkert, P. A. Henning, H. Jerjen, S. Juraszek, M. J. Kesteven, V. A. Kilborn, P. M. Knezek, B. Koribalski, R. C. Kraan-Korteweg, D. F. Malin, M. Marquarding, R. F. Minchin, J. R. Mould, R. M. Price, M. E. Putman, S. D. Ryder, E. M. Sadler, A. Schröder, F. Stootman, R. L. Webster, W. E. Wilson, and

T. Ye. The hi parkes all sky survey: southern observations, calibration and robust imaging. *Monthly Notices of the RAS*, 322(3):486–498, April 2001.

[5] Elias Brinks, Frank Bigiel, Adam Leroy, Fabian Walter, W. J. G. de Blok, Ioannis Bagetakos, Antonio Usero, and Robert C. Kennicutt. Hi in galactic disks. *Proceedings of the International Astronomical Union*, 4(S254):301–306, 2008.

[6] J. J. Condon, W. D. Cotton, E. W. Greisen, Q. F. Yin, R. A. Perley, G. B. Taylor, and J. J. Broderick. The nrao vla sky survey. *The Astronomical Journal*, 115(5):1693, 5 1998.

[7] James J. Condon and Scott M Ransom. Nrao lectures chapter 3: Radio telescopes and radiometers, 2018. url: https://www.cv.nrao.edu/ sransom/web/Ch3.html.

[8] W. J. G. de Blok, F. Walter, E. Brinks, C. Trachternach, S-H. Oh, and R. C. Kennicutt. High-resolution rotation curves and galaxy mass models from things. *The Astronomical Journal*, 136(6):2648–2719, November 2008.

[9] DESI Collaboration Et Al. The early data release of the dark energy spectroscopic instrument, 2023.

[10] Lin Du, Wei Du, Cheng Cheng, Ming Zhu, Haiyang Yu, and Hong Wu. Almost optically dark galaxies in decals (i): Detection, optical properties and possible origins, 2024.

[11] Riccardo Giovanelli and Martha P. Haynes. Extragalactic HI surveys. *The Astronomy and Astrophysics Review*, 24(1), 12 2015.

[12] James E. Gunn and III Gott, J. Richard. On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution. *Astrophysical Journal*, 176:1, August 1972.

[13] ChangHoon Hahn, Michael J. Wilson, Omar Ruiz-Macias, Shaun Cole, David H. Weinberg, John Moustakas, Anthony Kremin, Jeremy L. Tinker, Alex Smith, Risa H. Wechsler, Steven Ahlen, Shadab Alam, Stephen Bailey, David Brooks, Andrew P. Cooper, Tamara M. Davis, Kyle Dawson, Arjun Dey, Biprateep Dey, Sarah Eftekharzadeh, Daniel J. Eisenstein, Kevin Fanning, Jaime E. Forero-Romero, Carlos S. Frenk, Enrique Gaztañaga, Satya Gontcho A Gontcho, Julien Guy, Klaus Honscheid, Mustapha Ishak, Stéphanie Juneau, Robert Kehoe, Theodore Kisner, Ting-Wen Lan, Martin Landriau, Laurent Le Guillou, Michael E. Levi, Christophe Magneville, Paul Martini,

Aaron Meisner, Adam D. Myers, Jundan Nie, Peder Norberg, Nathalie Palanque-Delabrouille, Will J. Percival, Claire Poppett, Francisco Prada, Anand Raichoor, Ashley J. Ross, Sasha Gaines, Christoph Saulder, Eddie Schlafly, David Schlegel, David Sierra-Porta, Gregory Tarle, Benjamin A. Weaver, Christophe Yèche, Pauline Zarrouk, Rongpu Zhou, Zhimin Zhou, and Hu Zou. The desi bright galaxy survey: Final target selection, design, and validation. *The Astronomical Journal*, 165(6):253, May 2023.

[14] Martha P. Haynes, Ricardo Giovanelli, and Guido L. Chincarini. The Influence of Envirionment on the H I Content of Galaxies. *Annual Review of Astron and Astrophysis*, 22:445–470, January 1984.

[15] Martha P. Haynes, Riccardo Giovanelli, Brian R. Kent, Elizabeth A. K. Adams, Thomas J. Balonek, David W. Craig, Derek Fertig, Rose Finn, Carlo Giovanardi, Gregory Hallenbeck, Kelley M. Hess, G. Lyle Hoffman, Shan Huang, Michael G. Jones, Rebecca A. Koopmann, David A. Kornreich, Lukas Leisman, Jeffrey Miller, Crystal Moorman, Jessica O'Connor, Aileen O'Donoghue, Emmanouil Papastergis, Parker Troischt, David Stark, and Li Xiao. The Arecibo Legacy Fast ALFA Survey: The ALFALFA Extragalactic H I Source Catalog. *Astrophysical Journal*, 861(1):49, July 2018.

[16] Martha P. Haynes, Riccardo Giovanelli, Ann M. Martin, Kelley M. Hess, Amélie Saintonge, Elizabeth A. K. Adams, Gregory Hallenbeck, G. Lyle Hoffman, Shan Huang, Brian R. Kent, Rebecca A. Koopmann, Emmanouil Papastergis, Sabrina Stierwalt, Thomas J. Balonek, David W. Craig, Sarah J. U. Higdon, David A. Kornreich, Jeffrey R. Miller, Aileen A. O'Donoghue, Ronald P. Olowin, Jessica L. Rosenberg, Kristine Spekkens, Parker Troischt, and Eric M. Wilcots. The Arecibo Legacy Fast ALFA Survey: The $\alpha.40$ H I Source Catalog, Its Characteristics and Their Impact on the Derivation of the H I Mass Function. *Astronomical Journal*, 142(5):170, November 2011.

[17] Shan Huang, Martha P. Haynes, Riccardo Giovanelli, and Jarle Brinchmann. The Arecibo Legacy Fast ALFA Survey: The Galaxy Population Detected by ALFALFA. *Astrophysical Journal*, 756(2):113, September 2012.

[18] Michael G. Jones, Martha P. Haynes, Riccardo Giovanelli, and Crystal Moorman. The ALFALFA H I mass function: a dichotomy in the low-mass slope and a locally suppressed 'knee' mass. *Monthly Notices of the RAS*, 477(1):2–17, June 2018.

[19] J. Kamphuis and R. Sancisi. Widespread high velocity gas in the spiral galaxy NGC 6946. *Astronomy and Astrophysics*, 273:L31–L34, June 1993.

[20] B. S. Koribalski, L. Staveley-Smith, V. A. Kilborn, S. D. Ryder, R. C. Kraan-Korteweg, E. V. Ryan-Weber, R. D. Ekers, H. Jerjen, P. A. Henning, M. E. Putman, M. A. Zwaan, W. J. G. de Blok, M. R. Calabretta, M. J. Disney, R. F. Minchin, R. Bhathal, P. J. Boyce, M. J. Drinkwater, K. C. Freeman, B. K. Gibson, A. J. Green, R. F. Haynes, S. Juraszek, M. J. Kesteven, P. M. Knezek, S. Mader, M. Marquarding, M. Meyer, J. R. Mould, T. Oosterloo, J. O'Brien, R. M. Price, E. M. Sadler, A. Schröder, I. M. Stewart, F. Stootman, M. Waugh, B. E. Warren, R. L. Webster, and A. E. Wright. The 1000 Brightest HIPASS Galaxies: H I Properties. *Astronomical Journal*, 128(1):16–46, July 2004.

[21] Bärbel S. Koribalski, L. Staveley-Smith, T. Westmeier, P. Serra, K. Spekkens, O. I. Wong, K. Lee-Waddell, C. D. P. Lagos, D. Obreschkow, E. V. Ryan-Weber, M. Zwaan, V. Kilborn, G. Bekiaris, K. Bekki, F. Bigiel, A. Boselli, A. Bosma, B. Catinella, G. Chauhan, M. E. Cluver, M. Colless, H. M. Courtois, R. A. Crain, W. J. G. de Blok, H. Dénes, A. R. Duffy, A. Elagali, C. J. Fluke, B. Q. For, G. Heald, P. A. Henning, K. M. Hess, B. W. Holwerda, C. Howlett, T. Jarrett, D. H. Jones, M. G. Jones, G. I. G. Józsa, R. Jurek, E. Jütte, P. Kamphuis, I. Karachentsev, J. Kerp, D. Kleiner, R. C. Kraan-Korteweg, Á. R. López-Sánchez, J. Madrid, M. Meyer, J. Mould, C. Murugeshan, R. P. Norris, S. H. Oh, T. A. Oosterloo, A. Popping, M. Putman, T. N. Reynolds, J. Rhee, A. S. G. Robotham, S. Ryder, A. C. Schröder, Li Shao, A. R. H. Stevens, E. N. Taylor, J. M. vanÂ der Hulst, L. Verdes-Montenegro, B. P. Wakker, J. Wang, M. Whiting, B. Winkel, and C. Wolf. WALLABY - an SKA Pathfinder HI survey. *Astrophysics and Space Science*, 365(7):118, July 2020.

[22] Vincent MacKay, Mark Lai, Peter Shmerko, Dallas Wulf, Leonid Belostotski, and Keith Vanderlinde. Low-cost, low-loss, ultra-wideband compact feed for interferometric radio telescopes, 2023.

[23] Houjin Mo, Frank van den Bosch, and Simon White. *Galaxy Formation and Evolution*. Cambridge University Press, 2010.

[24] D. Cs. Molnár, P. Serra, T. van der Hulst, T. H. Jarrett, A. Boselli, L. Cortese, J. Healy, E. de Blok, M. Cappellari, K. M. Hess, G. I. G. Józsa, R. M. McDermid, T. A. Oosterloo, and M. A. W. Verheijen. The Westerbork Coma Survey. A blind, deep, high-resolution H I survey of the Coma cluster. *Astronomy and Astrophysics*, 659:A94, March 2022.

[25] R. Morganti, P. T. de Zeeuw, T. A. Oosterloo, R. M. McDermid, D. Krajnović, M. Cappellari, F. Kenn, A. Weijmans, and M. Sarzi. Neutral hydrogen in nearby elliptical and lenticular galaxies: the continuing formation of early-type galaxies. *Monthly Notices of the RAS*, 371(1):157–169, September 2006.

[26] Amé lie Saintonge. The arecibo legacy fast ALFA survey. IV. strategies for signal identification and survey catalog reliability. *The Astronomical Journal*, 133(5):2087–2096, 3 2007.

[27] Paolo Serra, Tobias Westmeier, Nadine Giese, Russell Jurek, Lars Flöer, Attila Popping, Benjamin Winkel, Thijs van der Hulst, Martin Meyer, Bärbel S. Koribalski, Lister Staveley-Smith, and Hélène Courtois. SoFiA: a flexible source finder for 3D spectral line data. *Monthly Notices of the Royal Astronomical Society*, 448(2):1922–1929, 02 2015.

[28] Kendrick Smith. Chord gpu upchannelization. Internal Document, 2024.

[29] John S. Townsend. *A Modern Approach to Quantum Mechanics*. University Science Books, London, 2012.

[30] Keith Vanderlinde, Adrian Liu, Bryan Gaensler, Dick Bond, Gary Hinshaw, Cherry Ng, Cynthia Chiang, Ingrid Stairs, Jo-Anne Brown, Jonathan Sievers, Juan Mena, Kendrick Smith, Kevin Bandura, Kiyoshi Masui, Kristine Spekkens, Leo Belostotski, Matt Dobbs, Neil Turok, Patrick Boyle, Michael Rupen, Tom Landecker, Ue-Li Pen, and Victoria Kaspi. The canadian hydrogen observatory and radio-transient detector (chord). 2019.

[31] Fabian Walter, Elias Brinks, W. J. G. de Blok, Frank Bigiel, Robert C. Kennicutt, Michele D. Thornley, and Adam Leroy. Things: The hi nearby galaxy survey. *The Astronomical Journal*, 136(6):2563, 11 2008.

[32] T. Westmeier, R. Jurek, D. Obreschkow, B. S. Koribalski, and L. Staveley-Smith. The busy function: a new analytic function for describing the integrated 21-cm spectral profile of galaxies. *Monthly Notices of the RAS*, 438(2):1176–1190, February 2014.

[33] T. Westmeier, S. Kitaeff, D. Pallot, P. Serra, J. M. van der Hulst, R. J. Jurek, A. Elagali, B. Q. For, D. Kleiner, B. S. Koribalski, K. Lee-Waddell, J. R. Mould, T. N. Reynolds, J. Rhee, and L. Staveley-Smith. SOFIA 2 - an automated, parallel H I source finding pipeline for the WALLABY survey. *Monthly Notices of the RAS*, 506(3):3962–3976, September 2021.

[34] Chuan-Peng Zhang, Ming Zhu, Peng Jiang, Cheng Cheng, Jing Wang, Jie Wang, Jin-Long Xu, Xiao-Lan Liu, Nai-Ping Yu, Lei Qian, Haiyang Yu, Mei Ai, Yingjie Jing, Chen Xu, Ziming Liu, Xin Guan, Chun Sun, Qingliang Yang, Menglin Huang, and Qiaoli Hao. The fast all sky h i survey (fashi): The first release of catalog. *Science China Physics, Mechanics & Astronomy*, 67(1), 12 2023.

APPENDICES

Appendix A

Matched Filter Proof

The situation: There are two 1-dimensional vectors d and t. We want to compare two hypotheses:

H_0: Each element of the data vector is produced by a Gaussian distribution (possibly correlated with other elements) centered at 0.

H_1: Each element of the data vector is produced by a Gaussian distribution (possibly correlated with other elements) centered at 0 plus the template rescaled by some parameter μ.

The probability $P(d|H_0)$ is proportional to $e^{-\frac{1}{2}(\sum_i d_i^2/\sigma_i^2+\sum_{i,j} c_{i,j}d_i d_j)}$. Where $c_{i,j}$ are correlation variables. We'll simplify this equation by combining the σ_i and $c_{i,j}$ into a covariance matrix N.

$$P(d|H_0) \propto e^{-\frac{1}{2}d^T N^{-1} d}$$

In the H_1 case, the probability is

$$P(d|H_1) \propto e^{-\frac{1}{2}(\mu t-d)^T N^{-1}(\mu t-d)}$$

For our hypothesis test, we are interested in the probability ratio

$$\begin{aligned}\frac{P(H_1|d)}{P(H_0|d)} &= \frac{P(d|H_1)P(H_1)}{P(d)}\frac{P(d)}{P(d|H_0)P(H_0)} \\ &= \frac{P(d|H_1)}{P(d|H_0)}\end{aligned}$$

using Bayes' theorem assuming the priors are the same for both hypotheses. Substituting the probabilities gives

$$\frac{P(H_1|d)}{P(H_0|d)} \propto e^{-\frac{1}{2}(\mu t-d)^T N^{-1}(\mu t-d)+\frac{1}{2}d^T N^{-1}d}$$

We want to optimize this probability, so we can consider the log of this ratio:

$$\begin{aligned} 2\log(\text{prob}) + 2\log(\text{const}) &= -(\mu t - d)^T N^{-1}(\mu t - d) + d^T N^{-1} d \\ &= -\mu^2 t^T N^{-1} t + \mu t^T N^{-1} d + \mu d^T N^{-1} t - d^T N^{-1} d + d^T N^{-1} d \end{aligned}$$

N^{-1} is a symmetric matrix, so we have $t^T N^{-1} d = d^T N^{-1} t$.

$$= \mu^2 t^T N^{-1} t - 2\mu t^T N^{-1} d$$

A matched filter is the optimal filter for optimizing probability, so we can set μ to the value that optimizes $2\log(\text{prob}) + 2\log(\text{const})$. This is a parabola with respect to μ, so the optimal μ is $\mu = \frac{t^T N^{-1} d}{t N^{-1} t}$. If we plug this μ back in, we get

$$\text{opt}(2\log(\text{prob}) + 2\log(\text{const})) = \frac{(t^T N^{-1} d)^2}{t^T N^{-1} t}$$

The left hand side of this equation is the delta χ^2 between the two hypotheses. The delta degrees of freedom between the two hypothesis is 1 because H_1 only has the extra μ parameter. This the case where ddof = 1, the 1 sigma probability is reached at χ, which is $\frac{t^T N^{-1} d}{\sqrt{t^T N^{-1} t}}$. Therefore, the matched filter variable has units of SNR.

Appendix B

Omitting the Re() in the matched filter equation

If you're running a matched filter on a complex vector, you want the result to be the same as if you ran the matched filter on a real vector that is twice as long. To do this, for every element of the complex vectors we want to multiply the real and complex part of the template with those of the data vector as if everything is real. This is accomplished by taking the real part of the conjugate transpose multiplication. The matched filter equation generalized for complex-valued vectors should be written:

$$M = \frac{\mathrm{Re}(t^\dagger d)}{\sqrt{t^\dagger t}}$$

This Re() is not necessary if $t^\dagger d$ is guaranteed to be real, which it is if t and d are hermitian matrices, as is the case with visibilities.

Define T and D as the "real" and unwrapped versions of t and d, which are complex Hermitian matrices of size $N \times N$. That is to say the real and complex parts of t and d get placed into a 1-dimensional vector that's twice as long as the original. We want to prove

that $t^\dagger d = T^T D$.

$$
\begin{aligned}
T^T D &= \sum_{i=1}^{2\times N\times N} T_i D_i \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{Re}(t^*_{ij} d_{ij}) \\
&= \mathrm{Re}(\mathrm{Tr}(t^\dagger d)) \\
&= \tfrac{1}{2}(\mathrm{Tr}(t^\dagger d) + \mathrm{Tr}(t^\dagger d)^*) \\
&= \tfrac{1}{2}(\mathrm{Tr}(t^\dagger d) + \mathrm{Tr}(t^{\dagger *} d^*)) \\
&= \tfrac{1}{2}(\mathrm{Tr}(t^\dagger d) + \mathrm{Tr}(t^{\dagger T} d^T)) \qquad \text{using Hermitian} \\
&= \tfrac{1}{2}(\mathrm{Tr}(t^\dagger d) + \mathrm{Tr}(t^\dagger d)) \qquad \text{Trace properties} \\
&= \mathrm{Tr}(t^\dagger d)
\end{aligned}
$$

Appendix C

Closest Alias Locations in the Instantaneous Case

The full analytic matched filter equation is from equation 2.7:

$$R_{uu_s} = \frac{1}{m_1^2 m_2^2} \frac{\sin\left(\frac{\pi m_1 L_1}{\lambda} \hat{e}_1 \cdot (-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_1}{\lambda} \hat{e}_1 \cdot (-\vec{u} + \vec{u}_s)\right)^2} \frac{\sin\left(\frac{\pi m_2 L_2}{\lambda} \hat{e}_2 \cdot (-\vec{u} + \vec{u}_s)\right)^2}{\sin\left(\frac{\pi L_2}{\lambda} \hat{e}_2 \cdot (-\vec{u} + \vec{u}_s)\right)^2}$$

The true source is at $\vec{u}_s$ where R_{uu_s} evaluates to 1. We want to find other values of $\vec{u}$ such that R_{uu_s} evaluates to 1. Figure 2.1 shows that these alias positions are on a rectangular grid. For an example, we'll find the location of the first northward alias.

Since the sin is squared and m_1 is an integer, the first sin squared fraction will evaluate to m_1 when

$$\frac{L_1}{\lambda} \hat{e}_1 \cdot (-\vec{u} + \vec{u}_s) \in \mathbb{Z}$$

This translates to a condition on $\vec{u}$ for the first northward alias:

$$\hat{e}_1 \cdot \vec{u} = \hat{e}_1 \cdot \left(\vec{u}_s - \frac{\lambda}{L_1} \hat{e}_1\right)$$

We require that the $\vec{e}_2$ component be the same as the source:

$$\hat{e}_2 \cdot \vec{u} = \hat{e}_2 \cdot \vec{u}_s$$

Lastly we require that $\vec{u}$ be a unit vector:

$$\begin{aligned} & \vec{u} \cdot \vec{u} = 1 \\ \implies & \hat{e}_3 \cdot \vec{u} = \sqrt{1 - \left(\hat{e}_1 \cdot \left(\vec{u}_s - \frac{\lambda}{L_1} \hat{e}_1 \right) \right)^2 - (\hat{e}_2 \cdot \vec{u}_s)^2} \end{aligned}$$

This indicates that a northward alias does not always exist. For example, if $\vec{u}_s = \hat{e}_2$, then the radicand becomes negative.

$\hat{e}_1, \hat{e}_2, \hat{e}_3$ form a basis set, so these three conditions fully define $\vec{u}$.

Appendix D

Correlation Theorem Proof

We can implement discrete correlations by doing Fourier transforms. This allows us to take advantage of fast Fourier transforms. Take two arrays f and g of length N. F and G are the Fourier transforms. Overline means complex conjugate.

$$\begin{aligned}
\mathrm{corr}(f,g)[n] &= \sum_{m=0}^{N-1} \overline{f}[m] g[n+m] \\
&= \sum_{m=0}^{N-1} \left(\sum_{k=0}^{N-1} \overline{F[k] e^{2\pi i k m/N}} \right) \left(\sum_{\ell=0}^{N-1} G[\ell] e^{2\pi i \ell (m+n)/N} \right) \\
&= \sum_{m=0}^{N-1} \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \overline{F}[k] G[\ell] e^{2\pi i \ell n/N} e^{2\pi i (\ell-k) m/N} \\
&= \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \overline{F}[k] G[\ell] e^{2\pi i \ell n/N} \sum_{m=0}^{N-1} e^{2\pi i (\ell-k) m/N} \\
&= \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \overline{F}[k] G[\ell] e^{2\pi i \ell n/N} \, (N \delta_{\ell k}) \\
&= N \sum_{\ell=0}^{N-1} \overline{F}[\ell] G[\ell] e^{2\pi i \ell n/N} \\
&= N \ \mathrm{idft}(\overline{\mathrm{F}}\mathrm{G})[\mathrm{n}]
\end{aligned}$$

Overall factor of N depends on Fourier transform convention.